Native Real-Time Interactive Omni Intelligence: A Unified Framework

Yifan Shen[1,*], Zhuoqing Zhong[2,*], Xinzhuo Li[1,*], Jiateng Liu[1,*], Bingxuan Li[1,*], Wei Cao[1,*], Pei Tian[3], Lin Zhu[1], Jian Xu[1], Zewei Chen[1], Tianjiao Yu[1], Haichao Zhang[4], Jiawen Zhang[1], Yuner Zhang[5], Brian Nlong Zhao[1], Sen Fang[6], Nan Huang[1], Bowen Fang[3], Chen Fang[1], Boyi Li[1], Onkar Susladkar[1], Frank Yang[1], Yuanzhe Liu[1], Jingyuan Zhu[5], Cheng Qian[1], Yijiang Li[7], Xinyang Han[8], Yuan Shen[1], Xiang Li[1], Bolin Lai[1], Xu Cao[1], Chengde Wan[9], Yiwen Song[10], Yue Guo[1], Heng Ji[1], Xuan Di[3], Yiwei Wang[11], Yaoyao Liu[1], Dimitris N. Metaxas[6], Yun Fu[4], James M. Rehg[1], Chengxiang Zhai[1], Ismini Lourentzou[1,†]

[1]University of Illinois Urbana-Champaign [2]Carnegie Mellon University
[3]Columbia University [4]Northeastern University [5]University of Pennsylvania
[6]Rutgers University [7]University of California, San Diego [8]University of California, Berkeley
[9]Independent Researcher [10]Google [11]University of California, Merced

[*]Core Contributor [†]Corresponding Author

⭐ **Awesome Omni** · GitHub

Abstract

The rapid progress of large language models has established strong foundations for reasoning, knowledge, and generation, while the integration of vision, audio, video, and other modalities has extended these capabilities toward increasingly unified intelligence across modalities, which we refer to as Omni Intelligence. Yet, real-world assistants must do more than process completed multimodal prompts. They must perceive evolving streams, respond while new evidence is arriving, handle hesitation, overlap, interruption, and changing intent, and coordinate reasoning or tool use that may proceed asynchronously. We refer to this broader capability as **Interactive Omni Intelligence**: the ability to maintain and exercise omni intelligence under the temporal and adaptive demands of real-time interaction. The question guiding this paper is: *How can foundation-model capabilities be carried into real-time interaction while preserving both intelligence and interactivity?*

We develop a conceptual framework connecting two development paths. One adapts pretrained language or omni models through streaming interfaces, modality modules, post-training, and orchestration. The other develops model-native systems that learn temporal structure, continuous multimodal processing, generation, and interaction control. Across both paths, we compare representations, architectures, fusion, data, training objectives, serving, and evaluation. Their comparison reveals an Intelligence–Interactivity Trade-off between preserving reasoning and knowledge, and meeting the timing, coordination, and resource demands of continuous operation. We examine Multi-Rate Interactive Intelligence as a cross-cutting direction, including foreground–background designs that coordinate fast interaction with slower reasoning, retrieval, memory, and tools. Finally, we identify open problems in evolving state, persistent and proactive interaction, joint evaluation, and safe open-world operation. The resulting framework connects capability preservation, interaction control, and multi-rate coordination.

Keywords: omni intelligence; omni-models; multimodal foundation models; real-time interaction; full-duplex systems; streaming generation; multi-rate intelligence

Contents

1 Introduction **3**

2 A Framework for Interactive Omni Intelligence **5**

2.1 From Omni Intelligence to Real-World Interactivity . . . 5

2.2 Taxonomy of System Families . . . 7

3 Streaming Multimodal Representations **8**

3.1 Audio Representations . . . 9

3.2 Video Representations . . . 11

3.3 Time, Silence, and Overlap as Model Inputs . . . 12

3.4 Unified Stream Formats . . . 14

3.5 Speech Synchronization and User Interruption . . . 15

4 Building Native Omni Models for Real-Time Interaction **18**

4.1 Architecture Families and Modality Fusion . . . 18

4.2 Data . . . 23

4.3 Training Objectives . . . 27

5 Maintaining Omni Intelligence **29**

5.1 Temporal Grounding . . . 29

5.2 Memory and Session State . . . 30

5.3 Reasoning and Tool Use under Real-Time Constraints . . . 32

5.4 Factual Grounding and Correctness . . . 34

5.5 Social and Pragmatic Understanding . . . 35

5.6 Synthesis: Interdependent Capabilities . . . 36

6 Realizing Omni Interactivity **37**

6.1 Latency and Responsiveness . . . 37

6.2 Turn-Taking, Silence, and Backchannels . . . 38

6.3 Interruption and Repair . . . 39

6.4 Proactive Interaction . . . 40

6.5 Adaptation Across Turns . . . 40

6.6 Full-Duplex Serving and Client Coordination . . . 41

7 Evaluating Intelligence and Interactivity **42**

7.1 Evaluation Metrics . . . 42

7.2 Comparative Benchmark Coverage . . . 45

7.3 Limitations of Existing Benchmarks . . . 48

8 The Intelligence–Interactivity Trade-off **50**

8.1 Capability Preservation under Real-Time Omni Adaptation . . . 50

8.2 Runtime Trade-offs . . . 52

8.3 Safety, Privacy, and Evaluation Implications . . . 53

9 Design Principles and Open Problems **54**

9.1 Multi-Rate Interactive Intelligence . . . 54

9.2 Persistent, Proactive, and Social Interaction . . . 55

9.3 Safe and Efficient Operation in Open-World Settings . . . 56

9.4 Outlook . . . 57

10 Conclusion **57**

1. Introduction

Artificial intelligence has progressed from language-centered models toward multimodal foundation models that can perceive, reason over, and generate content across multiple modalities such as text, images, audio, video, 3D geometry, dynamic 4D representations, and other forms of contextual information [19, 27, 29, 256, 315, 317, 320–323]. Recent omni models integrate several of these modalities within a shared backbone [1, 89, 204, 295]. We use **Omni Intelligence** to describe this breadth of multimodal perception, reasoning, knowledge, and generation.

Many current omni models still formulate multimodal interaction as completed prompts or conversational turns [1, 89, 295]. The model receives a sufficiently complete context and produces a response afterward. However, in real-world interaction, audio and video continue to arrive while the model is reasoning or generating a response, and results from retrieval, tools, or external computation may arrive after the interaction state has changed.

We define **Interactive Omni Intelligence** as the ability to maintain and use omni intelligence while multimodal input, system output, and interaction state continue to evolve. The model must judge when the available evidence is sufficient, whether silence signals hesitation or turn completion, whether speech is addressed to the system, and how new input should affect an active response. Pauses, overlap, backchannels, interruptions, and revisions therefore carry both semantic and control information [70, 116, 198].

Current systems support live interaction through three primary ways. Adapted systems add streaming interfaces, modality modules, and interaction controllers to pretrained language or omni models. Cascaded VAD–ASR–LLM–TTS systems are a common instance of this category [134, 222, 225]. Model-native systems represent temporal events explicitly and learn speech timing, overlap, interruption handling, or multimodal generation within the model. This group includes speech-native full-duplex models and streaming omni input-output models [61, 198, 220, 301, 302]. Foreground–background systems connect a low-latency interaction process to slower reasoning, retrieval, memory, or tool execution [11, 48, 112]. Such systems can be combined with adapted or model-native systems to coordinate fast interaction with slower computation.

These designs balance intelligence and interactivity differently. Adapted systems retain strong pretrained reasoning and knowledge, although their interaction loop may depend on coordination across several interfaces. Model-native systems incorporate temporal events directly into their state updates. Their requirements include synchronized interaction data, additional streaming computation, and training that preserves existing capabilities [55, 304]. Foreground–background systems allow acknowledgements, turn management, and interruption handling to proceed while slower computation continues. Returned results must then be checked against the latest interaction state. These competing demands during training and runtime give rise to the **Intelligence–Interactivity Trade-off** [147, 252, 286].

The central challenge is therefore to build systems that can continuously update their understanding of the world, coordinate processes operating at different temporal scales, and express their capabilities at the appropriate moment without sacrificing reasoning quality, responsiveness, or safety. By organizing existing work around this joint intelligence–interactivity perspective, we aim to provide a common conceptual foundation for comparing current systems, identifying missing capabilities, and guiding the development of omni models that can participate naturally and reliably in persistent, open-ended interaction.

We analyze two closely related requirements. **Omni Intelligence** concerns the system's ability to integrate information across modalities, understand the situation, and reason about it. **Omni Interactivity** concerns its ability to respond appropriately as the interaction unfolds: deciding when to respond or act, when to wait, and when to stop or revise an ongoing response. These requirements are coupled because incoming

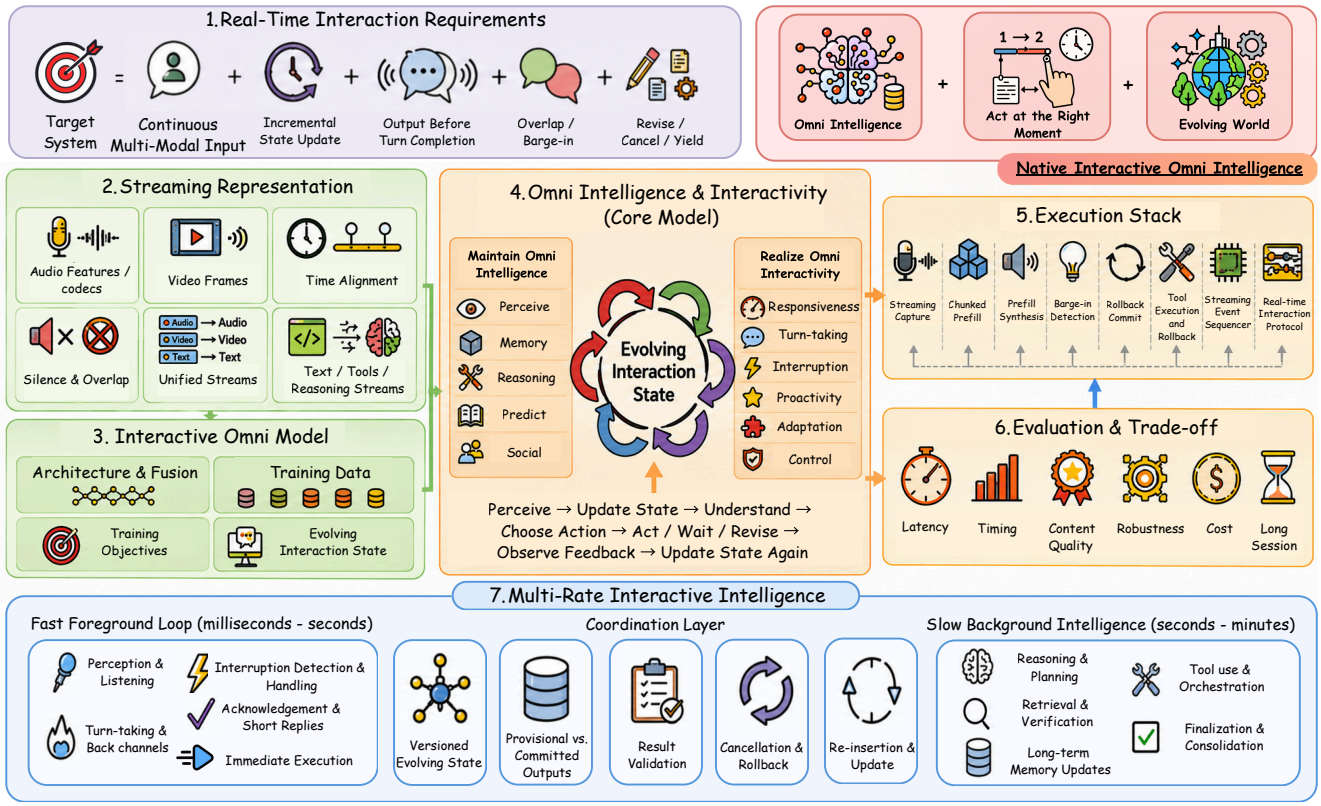

Figure 1: Native Interactive Omni Intelligence architecture. The framework integrates continuous multimodal interaction, streaming representations, an interactive omni-model, and a core intelligence layer that maintains an evolving interaction state. It supports real-time capabilities such as turn-taking, interruption handling, adaptation, tool use, and streaming execution. A multi-rate design coordinates fast foreground interaction with slower background reasoning, retrieval, memory updates, and orchestration, while balancing latency, quality, robustness, cost, and long-session performance.

information can change both what the system understands and how it should respond. Figure 1 summarizes the framework, from real-time interaction requirements and streaming representations to model construction, execution, evaluation, and multi-rate coordination.

We first define the field and position it relative to multimodal language models, spoken dialogue systems, streaming models, and interactive agents in Section 2. With this foundation, we further address three questions. First, how should continuous multimodal streams be represented, and how should models be built and trained to process them? Sections 3 and 4 examine representations, architectures, data, and training objectives. Second, how can a system maintain multimodal understanding and coordinate its responses during ongoing interaction? Sections 5 and 6 cover evolving state, grounding, reasoning, memory, responsiveness, turn-taking, interruption, and serving. Third, how should intelligence and interactivity be evaluated together, and how can their competing demands be managed? Finally, Section 7 covers evaluation, Section 8 analyzes capability and runtime tensions, and Section 9 develops the open research agenda.

Contributions

- **A framework for Interactive Omni Intelligence.** We define Interactive Omni Intelligence through multimodal understanding and response control during ongoing interaction, position it relative to adjacent fields, and compare adapted, model-native, and foreground–background implementations.
- **An end-to-end synthesis of native interaction.** We connect streaming representations, architectures, modality fusion, training data and objectives, state maintenance, response control, and real-time serving as successive layers of native interaction.
- **Joint evaluation of intelligence and interactivity.** We organize evaluation around latency, timing, content, robustness, long-session quality, and repair, and map current benchmarks to the combinations of intelligence and interactive behavior they evaluate.
- **The Intelligence–Interactivity Trade-off and a multi-rate research agenda.** We synthesize evidence that real-time adaptation and execution create capability-preservation and runtime coordination challenges, and identify Multi-Rate Interactive Intelligence, evolving state, persistent and proactive interaction, and safe open-world operation as central directions toward systems that can preserve intelligence while supporting natural interaction.

2. A Framework for Interactive Omni Intelligence

As shown in Figure 1, our proposed conceptual framework organizes Interactive Omni Intelligence around two coupled components: Omni Intelligence and Omni Interactivity. This section develops the framework by distinguishing modality coverage from real-time interaction requirements and by positioning related system families according to how they implement interaction control.

2.1. From Omni Intelligence to Real-World Interactivity

Omni Intelligence concerns the breadth and integration of model capabilities across language, vision, audio, and other modalities [19, 317]. Interactive Omni Intelligence adds a temporal dimension: the ability to use these capabilities and adapt responses as new information arrives and interaction unfolds. Research toward this goal draws on spoken dialogue systems [9, 44, 122, 251], multimodal foundation models [19, 27], streaming generation [61, 298, 301], low-latency serving [4, 140], and interactive agents [277, 287]. Although these areas often share modalities and components, they differ in their assumptions about input availability, response timing, and whether ongoing responses can be interrupted or revised. We therefore distinguish the modalities a system supports from how it processes incoming information and coordinates its responses over time.

We characterize real-time Interactive Omni Intelligence through three properties. First, **Streaming Input**: input can arrive incrementally through audio and video streams, text updates, changes in screen state, tool feedback, and other signals. Second, **Evolving State**: the system can incorporate new information into its interaction state while processing earlier input or producing a response. Third, **Online Response Control**: it can initiate, revise, or stop outputs while interaction is ongoing, including speech, text, tool calls, and interface actions, without waiting for a fixed end-of-turn boundary. Together, these properties distinguish continuous interaction from conventional turn-based multimodal inference, in which a system receives a completed input before generating a response. Figure 2 illustrates this distinction in the context of spoken interaction.

These properties describe interactive behavior without specifying how it is implemented. An adapted system can meet these requirements through streaming interfaces, external controllers, or orchestration around a pretrained model [166, 281]. We use *native interactive omni model* for models that learn temporal alignment,

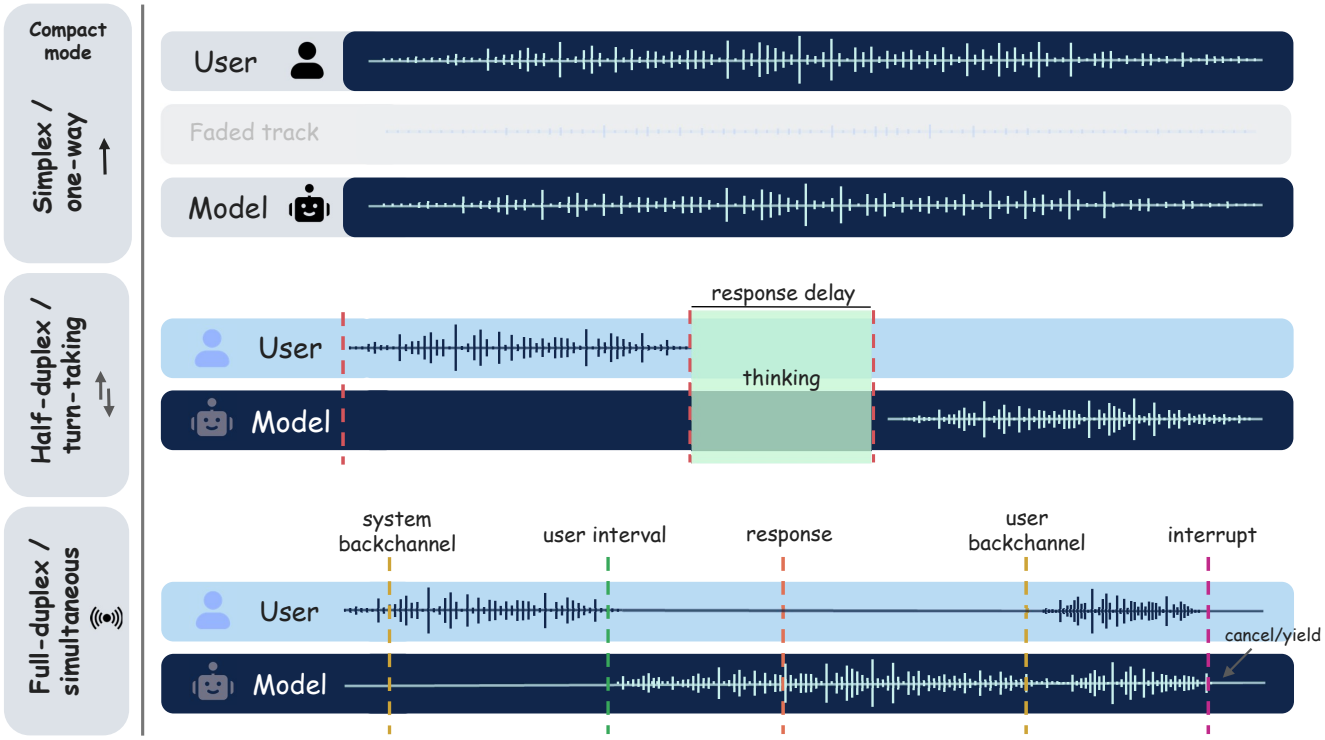

Figure 2: Spoken interaction in simplex, half-duplex, and full-duplex systems. Communication is unidirectional, alternates between participants, or allows overlapping speech and simultaneous listening and speaking, respectively.

incremental multimodal processing, and interaction control as integral parts of their architecture and training. Interactivity describes system behavior during ongoing interaction, while nativeness identifies where the mechanisms supporting that behavior are implemented and learned.

This distinction positions the survey at the intersection of two research directions: full-duplex spoken language modeling and general omni and multimodal modeling. Full-duplex spoken language modeling studies temporal coordination, overlapping speech, interruption, and turn-taking, primarily in spoken interaction [44, 122]. General omni and multimodal modeling supports a broader range of perceptual and generative capabilities, but many systems still assume completed inputs or discrete conversational turns [19, 27]. Interactive Omni Intelligence brings these directions together: it asks how models can maintain broad multimodal understanding and generation while continuously incorporating new information, coordinating their outputs, and adapting to changes in the interaction. This interaction couples updates to the system's understanding with response selection based on the updated state [331].

Within this conceptual framework, characterizing Omni Interactivity requires clarifying the relationship between response latency and continuous interaction. Endpoint detection [251] and inference optimizations such as KV-cache management [140, 282], speculative decoding [146], and latency-aware serving [4] can reduce response latency without changing a system's turn-based operation. Conversely, a system can process streaming inputs and support interruption while still requiring substantial computation. We therefore examine which input streams can update the system's state, when outputs can begin or change, and which components control multimodal fusion, response timing, and interruption. This also distinguishes our scope from surveys of spoken dialogue, spoken language models, and full-duplex speech interaction [9, 44, 122], and provides a common framework for comparing speech-native, multimodal, modular, and agentic systems.

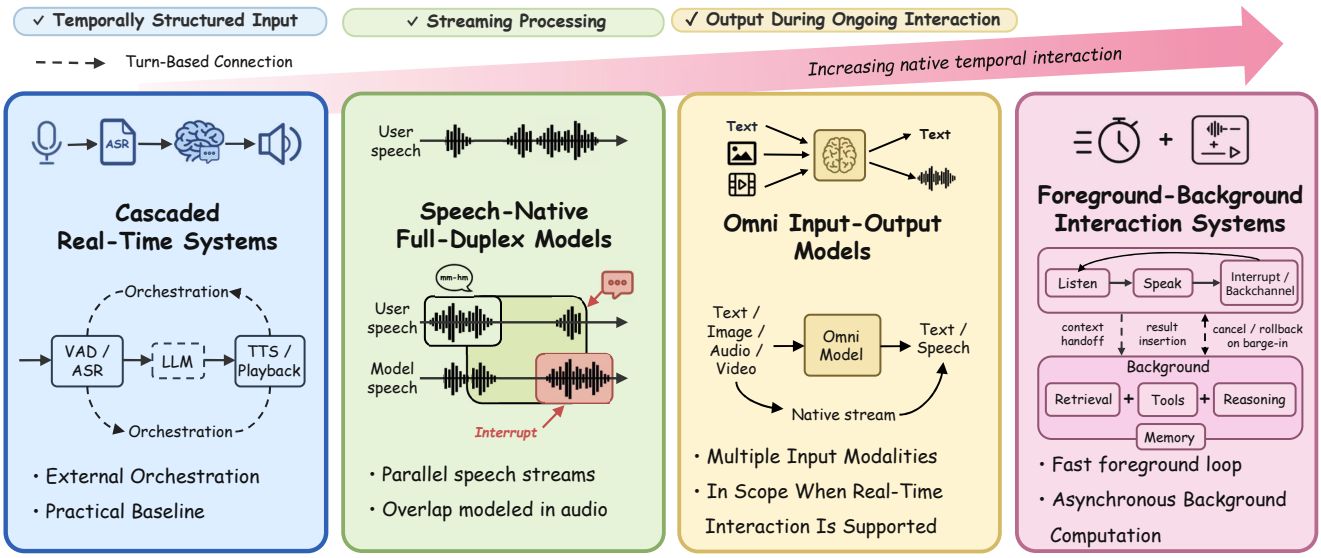

Figure 3: System families for real-time Interactive Omni Intelligence. Cascaded systems use external interaction controllers; speech-native full-duplex models learn speech timing and coordination; omni models support broader modalities with varying interactive capabilities; and foreground–background systems coordinate processes operating at different rates.

2.2. Taxonomy of System Families

We use the conceptual framework to examine four overlapping system families, followed by examples of how deployed systems combine them. They illustrate different combinations of modality coverage, temporal interaction capabilities, and the locus of interaction control. Figure 3 summarizes their relationships.

Cascaded real-time systems. A conventional real-time assistant chains separately trained components such as voice activity detection, streaming ASR, an LLM, and incremental TTS [44, 122]. Streaming recognition can build on online mechanisms such as RNN-Transducer and strong speech front-ends such as Whisper [95, 222]. These systems are modular, deployable, and easy to debug, and sufficiently aggressive streaming can provide low perceived latency. However, timing, turn-taking, interruption, and recovery are typically coordinated outside the language model [44, 166]. Textual interfaces between components may also discard information carried by prosody, laughter, hesitation, overlap, and backchannel timing [61, 198]. Cascaded systems therefore provide a modular baseline in which real-time interaction is coordinated primarily by components outside the pretrained model.

Speech-native full-duplex models. Speech-native full-duplex models represent user and model speech as synchronized or parallel streams and learn aspects of turn-taking, overlap, interruption, and backchanneling directly from conversational sequences. dGSLM [198], Moshi [61], and SyncLLM [268] illustrate this progression through two-channel dialogue modeling, parallel user–model speech streams, and language modeling synchronized to a continuous interaction clock. These models directly learn temporal coordination, primarily through speech, and provide methods that can inform broader multimodal interaction.

Omni input and output models. Omni models broaden perceptual and generative capabilities across text, images, audio, and video. Qwen2.5-Omni [301], for example, uses block-wise audio and visual processing, temporally aligned multimodal position encoding, and a Thinker–Talker architecture supporting streaming text and speech generation; Qwen2-VL provides an important vision–language predecessor [278]. LLaMA-

Omni [74], LLaMA-Omni2 [75], AnyGPT [333], Mini-Omni [298], Mini-Omni2 [299], VITA [80], and MiniMind-O [91] further illustrate efforts to unify multimodal perception with text or audio generation. However, modality breadth alone does not imply Interactive Omni Intelligence. Streaming output, for example, does not by itself establish that new input can continuously modify an ongoing response, that overlap and interruption are modeled, or that the system can revise its behavior as the interaction evolves. Our core scope therefore focuses on systems that combine these multimodal capabilities with the real-time interaction properties defined above.

Foreground–background and multi-rate systems. This system family cuts across the preceding categories and focuses on coordinating components that operate at different computational rates. A fast foreground process remains synchronized with the interaction, while slower processes perform retrieval, reasoning, memory access, planning, or tool use asynchronously. Building on retrieval-augmented generation [149], MoshiRAG [48] and Stream RAG [11] demonstrate how retrieval or tool use can proceed while spoken interaction remains active. STITCH interleaves speech generation with internal reasoning, while EgoMem coordinates asynchronous retrieval, dialogue, and long-term memory processes [47, 314].

We view this family as a broader pattern of Multi-Rate Interactive Intelligence. It directly addresses the Intelligence–Interactivity Trade-off developed in this survey: conversational interaction may require decisions on the scale of tens or hundreds of milliseconds, whereas reasoning, retrieval, memory consolidation, or tool execution may require substantially longer computation. Separating these processes allows slower computation to continue without blocking immediate responses, but requires synchronization, validation of returned results, cancellation, and updates to the interaction state. Multi-rate designs can appear in both adapted and model-native systems.

Deployed real-time systems. Recent industry systems support ongoing multimodal interaction. Google's Gemini Live [92] supports low-latency interaction over speech and visual inputs, allowing the system to respond as new speech and visual information arrives. Thinking Machines Lab's Interaction Model [262] advances this direction by representing interaction as continuous, time-aligned multimodal streams, processing audio, video, and text through fine-grained micro-turns so that pauses, interruptions, overlapping speech, and visual events remain part of the model's temporal context. OpenAI's GPT-Realtime-2 [205] targets real-time conversational applications with native text and audio input/output, image input, tool use, and low-latency streaming, although its currently documented interface does not include native video input. Alibaba's Qwen3.5-Omni-Realtime [6] extends real-time interaction to text, image, audio, and video inputs, supporting streaming spoken responses and multimodal grounding within a persistent interactive session. Similarly, ByteDance's SeedRealtime [26] is designed as a native audio-visual full-duplex model that jointly processes continuous audio, video, and text streams, with an emphasis on natural conversational timing, visual awareness, and integrated tool use. These systems increasingly integrate multimodal perception, response generation, and temporal coordination within real-time models.

3. Streaming Multimodal Representations

Native interactive models must process information as it arrives, without assuming that all relevant input is already available. Audio, video, text, and tool feedback, therefore, require representations that preserve their timing and support incremental processing by the model backbone. This section reviews these representations and four consequences of their design: the token rate the model must sustain, the information it can retain, the minimum delay before new input can influence a response, and the data and objectives available for training. These choices affect both pillars of the survey. Token rate and encoding delay constrain *omni interactivity*, while the preservation of content and temporal structure supports *omni intelligence*, including

prosodic understanding and temporal grounding. Training data and objectives shape both. Representation design is therefore an early source of the intelligence–interactivity trade-off, with implications for subsequent architectural and training choices. Each subsection examines how its mechanisms support the two pillars and what computational, informational, or latency costs they introduce.

We proceed from audio representations in Section 3.1 and streaming video representations in Section 3.2 to representations of time, silence, and overlap in Section 3.3. Section 3.4 then examines how multiple streams are serialized into model input sequences, while Section 3.5 covers the text, tool, and background reasoning streams that accompany ongoing interaction. The last group provides the representations needed to coordinate foreground interaction with background computation in the dual-model pattern discussed in Section 4.

3.1. Audio Representations

An interactive omni model converts continuous microphone input into features or tokens that the language-model backbone can process under a latency constraint. The audio front-end determines the token rate, the paralinguistic information retained, and the minimum encoding delay before the model can respond. Speech representations have been surveyed for spoken dialogue broadly [9, 122] and for discrete audio tokens specifically [97, 191]. We review mel-spectrogram features, self-supervised speech units, and neural codec tokens, then examine semantic and acoustic information, tokenization granularity, and the choice between discrete tokens and continuous features.

Mel features and their discretization. Log-mel spectrograms provide frame-based time-frequency features for audio encoders such as Whisper [222] and the block-wise audio towers of omni models [301]. They retain paralinguistic information and are commonly used for speech understanding. dMel directly discretizes mel filterbank energies into intensity bins, producing a training-free tokenizer that supports ASR and TTS within a single decoder [18]. MELLE instead predicts continuous mel frames with a regression head, providing an alternative to vector quantization [185].

Self-supervised speech units. A second family uses self-supervised learning to obtain representations of linguistic content without transcripts. Contrastive and masked-prediction objectives such as CPC [266], wav2vec 2.0 [16], HuBERT [106], w2v-BERT [52], and WavLM [38] produce frame-level embeddings whose intermediate layers align strongly with phonetic content [190, 289]. Quantizing these embeddings, either with learned codes [15] or with offline k-means as popularized by HuBERT, yields *discrete semantic units*. GSLM demonstrated that a language model trained purely on such units can generate intelligible speech, establishing the textless pipeline of speech encoder, unit language model, and vocoder [142]. Semantic units provide compact representations of linguistic content, but their compression may remove speaker and prosodic information needed for expressive full-duplex conversation.

Neural codec tokens. Neural audio codecs optimize waveform reconstruction at low bitrates, and their discrete code indices can be used as model tokens. SoundStream uses a fully convolutional causal encoder and decoder with residual vector quantization (RVQ), reconstruction and adversarial losses, and quantizer dropout. It compresses 24 kHz audio to 3–18 kbps in a single model and supports real-time streaming [326]. Subsequent codecs refine different parts of this design. EnCodec introduces multi-scale spectrogram discriminators and produces a 75 Hz token stream [60], DAC addresses RVQ codebook underuse with factorized, L_2-normalized codes and supports high-fidelity coding of 44.1 kHz general audio [138], and group-RVQ variants reduce the number of quantizers required [307], while lightweight inverse-STFT decoders accelerate token-to-waveform generation [248].

Because codec indices form a finite vocabulary, they can be modeled autoregressively. AudioLM applies autoregressive modeling to audio tokens [25], and VALL-E formulates zero-shot TTS as codec-token language modeling [39].

Semantic versus acoustic tokens. Semantic tokens emphasize linguistic content and achieve relatively low token rates, but retain less information for waveform reconstruction. Acoustic codec tokens prioritize reconstruction fidelity, although their linguistic structure is often less explicit [97, 122]. AudioLM combines the two in a coarse-to-fine hierarchy, generating semantic tokens for long-range coherence before acoustic tokens for waveform fidelity [25]. Subsequent tokenizers incorporate semantic information directly into codec representations. SpeechTokenizer distills HuBERT representations into the first RVQ level, leaving later levels to encode residual acoustic detail [344]. Mimi uses a split-RVQ design with a WavLM-distilled semantic codebook and parallel acoustic codebooks [61]. X-Codec incorporates SSL features to strengthen the semantic information carried by acoustic codes [316]. FACodec separates content, prosody, and timbre into quantized subspaces [129], while SemantiCodec represents general audio at fewer than 100 tokens per second [169]. For interactive models, these representations must often support both understanding and generation within the same audio-token interface.

Codebook depth, frame rate, and the latency versus quality trade-off. When codebooks are serialized, frame rate and codebook depth determine the token rate, while total sequence length additionally depends on audio duration. Deeper RVQ stacks can improve reconstruction fidelity but increase the number of tokens that the model must process. VALL-E, for example, applies autoregressive prediction to the first codebook and non-autoregressive refinement to seven additional codebooks [39]; the serialization of multiple codebooks is discussed in Section 3.4. Other designs reduce token rates by using fewer codebooks. WavTokenizer represents speech with a single codebook at 40–75 tokens per second [123], while BigCodec achieves 1.04 kbps with a single codebook by increasing codec capacity [300]. TS3-Codec provides streaming single-codebook coding without convolutional layers [294], SNAC distributes codebooks across multiple temporal resolutions [249], and GLM-4-Voice uses a single-codebook tokenizer operating at 12.5 Hz for streaming spoken dialogue [329]. Native interaction also requires causal processing and low encoding latency. Mimi, the codec used by Moshi, provides one example through causal convolutions, a 12.5 Hz frame rate at 1.1 kbps, and approximately 80 ms of codec latency [61]. Moshi uses this codec within a full-duplex system designed to respond within approximately 200 ms.

💡 **Discussion: discrete tokens or continuous features?** The choice between discrete tokens and continuous features remains open. Controlled comparisons report advantages for continuous features on semantic understanding tasks and attribute the gap to quantization loss [272, 273]. Dedicated benchmarks now quantify these gaps along complementary axes: Codec-SUPERB for reconstruction quality [293], DASB for the downstream utility of discrete tokens [192], and STAB for intrinsic tokenizer properties [267]. Discrete tokens support autoregressive speech generation, while continuous output requires an appropriate prediction head [185]. Some omni models therefore use continuous features for understanding and discrete codec tokens for generation [301]. Others tokenize both input and output to model the two speech streams jointly [61]. Continuous input features can preserve semantic and paralinguistic detail, while discrete output tokens support streaming generation. Fully tokenized designs must balance understanding quality with joint speech modeling. Section 4 extends this separation to foreground and background models. How these per-modality choices are arranged across model input and output streams is the subject of Section 3.4.

3.2. Video Representations

Streaming video models process frames as they arrive and may need to respond before the stream ends. This changes the focus from understanding a completed video [31, 258] to determining when visual information becomes available and when it should trigger a response [311]. We review frame tokenization, sampling, temporal position encoding, bounded visual memory, and response triggering. Section 3.4 discusses synchronization with audio and text.

Patch tokens and connectors. A standard video front-end divides each sampled frame into patches, encodes them with a vision transformer, and compresses the resulting tokens before passing them to the language model [66]. A single frame can produce hundreds of patch tokens, so video models use connectors to reduce the number processed by the language model. Cross-attention resamplers map variable-length spatial and temporal features to a fixed number of latent tokens [5, 119, 120]. Query-based bottlenecks such as the Q-Former provide another form of token compression [152]. More aggressive per-frame pooling can represent an entire frame with as few as two tokens, making hour-long video inputs more practical [155]. Multiplying the number of tokens per frame by the number of frames processed per second gives the visual token rate that the backbone must process. This is the video counterpart of the frame-rate and codebook-depth calculation discussed for audio in Section 3.1.

Frame sampling. Uniform sampling is computationally simple but can miss brief events. A common baseline samples approximately one or two frames per second, so events occurring between sampled frames may not be represented. Adaptive keyframe selection prioritizes frames according to query relevance and temporal coverage under a fixed token budget [257]. Streaming systems must make sampling decisions online as frames arrive, which favors lightweight policies that do not depend on future frames. In systems that do not retain raw frames, discarded observations cannot be recovered [311].

Temporal position encoding. Temporal position encoding represents when each observation occurred. Multimodal rotary position embeddings decompose position indices into temporal, height, and width components [278]. VideoRoPE examines which frequency bands should represent time and how temporal indices should be spaced between consecutive frames [288]. TMRoPE aligns temporal positions across audio and video by assigning the same temporal index to video patches and audio frames that share a timestamp [301]. This allows observations from different modalities to be aligned by absolute time. Section 3.3 examines explicit representations of time, silence, and overlap.

Streaming ingestion, memory, and token compression. Streaming video is processed incrementally, so stored context grows unless it is bounded by memory or compression. Attention sinks with contiguous position indices support long streams when training follows the same streaming setting [303], while memory modules condense earlier observations into queryable spatiotemporal states [339]. Compression exploits redundancy across frames. Removing unchanged tokens can eliminate more than 80% of visual tokens with little reported accuracy loss [312], while training-free token merging combines redundant visual tokens [24]. These methods support long-running streams while limiting the stored context. Memory and compression determine which past observations remain available and how much memory and computation are required.

When to respond. Streaming video also requires a decision about when to respond. Methods include per-frame prediction, dedicated response heads, separate perception and response modules, and activation thresholds. VideoLLM-online predicts either silence or the start of a response at each frame [37]. MM-Duet uses a dedicated response head that allows text to be inserted at different points during video playback [284]. Dispider separates perception, decision, and response generation [215], while StreamBridge and LiveStar

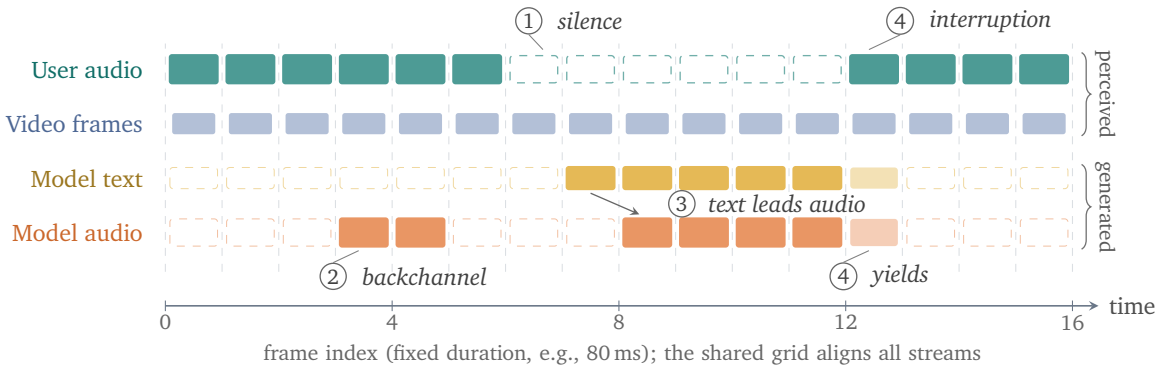

Figure 4: Time-aligned interaction streams. A shared time grid aligns input and output streams. The numbered callouts mark silence, a backchannel, text preceding audio, and interruption-driven yielding. A user interruption stops the model's ongoing speech, and faded blocks indicate generated output that is not played. Section 6 discusses interruption handling.

apply activation or uncertainty thresholds to determine when a response should begin [274, 310]. Early triggering can produce irrelevant responses, while conservative triggering can delay useful information. Section 8 examines this distinction through response-onset latency and time-to-informative-answer.

Discussion: coverage versus timeliness. Video sampling and compression balance information retention against processing cost and response delay. Discarding tokens can remove evidence needed later. Reliable response timing depends on knowing when relevant events occur, making temporal grounding central to proactive interaction. Streaming benchmarks evaluate response timing alongside answer accuracy on timestamp-anchored queries [176, 283], and current systems remain behind human performance on these evaluations [165]. Sections 5 and 6 further examine temporal grounding and response timing. The mechanisms reviewed here determine which visual events are encoded and when their representations become available to the model. Section 3.3 next examines how time, silence, and overlap are represented as model inputs.

3.3. Time, Silence, and Overlap as Model Inputs

Cascaded systems commonly use external speech segmentation and endpoint detection before passing a speaker-attributed transcript to the language model. Native models can instead represent pauses, overlapping speech, and timing within their input streams, as illustrated in Figure 4. Turn-taking has a long literature in conversational systems [34, 251], and full-duplex spoken language models have recently been surveyed [44]. This subsection examines the information required to learn response timing, while Section 6 discusses turn-taking behavior. In what follows, we first ground the discussion in human conversation science. We then trace the shift from reactive endpointing to predictive turn-taking, examine silence and overlap in turn, and review the mechanisms that make models time-aware; we conclude by discussing how temporal behavior is evaluated.

What human conversation implies for machines. Human turn-taking is systematic: turns are constructed from recognizable units, and speaker change becomes relevant at projectable completion points [231]. Across ten languages, response offsets typically fall within a narrow window of 0 to 200 ms, with both long gaps and substantial overlap generally avoided [255]. Because utterance planning takes considerably longer than 200 ms, human listeners must anticipate turn completion from prosodic and linguistic cues and prepare responses before the speaker finishes [148]. Interactive models similarly benefit from retaining these cues and preparing responses before an endpoint is detected.

From reactive endpointing to predictive turn-taking. Cascaded systems commonly reduce turn-taking to endpointing by declaring a turn complete after a fixed silence interval. However, a fixed silence threshold cannot reliably distinguish hesitation from turn completion across all conversational contexts. Continuous prediction provides a more flexible alternative. Recurrent models estimate future speech activity frame by frame [250], multiscale models combine acoustic and lexical cues at their respective rates [228], language models predict turn shifts from pragmatic completeness [69], and generative models learn distributions of response offsets [227]. Voice Activity Projection (VAP) predicts future speech activity for both speakers directly from two-channel audio without requiring turn annotations [70]. Unlike VAD, it predicts upcoming activity instead of detecting current activity and models both speakers jointly. Real-time and multilingual variants extend this approach to additional deployment and language settings [115, 116].

Silence as an interaction state. Acoustically similar silences can serve different interaction functions. A pause within a turn may hold the floor, while a gap between turns may transfer it. The floor-transfer offset measures the interval from the end of one speaker's speech to the start of the other's, with overlap represented by negative values [251]. Speakers can also use fillers such as "uh" to hold the turn during a pause [125]. Systems can use semantic and user-state information to decide whether a pause permits a response [166], a policy considered further in Section 6. In frame-based full-duplex models, silence can be represented as an output state, allowing the model to learn when not to generate speech [61].

Overlap and backchannels as normal input. Overlapping speech is a normal part of conversation. Backchannels such as "yeah" and "mm-hmm" signal engagement without claiming the floor. Models can predict their timing and type by combining acoustic and language-model cues [276] or by adapting VAP for continuous backchannel prediction [118]. During model speech, the system must distinguish user acknowledgments, interruptions, and background speech [44, 166]. This requires representations that distinguish incoming speech from the model's output. Codec-free designs such as SALMONN-Omni explicitly account for the model's own speech during simultaneous listening [325]. dGSLM demonstrates that some overlap and backchannel behavior can be learned from two-channel conversational audio without explicit turn labels [198].

Mechanisms for time awareness. Models represent temporal information in several ways. Temporal position encodings associate tokens with timestamps. TMRoPE, for example, assigns a temporal index to each fixed audio frame and aligns audio and video using absolute timestamps [301], as discussed in Section 3.2. Fixed-duration chunks provide another way to align input and output over time. SyncLLM inserts synchronization tokens at intervals of approximately 160 to 240 ms and predicts upcoming user chunks to compensate for processing delay [268]. Parallel layouts align user and model tokens at corresponding time steps. Moshi represents both speakers as parallel token streams [61], LSLM integrates a streaming listening channel with its speaking channel [180], and NTPP jointly predicts token pairs across the two dialogue channels [279]. Models can also be trained directly for incremental decoding, as demonstrated by Speech ReaLLM [235]. Synchronization may be learned within the model or implemented by an external controller that determines whether the system should continue listening, respond, or stop [44, 281]. Temporal representations support the reasoning about event timing examined in Section 5 and the response-timing decisions discussed in Section 6. Section 3.4 examines how these streams are serialized.

💡 **Discussion: measuring temporal behavior.** Conventional turn-based benchmarks often emphasize response content without evaluating continuous input processing or response timing. Dedicated evaluations provide broader coverage. Full-Duplex-Bench measures pause handling, backchanneling, turn-taking latency, and interruption behavior [160]; later versions add overlap handling [164] and multi-turn interaction with an automated examiner [163]. Talking Turns compares the turn-taking behavior of audio foundation

Table 1: Comparison of stream formats by first-output latency, information retention, training requirements, and representative systems. Teal and orange denote user and model tokens, respectively.

Format	First-output latency	Information retention	Training requirements	Representative systems
Turn-level concatenation	high waits for a full segment or text answer	limited speaker-timing detail; explicit text supervision	low text-LM objective on abundant turn data	SpeechGPT [336], Spectron [194], VoxtLM [183], USDM [133]
Time-aligned interleaving	low–medium chunk granularity plus look-ahead	temporal order kept at alignment granularity; text guidance retained	medium needs forced-aligned or synthetic interleaved data	SUTLM [49], SpiRit-LM [199], GLM-4-Voice [329], IST-LM [308]
Parallel user and model streams	lowest frame-level, full-duplex native	both timelines, incl. overlap and backchannels; lexical fidelity needs an added text stream	high scarce synchronized two-channel dialogue; architectures adapted beyond text LMs	dGSLM [198], Moshi [61], NTPP [279], PSLM [189]

models with conversational corpus statistics [10], while spoken-dialogue surveys summarize measures such as floor-transfer offsets, interruption success, and response delay [122]. Silence is usually studied within turn-taking and full-duplex systems. Evaluations should distinguish response onset from the delivery of useful content, since a fast initial response does not necessarily provide a timely answer. Learning temporal behavior requires representations that retain both speakers' timing, including pauses and overlap. Stream formats that discard either timeline lose this information. Representing both streams at each time step increases token count, including during silence, and requires training data beyond conventional text turns. Section 3.4 examines the resulting trade-offs.

3.4. Unified Stream Formats

The preceding sections reviewed modality-specific streams and their timing. We now examine how these streams are arranged as model inputs and outputs. We group related formats into turn-level concatenation [57, 122], time-aligned interleaving [9], and parallel streams [44], and compare their serialization mechanisms and trade-offs in Table 1.

Turn-level concatenation. Turn-level concatenation appends complete text or speech segments. SpeechGPT sequences recognized text, a textual response, and speech units [336], while Spectron transcribes speech, generates a textual continuation, and then synthesizes it as audio [194]. VoxtLM combines task tokens with text and speech units in a single decoder [183], while USDM continues pretraining on speech-text data before turn-based dialogue fine-tuning [133]. Complete segments allow these systems to reuse standard text-language-model objectives and turn-based datasets. However, waiting for a segment or textual answer to finish delays speech output and can remove detailed information about speaker timing.

Time-aligned interleaving. Time-aligned interleaving orders tokens from different streams within a sequence using timestamps or alignment boundaries. SUTLM compares speech-only, text-only, concatenated, and alternating layouts, and reports the strongest cross-modal transfer with word-level alternation [49]. SpiRit-LM interleaves text and speech units at word boundaries [199], while GLM-4-Voice uses chunk-level interleaving

for streaming synthesis and trains on large-scale synthetic interleaved data [329, 330]. A fixed text-to-speech chunk ratio controls look-ahead and latency, as demonstrated by IST-LM [308]. Interleaved models can be initialized from pretrained text models [99]; recent work also studies their scaling behavior [182] and dynamic turn-level interleaving [58]. These models require aligned training data and an explicit representation of both speakers to support overlap.

Parallel user and model streams. Parallel layouts represent both participants at aligned time steps, thereby preserving the timing relationship between streams described in Section 3.3. dGSLM uses a dual-tower architecture with cross-attention between two conversation channels [198], while Moshi represents the user's speech, the model's speech, and a time-aligned inner-monologue text stream as parallel sequences [61]. SyncLLM combines interleaving with synchronization of fixed-duration chunks from both speakers [268]. NTPP jointly predicts token pairs across two dialogue channels [279], PSLM uses parallel text and speech sequences to reduce latency [189], Mini-Omni decodes text and audio in parallel, with text guiding audio generation [298], and LSLM integrates a streaming listening channel with its speaking channel [180]. Parallel representations also precede LLM-based dialogue, as pGSLM models speech units, pitch, and duration in separate channels [132]. These categories can overlap. OmniFlatten, for example, progressively converts parallel dialogue streams into a single GPT sequence [342]. Parallel-stream models require synchronized two-channel data and architectures adapted beyond conventional text language models [44].

Serialization mechanics. Three serialization mechanisms recur across these formats. Multi-codebook outputs require a defined ordering. MusicGen's delay pattern offsets codebook sequences so that they can be generated in a single pass [54], while stack-and-delay changes their ordering to balance latency and reconstruction quality [145]. Moshi instead combines a temporal Transformer with a smaller depth Transformer that predicts tokens across codebooks [61]. Text and audio sequences generally differ in length, requiring padding or delay policies. These include fixed-length padding in LLaMA-Omni [74], trailing padding in Mini-Omni [298], and a learned delay between Moshi's text and audio streams [61]. Chunk ratio and look-ahead control how much text is available during speech generation [308]. Temporal position encodings can additionally align modalities through absolute time, as illustrated by TMRoPE in Section 3.2 [301].

💡 **Discussion: serialization trade-offs.** The formats impose different buffering requirements, including complete segments, aligned chunks with look-ahead, and synchronized frames. End-to-end latency also depends on implementation and compute. For example, Moshi operates with 80 ms codec frames, while SyncLLM uses synchronized chunks of approximately 160 to 240 ms [61, 268]. Temporal detail depends on alignment granularity, and linguistic consistency can depend on text supervision. Textless parallel models can capture overlap and laughter but may lose semantic information [198]. Text guidance, as in Moshi's inner-monologue stream, can improve linguistic quality within a parallel-stream design [61]. Concatenated formats use conventional turn-based data, interleaved formats require aligned sequences, and parallel formats require synchronized conversational streams [44]. Serialization therefore affects response timing, information retention, and compatibility with pretrained text models. The balance among these factors depends on the model architecture, training data, and implementation.

3.5. Speech Synchronization and User Interruption

Ongoing interaction can involve partial transcription, internal text generation, tool calls, and retrieval. These topics have dedicated surveys [88, 186, 200, 217, 277, 287]. Here, we examine how these streams are synchronized with speech and handled when the user interrupts. Table 2 summarizes the stream types and their synchronization requirements.

Partial transcript streams. Incremental transcripts can change as new audio arrives. Transducer models support streaming recognition through monotonic alignment [95, 211]. Stabilizing partial recognition results balances transcription accuracy against the delay before results are treated as final [236]. In addition to producing transcripts through a separate recognizer, native models can align text directly with an audio stream. Delayed Streams Modeling pairs input audio with a time-aligned delayed transcript [327]. Moshi uses a related alignment mechanism for a different purpose, pairing inner-monologue text with the model's own speech [61]. Systems must decide when a transcript prefix is reliable enough to use for retrieval or tool selection.

Inner monologue: reasoning while listening and speaking. Text generated just ahead of audio can improve linguistic consistency, as demonstrated by Moshi [61]. Some models also generate separate, unspoken reasoning sequences while input or audio playback continues. SHANKS generates chain-of-thought while the user is speaking and reports that more than half of its required tool calls can be issued before the user finishes [46]. Think-While-Listening evaluates text-space reasoning for spoken responses and estimates when a question is sufficiently complete to begin reasoning [243]. STITCH alternates unspoken reasoning chunks with spoken chunks, allowing reasoning to continue during audio playback [47]. Other methods reason in hidden states during the user's turn [292] or use pauses in the model's speech for additional reasoning [67]. These methods perform reasoning in text or hidden states while receiving input or playing speech, allowing computation to overlap with ongoing interaction.

Tool-call and tool-result streams. Tool use adds outgoing call streams and incoming result streams. Text-based agents commonly interleave reasoning, actions, and returned observations [196, 313], while Toolformer and ToolLLM train models to decide when and which tools to invoke [216, 234]. Tool-learning surveys further organize this process into planning, tool selection, invocation, and result integration [217]. Spoken interfaces introduce additional errors. Speech recognition can produce incorrect tool arguments, and evaluations compare the resulting tool-use performance with text-based inputs. VoiceAgentBench reports that cascaded ASR-plus-LLM systems outperform end-to-end speech models on voice-based agentic tasks [121], while conversions of text tool benchmarks into spoken form reveal a consistent modality gap [143]. Full-duplex benchmarks also evaluate multi-step tool use under realistic speech disfluency [162].

Background retrieval and reasoning. Retrieval augmentation grounds responses in external knowledge [149]. Retrieval timing determines when external information becomes available. FLARE triggers retrieval when a look-ahead draft is uncertain [127], while Self-RAG uses reflection tokens to decide when to retrieve or critique generated content [13]. MoshiRAG launches retrieval asynchronously and supplies the returned text to the foreground model while speech interaction continues [48]. This can overlap retrieval with early response generation. Section 8 discusses the resulting distinction between response-onset latency and time-to-informative-answer. Stream RAG predicts a retrieval query while the user is still speaking [11]. Speculative interaction agents issue tool calls early and process input and output asynchronously [103], while streaming retrieval methods estimate when a speculative query is sufficiently stable to execute [83]. Other asynchronous designs allow a voice foreground to continue speaking while a background model reasons [167]. These systems coordinate a fast foreground interaction process with slower background retrieval, tool use, or reasoning. The design can reduce blocking delays, but requires a mechanism for validating and incorporating results during ongoing interaction. Section 4 examines this foreground-background pattern [11, 48, 112, 262].

Interruption: cancel, delay, or roll back. Interruptions require explicit handling of generated but unplayed speech, running tool calls, and pending retrieval. Systems may cancel pending work, retain useful results, or undo reversible effects. ChipChat tracks which generated tokens have been played, allowing the context

Table 2: Text, tool, and reasoning streams: function, temporal relationship to speech, and handling of user interruptions. Text streams (amber) are speech-aligned; tool and background streams (violet) can run asynchronously, a foreground–background split that Section 4 develops. Legend below the table.

Stream	Function	Temporal relationship to speech	On user interruption
Partial transcript	incremental text of the user's speech	follows the user's speech with a short lag	keep continues with the new user speech
Inner monologue	text generated ahead of the model's speech, or unspoken reasoning	runs slightly ahead of the model's speech	discard unspoken remainder discarded; context rolled back to what the user heard
Tool call	request to an external tool	may be issued before the user finishes speaking	discard cancel if pending; deduplicate if already issued
Tool result	result returned from a tool	returns asynchronously after call	hold keep for later use; discard if the request is outdated
Background retrieval	information retrieved asynchronously	runs alongside speech; results return later	hold revise the retrieval query; ignore results of the old query

ongoing speech · text stream · tool / background · speech onset · call issued early

keep retain current work · hold pause pending work · discard stop or ignore pending work

to be rolled back to the output that the user actually heard [159]. Interruptible systems may stop spoken output, steer or abandon ongoing reasoning, and update their behavior after new user input [46, 167], while full-duplex benchmarks evaluate tool use under interruption [162]. Open problems include preventing duplicate calls and managing state after interrupted execution. Section 6 discusses serving-side operations such as stopping playback and rolling back caches, while Section 9 considers cancellation, deduplication, and reversible side effects as broader research problems.

Discussion: speech-aligned and asynchronous streams. Table 2 summarizes these streams by function, timing relative to speech, and interruption handling. Text streams can be aligned with speech, while tools and retrieval complete asynchronously. Their requests and results must be tied to the interaction state and inserted at appropriate points. Partial transcripts and inner monologues follow the speech-aligned formats discussed in Section 3.4. Dual-model architectures discussed in Section 4 separate immediate interaction from slower computation, introducing coordination and cancellation requirements. Failures include premature commitment, delayed reasoning, outdated intent, and inconsistent state after interruption. Representation choices determine which activities can be interrupted and resumed consistently. Section 9 examines the remaining cancellation and state-management problems.

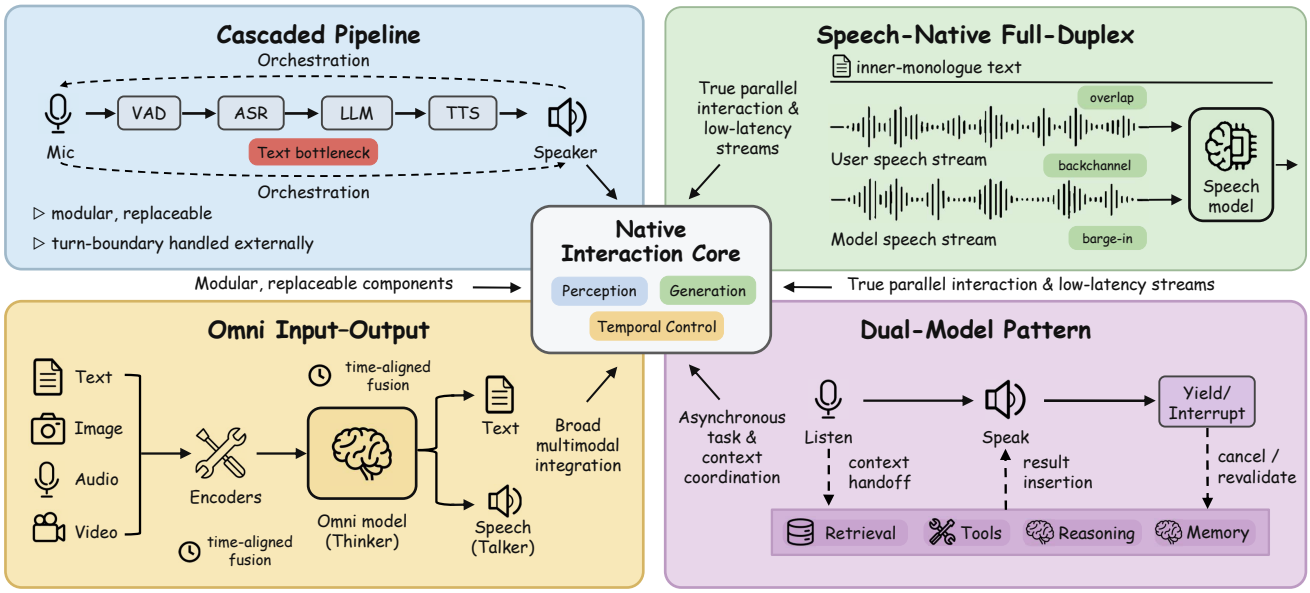

Figure 5: Architecture families for native real-time interaction. The figure maps four architectural patterns by where they place stream fusion, speech generation, temporal control, and background computation.

4. Building Native Omni Models for Real-Time Interaction

Native real-time omni models operate while interaction is still unfolding. User speech may continue after response generation begins, visual evidence may become relevant midway through a response, and results from reasoning or tool use may arrive after the conversational state has changed. The central design problem is to update the interaction state with new multimodal evidence while controlling ongoing output, interruption, and revision. This section examines the problem through architecture, data, and training. We first compare architecture families and fusion mechanisms, with attention to where the live state is maintained and how low-latency interaction is coordinated with slower computation. We then discuss the data and objectives required to learn temporally grounded behavior, including silence, overlap, interruption, and foreground–background coordination.

4.1. Architecture Families and Modality Fusion

We compare architectures by which components observe incoming events, maintain the interaction state, and control response timing. Four families organize the current design space: cascaded baselines, speech-native full-duplex models, omni input-output models, and dual-model interaction systems. Figure 5 shows how these four families implement fusion, speech generation, temporal control, and background computation.

Cascaded baselines. The most widely deployed real-time voice architecture chains independently trained modules: voice activity detection (VAD) or endpointing, streaming ASR, an LLM, incremental TTS, and playback control. Online recognition mechanisms such as RNN-Transducer support streaming input [95], while ASR front-ends such as Whisper provide strong transcription quality for speech interfaces [222]. Incremental TTS can begin synthesis from partial text and reduce audible response delay [178, 232]. The strength of this design is modularity: each component has a defined interface, can be improved independently, and can be monitored or replaced without retraining the whole interaction system.

These interfaces also determine where the live interaction state can reside. At the ASR-to-LLM boundary, converting audio into text hypotheses and metadata can remove prosody, hesitation, laughter, overlap, and backchannel timing before they reach the language model. At the LLM-to-TTS boundary, generated text becomes an audio queue whose timing is further shaped by synthesis and playback buffering. Interruption is usually handled by runtime logic that stops playback, discards queued speech, and cancels or restarts generation [44, 247]. Cascades can therefore be responsive when their components are carefully scheduled, but turn-taking and overlap remain forms of engineered synchronization outside the model [44]. Work such as Voice Activity Projection shows that conversational timing can instead be learned through predictions of future voice activity [70], while full-duplex benchmarks directly evaluate pause handling, backchanneling, turn-taking, and interruption management [160]. A useful middle ground consolidates the front end itself: UAF combines several front-end functions within a single model, including VAD, turn detection, speaker recognition, ASR, question answering, and interruption-related control signals, predicting them within a single streaming autoregressive model while retaining a separate generation module [154]. Native architectures can learn these functions within shared stream representations and interaction state.

Speech-native full-duplex architectures. Speech-native full-duplex models represent user speech and model speech as time-aligned streams that can be active during the same interval [61, 198, 268]. Silence, pauses, overlap, backchannels, and interruptions therefore become temporal conditions visible to the model. Their state can update from incoming audio while speech generation continues, allowing the model to condition turn-taking and interruption responses on both streams. Interaction events can consequently be represented as stream events: a backchannel is aligned with ongoing user speech, an interruption occurs while the model-speech stream is active, and yielding changes the model-speech stream in response to the user.

dGSLM provides an early foundation by modeling two-channel dialogue directly from conversational audio, showing that turn-taking can be learned from paired speech streams without first reducing the exchange to text [198]. Moshi jointly models user speech, model speech, and an inner-monologue text stream over neural codec tokens [61], while SyncLLM uses fixed-duration dialogue chunks [268]. Freeze-Omni preserves a text-capable LLM backbone and adds streaming speech modules with chunk-level state prediction [281]. SALMONN-omni uses continuous speech embeddings without codec tokens [324]. Mini-Omni and Mini-Omni2 provide smaller open-source implementations with speech encoders, speech-token decoders, delayed parallel generation, and lightweight interruption signals [298, 299]. BayLing-Duplex shows that full-duplex interaction control need not require a redesigned representation: it is learned within an otherwise standard autoregressive speech–language model using only a small set of control tokens, so that a single model decides when to listen, speak, and stop without a separate turn-taking module while retaining the backbone's response capabilities [76]. These designs differ in stream coupling, in the use of text to guide speech generation, and in the choice of speech representation: discrete units, codec tokens, or continuous embeddings. Their main limitation is modality breadth: most remain centered on audio, while video, screen state, tools, memory, and background reasoning are external or absent.

Omni input-output architectures. Omni input-output systems must coordinate text, visual, and audio information arriving at different rates, extending the temporal problem from speech-centered interaction to images, video, screen context, and spoken response [220, 301, 302]. Audio is processed in chunks, video as frames or visual tokens, and text as instructions or intermediate representations. A visual event may change when the model should speak, while an acoustic cue may determine whether it should continue or pause. A central design challenge is to align these inputs while keeping speech generation responsive to new evidence [112, 301].

The Qwen-Omni family illustrates how this architecture has evolved. Qwen2-VL introduced dynamic-resolution visual processing and multimodal RoPE [278]. Qwen2.5-Omni extended this approach with

block-wise audio and visual encoders, time-aligned multimodal positional encoding, and a Thinker–Talker architecture for streaming text and speech generation [301]. The Thinker produces text-oriented hidden states that condition the Talker, allowing speech to reflect the current multimodal state before a complete text response has been emitted. Qwen3-Omni replaces fixed two-second audiovisual blocks with temporal IDs anchored to absolute time, directly aligning audio and video representations and supporting streaming inputs of arbitrary duration [302]. Qwen3.5-Omni retains the Thinker–Talker structure while introducing Hybrid MoE backbones for long audio-video sequences and ARIA, which adaptively interleaves text and speech units during streaming generation [220].

Other systems use different interfaces between multimodal perception, language modeling, and speech generation. LLaMA-Omni and LLaMA-Omni2 connect speech encoders and LLM backbones to streaming speech decoders [74, 75], while VITA, AnyGPT, and MiniMind-O explore broader or smaller-scale routes to multimodal input and spoken output [80, 91, 333]. These systems vary in modality coverage, scale, and how speech decoders receive the model's current representations. Talker modules, hidden-state handoffs, codec-token paths, and streaming speech decoders provide different degrees of access to the live model state, and their interfaces affect how quickly new visual or acoustic information can influence generated speech, with corresponding costs in alignment, latency, and cancellation. Recent work also begins to extend this output path beyond text and speech: Ex-Omni-2D generates coordinated text, personalized speech, and reference-conditioned streaming video, using speech units as a shared acoustic–temporal interface between the dialogue and visual-generation pathways [341]. These choices form part of the broader fusion and coupling problem discussed next.

Dual-model interaction systems. Dual-model interaction systems separate immediate interaction from slower computation [112, 262]. A foreground model processes ongoing input and controls speech, pauses, acknowledgments, and interruptions, keeping the user-facing state coherent. A background model handles longer reasoning, retrieval, browsing, tool use, planning, or memory operations without blocking the live exchange. Thinking Machines Lab [262] provides a direct formulation of this pattern. Its time-aware interaction model processes continuous audio, video, and text as micro-turns, while an asynchronous background model handles longer-horizon reasoning and tool use. The foreground remains active as this work proceeds and decides when returned results can enter the conversation.

DuplexOmni provides a recent research implementation of the same separation. Its interaction layer processes streaming audio and video while generating text and speech in real time, whereas a replaceable reasoning component returns reasoning and tool-use results asynchronously [112]. Other systems use specialized background components. MoshiRAG combines a compact full-duplex speech model with selective asynchronous retrieval, allowing response onset to precede the arrival of external evidence [48]. Stream RAG predicts tool queries during ongoing user speech and later fuses retrieved text into spoken responses [11]. STITCH interleaves unspoken reasoning chunks with spoken chunks so that reasoning can continue during speech playback [47], while EgoMem adds asynchronous retrieval and memory management to real-time omnimodal dialogue [314]. StreamArena introduces StreamMind, a streaming architecture that separates latency-critical interaction from persistent multimodal recall and longer-horizon processing over hour-scale video [343]. These systems do not all use two independently scaled models, but they share the same temporal separation between immediate interaction and slower computation.

The defining challenge is the interface between the live and background processes, as illustrated in Figure 6. The foreground must decide when to delegate, what context to send, and whether returned results should be inserted, revised, delayed, or discarded as the live state changes, and must validate them against the current interaction state before incorporating them. Starting audio quickly does not necessarily reduce the

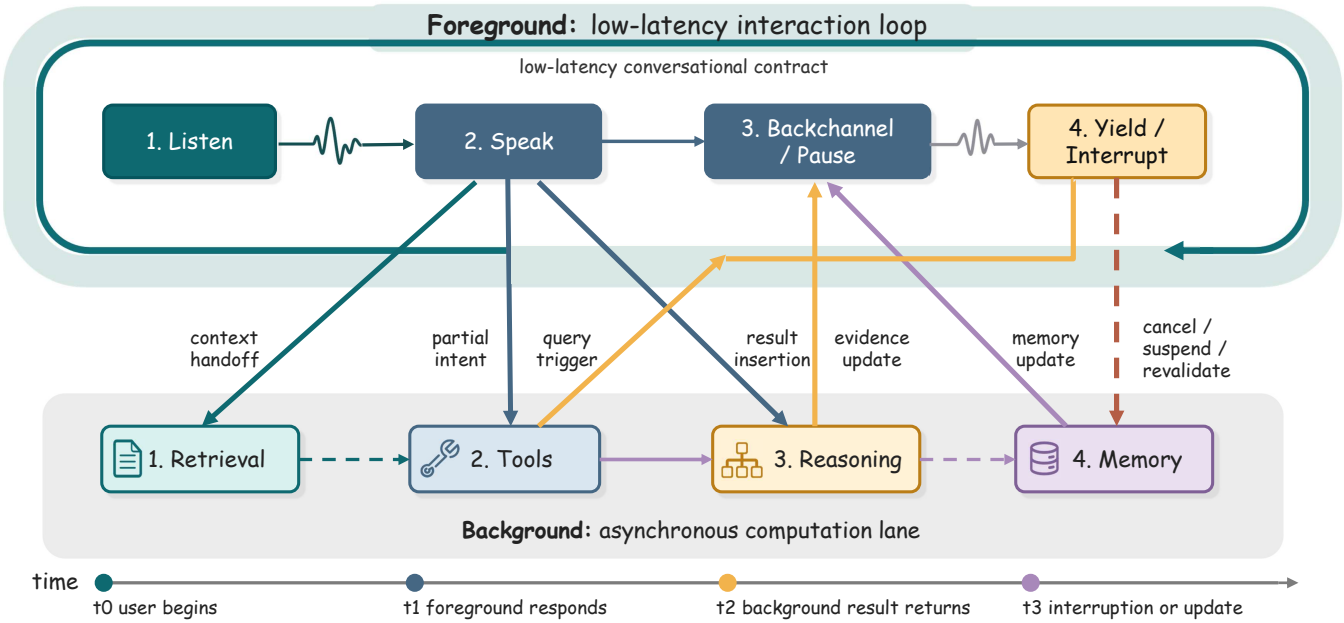

Figure 6: Foreground–background coordination in a dual-model interaction system. A foreground process handles ongoing interaction while retrieval, tools, reasoning, and memory operations run asynchronously. Returned results are incorporated, deferred, revalidated, or discarded according to the current state.

time until the user receives a useful answer, so explicit policies for insertion, validation, cancellation, and recovery are required.

Modality Fusion and Coupling Approaches. Fusion describes how modality representations are combined, whereas coupling describes how strongly and continuously streams influence one another's content and timing. These design choices can be combined in different architectures. A central question is how much temporal and modality-specific structure remains visible across architectural boundaries.

Encoder-based fusion. Encoder-based fusion converts speech, audio, images, or video into learned features before projecting them into an LLM or multimodal backbone. Pretrained encoders such as Whisper, CLIP, and SigLIP provide strong semantic representations [221, 222, 332]. LLaMA-Omni, the Mini-Omni series, and Qwen2.5-Omni follow this pattern [74, 298, 299, 301]. Its main advantage is stable and sample-efficient training. The encoder determines which temporal details are retained and how frequently the model state can update, from once per turn to once per chunk or frame: turn-level features suppress within-turn timing, whereas chunk- or stream-level features preserve more local evidence at greater alignment and memory cost. Separate modality encoders can also process audio and video at mismatched rates. OmniEncoder jointly encodes audio and video at 25 frames per second [17].

Discrete-token and codec-token fusion. Discrete-token fusion converts continuous signals into token sequences or parallel codebook streams. Neural codecs such as SoundStream and EnCodec provide this interface for speech [60, 326]; Moshi uses codec tokens for parallel user and model speech streams [61], while AnyGPT extends discrete-token modeling across several modalities [333]. A common token interface makes autoregressive perception and generation easier to combine, but high token rates and multiple codebooks increase decoding and memory costs as well as first-audio latency. Time-aligned user and model streams can represent silence, overlap, interruption, and yielding within the same representation [61, 198]. Codec-token

language modeling is not the only route to overlapping conversational audio: Alignment-Free Text-Audiobox instead uses a latent diffusion formulation with flow matching and learns text–speech alignment through cross attention rather than forced alignment, supporting long-form full-duplex dialogue synthesis with turn-taking, backchannels, and emotional dynamics [40].

Early or encoder-free fusion. Early fusion uses lightweight projections of lower-level inputs and jointly trains the backbone to learn their perceptual representations, preserving local acoustic structure, visual change, and fine timing. Native multimodal scaling and encoder-free vision-language studies suggest that this approach can remain competitive when supported by sufficient data and joint training [64, 244]. Thinking Machines Lab's interaction model provides a real-time example by feeding dMel audio features and patch-level image features through lightweight embeddings into a jointly trained transformer [262], allowing incoming observations to update the model state more directly, while requiring more extensive joint training.

Intermediate fusion and expert routing. Intermediate fusion can operate on encoder features or discrete tokens, retaining separate computation for each modality while allowing information exchange through shared attention, cross attention, adapters, or shared experts. JavisDiT illustrates this pattern in joint audiovisual generation: audio and video tokens interact through shared self-attention before passing through separate FFNs for each modality [171]. For speech and text, MoST combines dedicated experts with shared experts to support information transfer across modalities [175]. DeepOmni assigns experts by their routing statistics: for each MoE layer it measures the fraction of speech tokens and of text tokens routed to each expert, designates as speech experts those with high speech load and low text load, and keeps the remainder as text experts, alongside a small set of shared experts that both modalities use [237]. Cross-modal exchange occurs through these shared experts and through the attention layers, which all tokens pass through regardless of modality. Training proceeds in stages: the router is frozen while speech and text experts are trained separately on their own modality, then unfrozen for joint training on cross-modal data, which limits the loss of language capability during speech adaptation. These mechanisms preserve modality specialization and support interaction within the backbone, but introduce choices about which layers exchange information, routing balance, and temporal synchronization; related routing mechanisms have also been applied to social evidence in omni models [239].

Coupling strength and speech-output paths. Loose coupling transfers completed segments, such as ASR text, visual summaries, or tool results. Block-level coupling updates the state as chunks arrive. Shared-temporal-grid coupling aligns multiple active streams that arrive at different rates, such as audio chunks every few tens of milliseconds and video frames at a lower rate, on one time grid, preserving overlap and within-turn events. Qwen3-Omni aligns audio and video through temporal IDs anchored to absolute time [302], while Thinking Machines uses interleaved 200-ms input-output micro-turns [262]. ROMA instead groups frequent audio chunks and less frequent video observations into synchronized multimodal units and adds a speak head that predicts response timing separately from linguistic content [263]. These mechanisms increase temporal control but also increase synchronization, memory, and scheduling costs.

Speech decoders can condition on generated text, hidden states, or codec tokens. Text-to-TTS paths provide modularity but pass through a completed or partial text boundary, whereas hidden-state handoffs, Talker modules, codec-token paths, and streaming speech decoders connect speech more directly to the active model state. Qwen3.5-Omni's ARIA further adapts the interleaving rate between text and speech units during streaming generation [220]. More frequent updates can let new evidence influence speech sooner, while requiring explicit handling of already generated or queued audio after interruption.

Table 3: Fusion and coupling mechanisms for native real-time omni models.

Mechanism	State interface	Temporal unit	Main benefit	Main risk
Encoder features	Pretrained or unified encoder output	Segment or chunk	Stable, sample-efficient training	Loss of fine temporal detail
Discrete/codec tokens	Token or codebook streams	Frame or token	Unified autoregressive interface	High token and decoding load
Early/encoder-free	Lightweight input projection	Fine-grained signal	Preserves lower-level input detail	Requires extensive joint training
Intermediate/expert-routed	Shared layers or routed experts in the backbone	Layer or token	Specialization with cross-modal exchange	Poor layer placement or routing imbalance
Shared temporal grid	Aligned active streams	Fixed time step	Overlap and event-level control	Misalignment across streams and memory cost
Speech-output coupling	Text, hidden-state, or token handoff	During generation	Speech conditioned on current state	Possible inconsistency and difficult cancellation
Dual-model coupling	Context and result channel	Asynchronous return	Fast onset with slower reasoning	Stale or contradictory results

Dual-model coupling. Dual-model coupling defines how the foreground sends context to background processes and incorporates returned results. Context leaves the foreground through a request or context package, while reasoning, retrieval, tool, or memory results return through prompt updates, tool-result streams, or other conditioning inputs. Because the live conversation may change before a result returns, the interface must support insertion, validation, cancellation, and revalidation [112, 262].

These choices affect temporal detail, modality specialization, training requirements, decoding cost, routing stability, and coordination. Table 3 compares their interfaces, timing, benefits, and risks.

4.2. Data

Training data determine which interaction behaviors and representations a model can learn. General speech and audiovisual corpora provide abundant supervision for perception and generation [94, 187, 207, 328], but native interaction requires temporal supervision such as separate speaker tracks, overlap and silence, event timestamps, response windows, interruption outcomes, and asynchronous execution traces. Figure 7 summarizes this hierarchy, while Table 4 reports representative datasets and training-data pipelines.

Speaker-separated dialogue data. Speaker-separated dialogue provides the most direct supervision for turn-taking, overlap, interruption, backchannels, and conversational silence. Fisher is a widely used example, with more than 16,000 conversations and roughly 2,000 hours of two-channel telephone audio [53]. This dataset has been used by dGSLM, Moshi, and SyncLLM to learn multistream conversational timing [61, 198, 268]. Switchboard and CallHome provide related telephone conversations [90, 168], while the Open-source Full-Duplex Dataset, DialoSpeech, and HumDial add human-recorded or large-scale dual-track dialogue, interaction annotations, and interruption and overlap data [271, 297, 349]. DuplexChat increases the available data by automatically extracting and restoring two-speaker tracks from public podcasts, reporting approximately 282,634 hours of English and 132,723 hours of Japanese dialogue [197]. Despite this increase in scale, the category retains important limitations. Classic corpora often contain narrowband, two-person telephone conversations, primarily in English, while automatically recovered tracks are not equivalent to directly recorded multichannel interaction. High-fidelity multilingual and multiparty recordings, together with fine-grained annotations of addressee, interruption intent, backchannel function, and floor negotiation, remain scarce.

Table 4: Representative datasets and training mixtures for pretraining, instruction tuning, streaming interaction, synthetic dialogue, and preference optimization in omni models.

Dataset	Type	Modality	Language	Size	Date
Fisher [53]	Two-channel telephone speech	Audio, Text	English	1.96k h	2004
Switchboard [90]	Two-channel telephone speech	Audio, Text	English	0.26k h	1992
CallHome [168]	Multilingual two-channel calls	Audio, Text	6 languages	100–120 calls/lang.	1996/1997
DialoSpeech [297]	Dual-track spoken dialogue	Audio, Text	EN, ZH	10k h	2025
Full-Duplex Datasets [349]	Human-recorded dual-track dialogue	Audio, Text	EN, ZH	15 h	2025
HumDial [271]	Human-recorded dual-channel dialogue	Audio, Text	EN, ZH	>100 h	2026
DuplexChat [197]	Podcast-derived dual-track dialogue	Audio	EN, JA	282.6k/132.7k h	2026
LibriSpeech [207]	Read audiobook speech	Audio, Text	English	1.0k h	2015
Libri-Light [130]	Unlabeled audiobook speech	Audio	English	60k h	2020
LibriHeavy [131]	Audiobook speech with punctuation	Audio, Text	English	50k h	2024
LibriTTS [328]	Multi-speaker audiobook speech	Audio, Text	English	0.585k h	2019
LibriTTS-R [137]	Quality-restored audiobook speech	Audio, Text	English	0.585k h	2023
GigaSpeech [36]	Audiobook/podcast/YouTube speech	Audio, Text	English	10k h	2021
GigaSpeech 2 [309]	Multilingual low-resource speech	Audio, Text	TH, ID, VI	30k h	2025
The People's Speech [85]	Web speech	Audio, Text	English	30k h	2021
MLS [212]	Multilingual audiobooks	Audio, Text	8 languages	50k h	2020
Common Voice [8]	Crowdsourced read speech	Audio, Text	Multilingual	Variable	2020
VoxPopuli [269]	Parliamentary speech recordings	Audio, Text	23 languages	400k h unlab.; 1.8k h trans.	2021
YODAS [153]	YouTube speech with transcripts	Audio, Text	100+ languages	500k h	2024
Emilia [100]	In-the-wild speech recordings	Audio, Text	6 languages	101k h	2024
WenetSpeech [335]	YouTube/podcast speech	Audio, Text	Mandarin	10k h	2022
WenetSpeech4TTS [177]	WenetSpeech-derived TTS corpus	Audio, Text	Mandarin	12.8k h	2024
Parler-TTS [141]	Expressive TTS speech	Audio, Text	English	45k h	2024
Seed-TTS subset [7]	Five-speaker adaptation speech	Audio, Text	–	20 h	2024
Fish-Speech data† [158]	Multilingual TTS speech	Audio, Text	8 languages	720k h	2024
Moshi pretrain† [61]	Internal audio pretraining mixture	Audio	Mostly English	~7M h	2024
VoiceAssistant-400K [298]	TTS-rendered GPT-4o dialogues	Audio, Text	English	400k	2024
InstructS2S-200K [74]	TTS-synthesized instructions and responses	Audio, Text	English	200k	2025
SpeechInstruct [336]	Cross-modal instruction data	Audio, Text	English	38k	2023
AnyInstruct-108k [333]	GPT-4-generated multimodal dialogues	Audio, Text, Image, Music	English	108k	2024
UltraChat [65]	LLM-generated text dialogues	Text	English	1.5M	2023
Alpaca [260]	GPT-3.5-generated instructions	Text	English	52k	2023
WavCaps [184]	Audio-caption pairs	Audio, Text	English	403k	2024
Ego4D [94]	Egocentric video with audio	Video, Audio, Text	Multilingual	3.7k h	2022
HowTo100M [187]	YouTube instructional videos	Video, Audio, Text	English	136M clips	2019
AVSpeech [73]	Clean audiovisual speech	Video, Audio	Multilingual	~4.7k h	2018
VoxCeleb [195]	Celebrity interview video	Video, Audio	Multilingual	1k+ spk	2017
VoxCeleb2 [51]	Audiovisual speaker recordings	Video, Audio	Multilingual	1M+ uttr	2018
Perception-Test [213]	Multimodal reasoning video	Video, Audio, Text	English annotations	11.6k clips	2023
AudioSet-Strong [101]	Temporally annotated sound events	Audio, Text	N/A	67k clips	2021
Qwen3-Omni pretrain† [302]	Internal omni pretraining mixture	Audio, Video, Image, Text	Multilingual	~2T tok.; 20M h audio	2025
Qwen3.5-Omni pretrain† [220]	Internal omni pretraining mixture	Audio, Video, Image, Text	Multilingual	>100M h audio-video	2026
ROMA streaming data† [263]	Proactive responses, narration, reactive QA	Video, Audio, Text	English	676k samples	2026
Moshi synthetic† [61]	Internal synthetic multistream dialogue	Audio	English	>20k h	2024
SyncLLM synthetic [268]	Bark-rendered LLM dialogues	Audio	English	212k h	2024
OmniFlatten synthetic [342]	CosyVoice-rendered text dialogues	Audio	ZH, EN	~2k h	2025
LLaMA-Omni/2 synthetic [74, 75]	TTS-synthesized InstructS2S dialogues	Audio, Text	English	200k dialogues	2024/2025
OpenHermes-2.5 [261]	Instruction and chat data	Text	English	1M samples	2023
CosyVoice 2 data† [68]	Internal multilingual TTS mixture	Audio, Text	ZH, EN, JA, KO	166.8k h	2024
DuplexOmni† [112]	Writer–Director interaction traces	Video, Audio, Text	ZH, EN	3.02M conv.; 620k seeds	2026
DuplexGen† [136]	Preference-calibrated synthetic dialogue	Audio, Text	English	Not reported	2026
Qwen2.5-Omni DPO† [301]	WER-/pause-ranked speech candidates	Audio, Text	–	Not reported	2025
Qwen2-Audio DPO† [50]	Human-ranked audio responses	Audio, Text	–	Not reported	2024
Align-SLM [161]	AI-feedback preference pairs	Audio	English	~263k pairs	2025
SpeechAlign [337]	Reference/synthetic codec preferences	Audio	English	50k pairs; 250k ablation	2024
ASPIRin [105]	Speak/silence GRPO actions	Audio	English	43 h	2026
Multi-Faceted [201]	RL: pause/turn/backchannel/interruption	Audio	English	≤2k seg./axis/corpus	2026
RLAIF-SDS [12]	Switchboard DPO preferences	Audio, Text	English	166k pairs	2026

† Internal training mixtures or constructed pipelines reported in model papers; sizes follow those papers. "Not reported" means that the total scale is undisclosed, and "–" means that the language composition is unspecified, and "N/A" means that language is not applicable to the dataset.

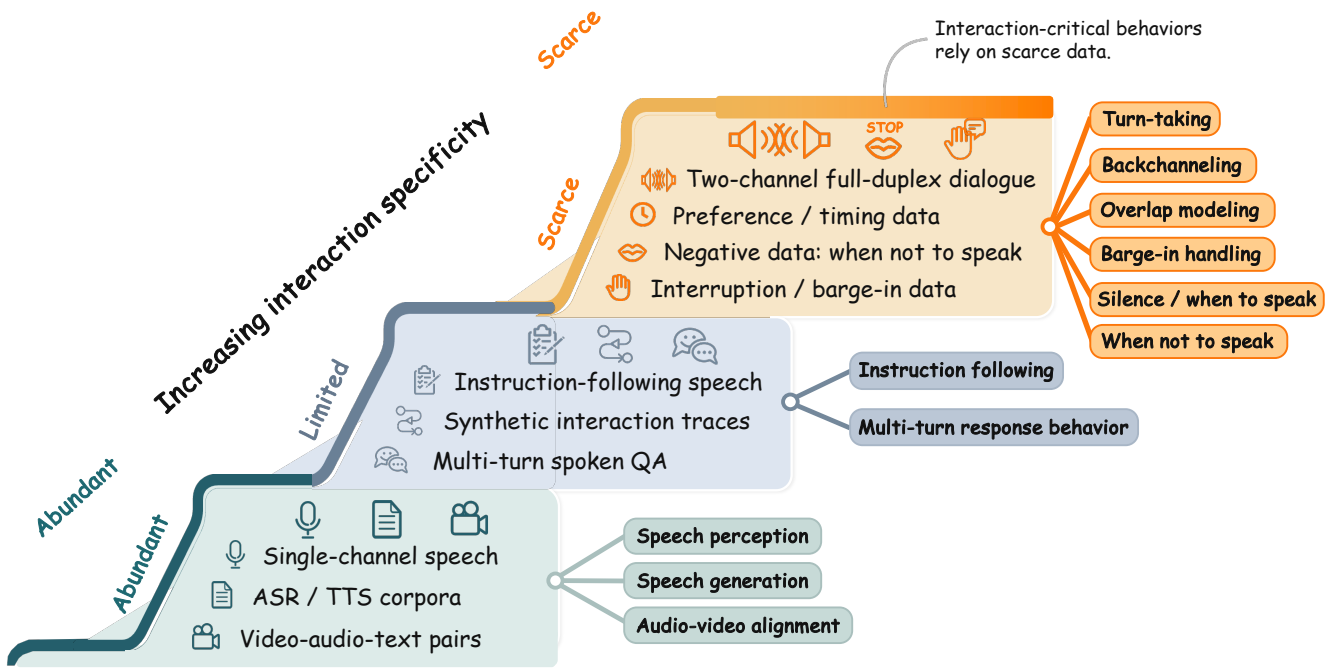

Figure 7: Training data for building native real-time omni models. Temporally annotated data for specialized interaction behaviors is less available than general speech data. These behaviors include turn-taking, acknowledgment, interruption handling, silence, proactive responses, and foreground–background coordination, all of which require temporally aligned interaction traces.

Foundational speech and spoken-instruction data. Single-channel speech supports training for ASR, TTS, audio codecs, and speech representation learning. Representative resources include LibriSpeech [207], Libri-Light [130], LibriHeavy [131], LibriTTS [328], GigaSpeech [36], Common Voice [8], VoxPopuli [269], YODAS [153], Emilia [100], and WenetSpeech [335]. At a larger scale, Moshi reports pretraining its audio stream on approximately seven million hours of mostly English speech transcribed with Whisper [61, 222]. These corpora support perception and generation but generally do not preserve separate speaker timelines or explicit turn-control annotations. Spoken-instruction datasets connect these capabilities to assistant behavior. VoiceAssistant-400K [298], InstructS2S-200K [74], SpeechInstruct [336], AnyInstruct-108K [333], and LLaMA-Omni2's multi-turn speech data [75] are largely produced by generating or rewriting text instructions and rendering them with TTS. These datasets support training for spoken instruction following and conversational response generation, but their clean request–response structure provides little direct supervision for overlap, interruption, or deciding not to speak.

Temporally aligned audiovisual data. Omni models require audio and visual streams whose temporal relationship is preserved. Ego4D [94] supplies egocentric video with audio and annotations, HowTo100M [187] provides narrated instructional video at web scale, and AVSpeech [73] and VoxCeleb [51, 195] associate visible speakers with speech. These resources support audiovisual grounding but often rely on clip-level captions or weak temporal alignment. Qwen2.5-Omni uses temporally interleaved audio-video training for synchronized multimodal processing [301], while ROMA curates streaming data for reactive and proactive audiovisual interaction [263]. Nevertheless, these model-specific mixtures do not resolve the broader data gap. Native interaction requires event-level supervision of what occurred, when it occurred, and whether it should trigger, delay, or suppress a response, while large-scale natural audiovisual conversations with such annotations remain scarce.

Synthetic interaction traces. Synthetic data supplies interaction patterns that are difficult to collect naturally at scale. A common pipeline generates or rewrites dialogue with a text model and renders it as speaker-separated speech with controlled silence, overlap, backchannels, speaker identity, and noise. Moshi reports more than 20,000 hours of synthetic multistream speech [61], while SyncLLM progressively introduces pauses, overlapping speech, and natural Fisher conversations during training [268]. OmniFlatten and LLaMA-Omni similarly synthesize multi-turn spoken dialogues through LLM rewriting and TTS [74, 75, 342].

DuplexOmni extends this approach with a Writer–Director pipeline for continuous multimodal interaction data [112]. The scarcity of synchronized conversational data is compounded by the fact that appropriate turn-taking behavior is scenario-dependent. DuplexGen calibrates synthetic human–AI dialogues using a small amount of human turn-taking preference supervision, producing task-specific timing patterns rather than one universal interaction distribution [136]. The resulting overlap and interruption patterns differ between cooperative and competitive settings, showing why training data should include the context in which a timing behavior is appropriate. More generally, synthetic traces provide scale and control over interaction patterns, but may not reproduce natural disfluency, prosodic coordination, interruption intent, or human timing distributions. Systems such as Moshi and SyncLLM therefore continue to mix synthetic traces with genuine speaker-separated recordings when available [61, 268].

Interaction-policy and safety data. Interaction-policy data supervises when the model should speak, wait, acknowledge, yield, or stop. This layer is smaller and less standardized than the preceding ones. Existing resources include ranked speech candidates for quality and pause control [301], preference pairs for audio-conditioned text responses [50], preference pairs for speech generation [12, 161, 337], and interaction-specific segments or trajectories for optimizing turn-taking, backchanneling, and interruption [105, 201].

Silence is also an explicit positive training target. Moshi, for example, continuously predicts both speech and silence [61]. Training examples should include appropriate silence, brief acknowledgment, yielding after interruption, and abandoning an outdated response, as well as cases where a response must be stopped or corrected when unsafe intent becomes apparent. Safety data must additionally cover streaming refusal, bystander speech, and privacy-sensitive audiovisual input. Most interaction-policy and streaming-safety datasets remain system-specific or unreleased.

Dual-model traces. Dual-model systems require execution traces that go beyond ordinary dialogue pairs. A trace must record when the foreground delegates, what context is sent, when background reasoning or tool results return, what has already been spoken, and whether the result is inserted, revised, delayed, or discarded. MoshiRAG and Stream RAG provide examples of such coordination traces for asynchronous retrieval and tool use [11, 48], while STITCH and EgoMem extend them to background reasoning and long-term memory [47, 314]. The scarcity of public traces limits the reproducibility of training for interruption, cancellation, stale-result detection, and recovery.

Cross-cutting data challenges. Across these categories, three limitations remain. First, speech datasets are growing, but high-quality multilingual and multiparty interaction remains scarce relative to single-channel speech. Second, timestamp and speaker-attribution errors directly corrupt supervision for silence, overlap, interruption, and audiovisual response timing. Third, current data is dominated by English and a small number of conversational settings, leaving other languages, accents, and social norms underrepresented; data should also better represent bystander speech and how participants coordinate turns.

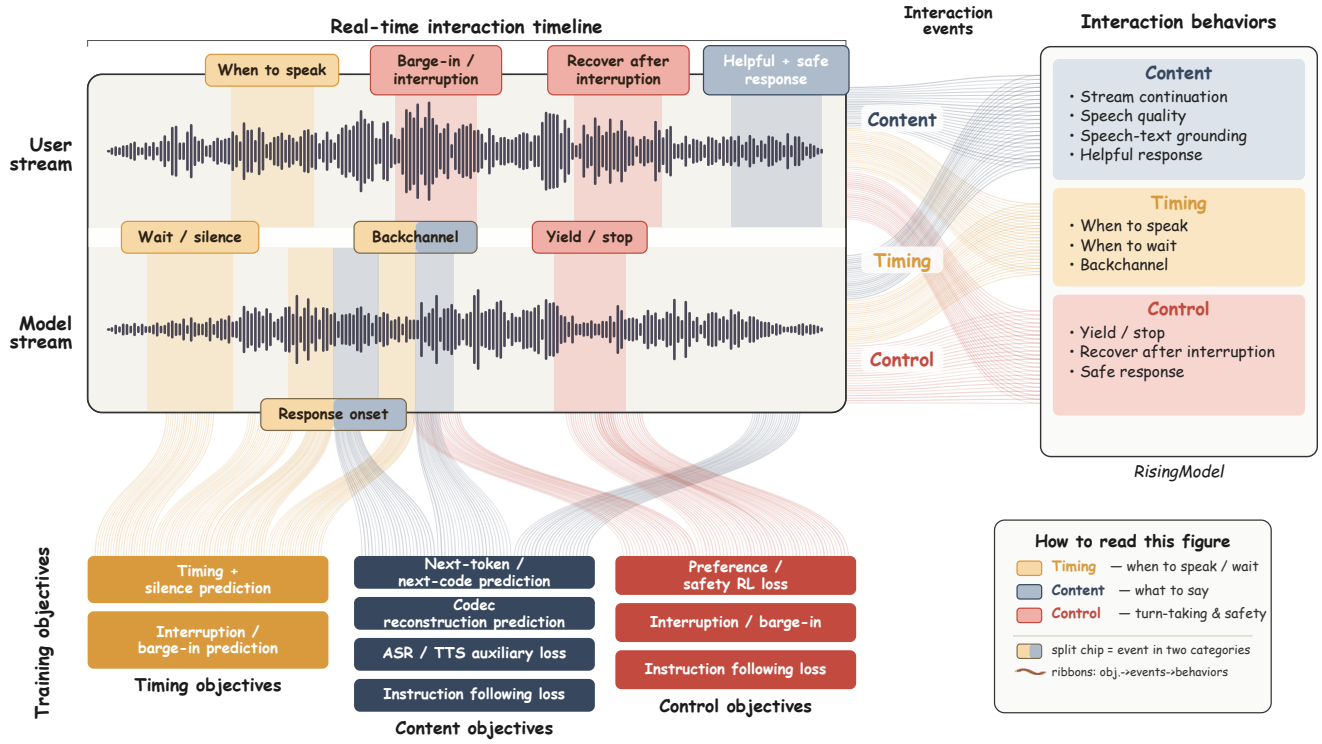

Figure 8: Training objectives for native real-time omni interaction. Objectives for content and cross-modal grounding, temporal control, interaction-policy alignment, and concurrent reasoning and coordination, together with safety objectives, supervise response onset, silence, acknowledgment, interruption handling, yielding, recovery, and safe response.

4.3. Training Objectives

Training objectives specify how models learn multimodal representations, response generation, and interaction decisions. Native real-time omni models must generate coherent text and speech, align evidence across modalities, control when to speak or remain silent, respond to interruption, and coordinate computation that may finish after the conversational state has changed. We group these objectives into four categories: content and cross-modal grounding, temporal control, interaction-policy alignment, and concurrent reasoning and coordination. Safety and balancing between objectives apply across these categories. Figure 8 maps these objective families onto events in a live user–model interaction.

Content and cross-modal grounding. Autoregressive likelihood remains the core objective, with targets drawn from text tokens, semantic speech units, codec tokens, acoustic tokens, and interaction-control symbols. Multi-codebook speech requires prediction across both time and codebook depth. Moshi factorizes these dimensions with temporal and depth Transformers [61], while Mini-Omni uses delayed parallel codebook prediction [249, 298]. Text representations can condition incremental speech generation, as in text planning streams and Thinker–Talker handoffs [61, 301, 329].

Reconstruction and speech–text alignment losses support this objective without defining the interaction policy themselves. Codec reconstruction losses train waveform fidelity so that predicted speech units are acoustically usable [60, 326], while auxiliary CTC losses align acoustic and linguistic representations in systems such as Freeze-Omni, VITA 1.5, and LLaMA-Omni [74, 81, 281]. These losses establish perception and generation capability; temporal and policy objectives determine how that capability is used during interaction.

Temporal control. Temporal objectives supervise when the model should speak, continue, wait, acknowledge, yield, or stop. dGSLM separates content and duration prediction, using cross-entropy for speech units and an L1 loss for their durations [198]. SyncLLM predicts fixed-duration chunks, including silent chunks, while Moshi continuously predicts speech or silence on aligned streams [61, 268]. These formulations make pauses, overlap, and response onset part of the prediction target.

Other systems supervise interaction state more explicitly. Freeze-Omni uses chunk-level state classification for continued listening, interruption, and response regeneration, while SALMONN-omni combines speech-boundary tokens with a weighted loss over ambiguous intervals during which the model should delay responding [281, 324]. For audiovisual streams, ROMA applies a weighted binary cross-entropy objective to a speak head at each multimodal time step, using event windows or segment boundaries as response-timing labels [263]. These objectives explicitly train response timing and interruption handling, moving them from runtime heuristics into supervised model decisions.

Interaction-policy alignment. Supervised likelihood encourages imitation of observed interaction sequences, but it does not directly optimize whether a pause, acknowledgment, or interruption response is well timed. Preference optimization first addressed speech quality and response behavior. Qwen2.5-Omni applies DPO to speech candidates ranked by word-error and pause-error measures [301], while Qwen2-Audio, Align-SLM, and SpeechAlign use preference signals for factuality, generation quality, or speech alignment [50, 161, 223, 238, 337].

Recent methods optimize interaction dynamics directly. ASPIRin uses speech–silence actions and timing-oriented GRPO rewards [105], while Multi-Faceted Interactivity Alignment applies reinforcement learning with axis-specific rewards for pause handling, turn-taking, backchanneling, and user interruption, and adds an LLM-based response-quality reward to prevent semantic degradation [12, 201]. DuplexPO further factorizes the reward across turn initiation, acknowledgments, yielding, and regularized participation [157]: the turn-initiation term decays with the delay between the reference and predicted response onset, the acknowledgment term rewards agent speech that falls inside the annotated backchannel interval, the yielding term penalizes speech that continues beyond a tolerance after a user interruption, and the participation term penalizes undesirable interaction patterns. Credit is assigned locally by optimizing only the tokens inside short windows around each annotated event, with the preceding history teacher-forced. These methods use separate rewards for interaction behaviors and for response quality, and DuplexPO additionally keeps the policy close to its supervised initialization through a KL term. Without an objective that preserves content quality, timing optimization can reduce response quality.

Concurrent reasoning and coordination. Training must also support reasoning during ongoing input or output. FLAIR trains a recursively updated latent reasoning state during user speech through an evidence-lower-bound objective and teacher forcing, avoiding the need for explicit reasoning traces [292]. This provides a single-model route to reasoning during perception.

Dual-model systems expose the coordination decisions directly. DuplexOmni uses temporally annotated control events for thinking requests, waiting, interruption, speech cutoff, silence, and delayed result delivery, and trains these events through likelihood objectives over Writer–Director traces [112]. More generally, supervision for dual-model systems covers background requests, the context sent, when returned results are incorporated, and how outdated work is canceled or revalidated. A background result may no longer match the live intent when it returns, so correctness depends on both its content and its validity at the time of insertion. Standardized objectives for stale-result detection and recovery remain limited. Table 5 summarizes the four objective categories by their training signals and the interaction decisions they supervise.

Table 5: Training-objective categories for native real-time omni interaction. Safety and balancing between objectives apply across all four categories.

Objective category	Main training signal	Core decision
Content and cross-modal grounding	Token/code likelihood, reconstruction, and speech–text alignment	What multimodal content should be represented and generated?
Temporal control	Duration, fixed-duration chunks, silence, interaction-state labels, and response windows	When should the model speak, wait, acknowledge, yield, or stop?
Interaction-policy alignment	Ranked candidates and interaction-specific rewards	Which timing and conversational behavior should be preferred?
Concurrent reasoning and coordination	Latent-state objectives and temporally annotated control events	How should reasoning, delegation, result insertion, and cancellation proceed during interaction?

Safety and objective balancing. Safety training should evaluate partially observed input and partially generated output, because both remain incomplete while interaction unfolds, including when new evidence requires cancellation or correction. More broadly, objectives for content accuracy, acoustic quality, latency, timing, reasoning, and safety may generate competing gradients. Projected action spaces, factorized rewards, and semantic-quality rewards reduce part of this interference [105, 157, 201], but current systems rarely report objectives that jointly cover continuous multimodal input, partial output, and asynchronous tool results.

5. Maintaining Omni Intelligence

This section examines the capabilities needed to maintain a coherent understanding as new observations, user corrections, and external results arrive during a session. A completed turn gives a model a stable block of context. A live interaction does not. New events continue to arrive, a user may correct an earlier statement, and retrieved evidence may return after the conversation has already moved forward. The model must therefore maintain a coherent representation of events, evidence, and participants as the interaction changes.

We organize this section around five capabilities. Temporal grounding tracks when events occur and how they relate. Memory and session state preserve events, entities, goals, and corrections across the interaction. Reasoning and tool use operate over partial input and continue while speech is being heard or played. Factual grounding ties a claim to the evidence available when that claim is made. Social and pragmatic understanding identifies speakers, intentions, emotional cues, and conversational roles. These capabilities define the intelligence side of native omni interaction. Sections 3 and 4 discuss the supporting representations and architectures; Sections 6–9 address interactive behavior, serving, evaluation, trade-offs, and open problems.

5.1. Temporal Grounding

Event time as an internal model state. Temporal grounding is the ability to represent when an event occurs and how that event is related to the rest of the interaction [109, 224, 240, 275]. The relevant state includes event boundaries, timestamps, duration, ordering, overlap, and relations such as before, during, and after. These quantities are separate from event identity. A model may recognize that a person picked up a cup, that a door closed, or that a user mentioned a meeting while assigning the observation to the wrong

interval. In a completed video, this may appear as a localization error [102, 109, 275]. In a live session, it changes which evidence is valid for the current response.

Earlier video-language models explicitly predict event boundaries and timestamps. VTimeLLM [109] introduces boundary-aware training for video moments, TimeChat [224] introduces a timestamp-aware frame encoder and instruction tuning for time-sensitive long-video understanding, and Grounded-VideoLLM [275] represents temporal locations with dedicated time tokens. In real-time interaction, temporal grounding must update using only observations available so far, rather than relying on a completed clip: at time t, the model can use events observed up to t and cannot assume that the remainder of the stream is known.

Temporal relations across modalities. An omni model also needs temporal relations across modalities [32]. A spoken reference can overlap a pointing gesture. A sound may begin before its visual source becomes clear. A tool result may arrive several seconds after the request that initiated it. FineMoLA provides a complementary weakly supervised mechanism for learning fine-grained frame–phrase temporal correspondence from clip-level descriptions [280]. The model must retain each observation's time interval and link observations that refer to the same event. This gives temporal grounding three levels of structure: localization within a stream, relations across streams, and whether an event is past, ongoing, or not yet observed.

The third level is the main change from turn-based modeling. Turn-based input preserves sequence order, but the model receives the sequence after the turn is complete. A native model must know whether a relevant event has already happened, is still unfolding, or has not yet become observable. This distinction is needed even before a response policy is applied. It determines whether the current state contains enough information to support a reference, a reasoning step, a tool argument, or a factual claim.

Systems with explicit temporal alignment. Several recent systems expose this requirement directly. SyncLLM inserts synchronization tokens at fixed temporal intervals so that language-model updates advance with real-world time [268]. MiniCPM-o 4.5 uses Omni-Flow to align continuous multimodal input and output along a shared temporal axis [56]. Aligned timestamps allow audio and visual observations to update a shared interaction state during speech generation. ROMA introduces a lightweight speak head that predicts response initiation from streaming multimodal state [263]. The speak decision belongs to interaction control, but its input must encode whether the event that justifies a response has occurred and whether the relevant evidence is complete enough to act on.

Figure 9 illustrates a time-indexed event representation with event boundaries, cross-modal links, and the current observation time. Temporal grounding identifies when each observation occurred and whether it is relevant to the current response. Section 3.3 discusses how timestamps, silence, and overlap enter the model representation. Section 7.1 defines timing and temporal measurements. Section 8 analyzes premature and delayed commitment. The role of temporal grounding here is narrower: it is the model's internal understanding of when the evidence belongs.

5.2. Memory and Session State

A transcript alone may not capture recurring entities, changing goals, and corrections in a multimodal session: people and objects recur, events create consequences, user goals change, and later corrections can replace earlier assumptions. The model must track these changes together with the source and time of the supporting observations. This requirement is especially important for continuous audio and video, where an observation may be a short sound, a gesture, a visible object state, or a relation between two modalities.

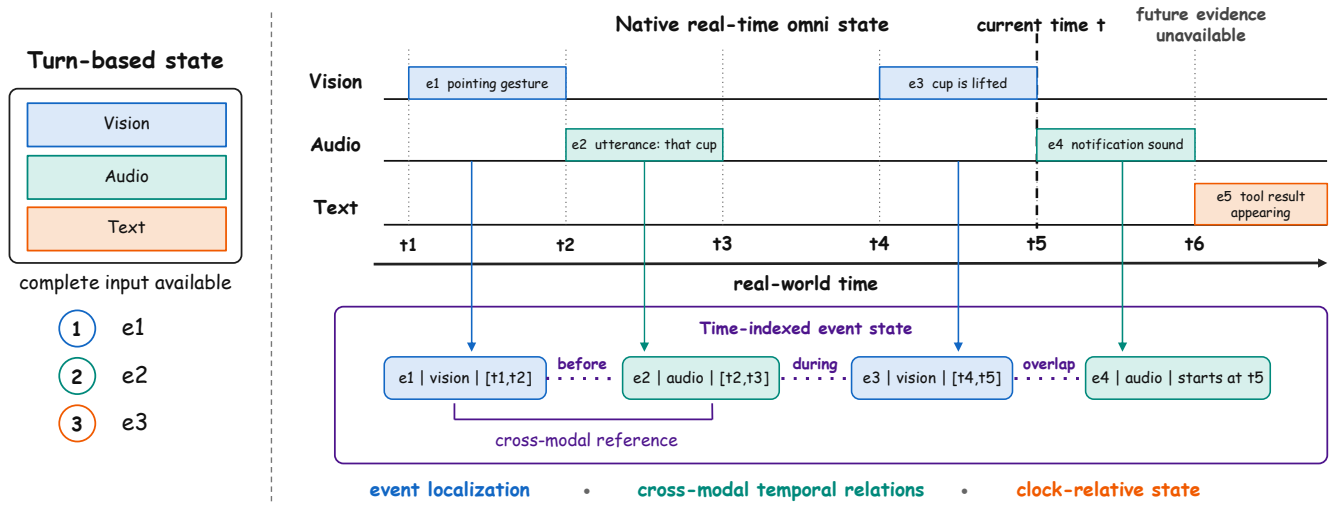

Figure 9: Temporal grounding during ongoing interaction. Observations retain their time intervals and cross-modal links within a time-indexed event state. The current time separates available observations from future input, so event localization and reasoning use only information observed so far.

Three semantic layers. We describe session state with three semantic layers. The *event state* stores temporally grounded observations. Each event records what occurred, when it occurred, its modality and source, and any confidence or cross-modal links needed by later reasoning [109, 275]. The *entity state* connects observations that refer to the same person, object, location, document, or external resource. It preserves identity across time and updates attributes or relations when new evidence arrives [314]. The *versioned task state* records the active goal, constraints, unresolved steps, corrections, tool results, and commitments. A correction creates a new task version and identifies which earlier conclusions depend on the superseded state [35, 290].

These layers answer different questions. Event state supports references to what happened at a particular time. Entity state supports continuity across repeated observations. Versioned task state supports revision. The distinction matters because the same new input can have different effects on each layer. A user correction may leave the identity of an object unchanged, add a new event, and replace one task constraint. Treating the entire session as an append-only text history hides this structure.

Working, episodic, and long-term memory. The semantic layers operate over several stores that differ in retention duration. Working memory keeps the recent multimodal context required for local references and immediate reasoning [108]. Episodic memory preserves events and task transitions across the current session [210]. Long-term memory stores selected user facts, preferences, entities, and recurring goals across sessions [291, 314]. The retention duration describes how long an item remains available; the semantic layer describes what the item means. A person's identity may be retained across sessions, while a temporary tool result may be relevant only to the current task.

Each stored item should retain its source, modality, timestamp, and relevant task version. Otherwise, a retrieved item may be semantically relevant and still refer to another user, an earlier time, or an outdated version of the task. Table 6 summarizes the three semantic layers and the update operations they require.

Queryable state over multimodal streams. Recent systems implement parts of this memory design. ReKV stores video key-value caches outside GPU memory and retrieves relevant segments when a question arrives [63]. Its contribution is primarily an efficient streaming mechanism, which belongs to the system discussion

Table 6: Session-state taxonomy for native omni intelligence. The state type specifies what is represented; the retention period specifies how long it is stored.

State type	Stored content	Source and timing information	Main update operations	Retention period
Event state	Time-bounded observations, utterances, actions, sounds, visual changes, and external results	Timestamp or interval, modality, source or speaker, confidence, cross-modal links	Append, merge, split, relabel, revise boundaries	Working; episodic
Entity state	Identities, attributes, and relations of recurring people, objects, locations, documents, and resources	Links to supporting event IDs, identity evidence, last observation time	Re-identify, update attributes, merge or separate identities, revise relations	Working; episodic; long-term
Versioned task state	Current goal, constraints, corrections, unresolved steps, tool results, commitments, and dependencies	Task version, supporting events and entities, tool source, validity status	Create new version, supersede, invalidate dependent state, revalidate, commit	Working; episodic

in Section 6, but it also shows why long visual history must remain queryable by temporal segment. InteractiveOmni extends multi-turn interaction to audio and vision, requiring earlier multimodal observations to remain usable later in the dialogue [264]. EgoMem associates information with users using face and voice cues, maintaining facts, preferences, and social relations across sessions in real-time omnimodal streams [314].

From retrievable history to evolving latent state. Streaming memory need not remain an external store that is repeatedly retrieved into the context window. FlexMem uses visual key–value caches as memory, applying two compression pathways together with task-dependent retrieval for long and streaming video without additional training [41]. LatentStream instead progressively consolidates historical observations into a compact, fixed-length latent memory for subsequent reasoning [219]. Taken together, these approaches store history as key–value caches, externally retrieved records, or compressed latent representations. Some also retain semantic or personalized information across sessions. StreamArena's StreamMind system illustrates the same separation at the system level, keeping latency-critical interaction distinct from persistent multimodal recall over hour-scale video [343]. Streaming interaction requires memory to remain bounded and revisable while new observations continue to arrive.

Session memory should support retrieval and revision while preserving source information. Sections 6, 8.2, and 9.2 discuss storage, consistency failures, and long-session memory, respectively. The event, entity, and versioned task layers defined here provide the shared vocabulary for those later discussions.

5.3. Reasoning and Tool Use under Real-Time Constraints

Reasoning and tool use share a constraint that does not appear in the same form for a completed prompt. Computation can start while the user is still speaking, continue while the model's audio is being played, and return after the user has changed the task. Intermediate results are therefore tied to the input prefix and task version that produced them. The model needs to reason causally over available evidence and preserve that association when reasoning invokes retrieval or another external tool.

Reasoning while listening. An internal text stream can support language processing while speech generation continues. Moshi maintains an inner-monologue text stream alongside the speech streams [61]. Qwen3-Omni adopts a Thinker–Talker architecture in which the Thinker processes multimodal inputs and provides semantic states, while the Talker generates streaming speech [302]. Their architectural details belong to Section 4. Semantic computation can retain its own state during speech generation. Cognitive Supersensing illustrates visual latent reasoning in an offline setting [150].

Other methods reason directly during ongoing input or speech. FLAIR updates a latent reasoning state while the user speaks, preserving causal processing [292]. STITCH interleaves internal reasoning with speech playback, using playback time to prepare later reasoning without inserting a full reasoning delay before speech [47]. Can Speech LLMs Think while Listening studies when a speech model should begin reasoning before an utterance is complete, using a question-completeness signal to control the accuracy-latency trade-off [243]. These methods differ in whether reasoning is explicit text or latent state and whether it runs during listening or speaking. Intermediate reasoning must remain consistent with the input observed so far.

Reasoning can also run asynchronously with perception. Video Streaming Thinking processes incoming clips while playback continues, rather than waiting for perception to terminate before deliberation begins [96]. R3-Streaming allocates additional reasoning to difficult queries after memory compression and response-readiness assessment, through a cascaded remember–respond–reason control process [170]. These designs allocate computation adaptively rather than forcing every input through the same reasoning budget, and they connect directly to the multi-rate view developed in Section 9.1.

Reasoning should retain uncertainty when the user's request is incomplete, and revise affected conclusions when constraints change, without treating all earlier observations as invalid. This property connects reasoning directly to the versioned task state in Section 5.2. Such reasoning concerns the model's current understanding; deciding when to speak is a separate interaction-control problem addressed in Section 6.

Tool use before the request is complete. External tools extend the same causal computation. Stream RAG predicts retrieval queries during ongoing user speech and integrates returned evidence into a later response [11]. Related work on multi-turn agentic scientific literature search uses workflow induction to organize external search across turns [151]. MoshiRAG performs retrieval asynchronously between response onset and the delivery of substantive information [48]. DuplexOmni pairs a low-latency interaction layer with a separate reasoning component that can provide reasoning and tool use in parallel [112]. These systems demonstrate that retrieval and external computation can overlap the spoken interaction.

The intelligence requirement concerns grounding and validity. A tool request should record its arguments, supporting evidence, task version, and launch time. Returned results must be checked against the current task before use. Query prediction and asynchronous execution can reduce idle time, but the returned information is useful only if it still answers the current task. Section 3.5 covers how reasoning, tool, and retrieval streams are represented. Section 8.2 analyzes runtime errors when reasoning or external results become inconsistent with the live state. Section 9.1 treats cancellation, rollback, and multi-rate computation as open design questions.

Figure 10 follows a session in which a correction updates the task version. Events update event and entity state, background results retain their originating version and are checked against the current state before the assistant uses them in a response.

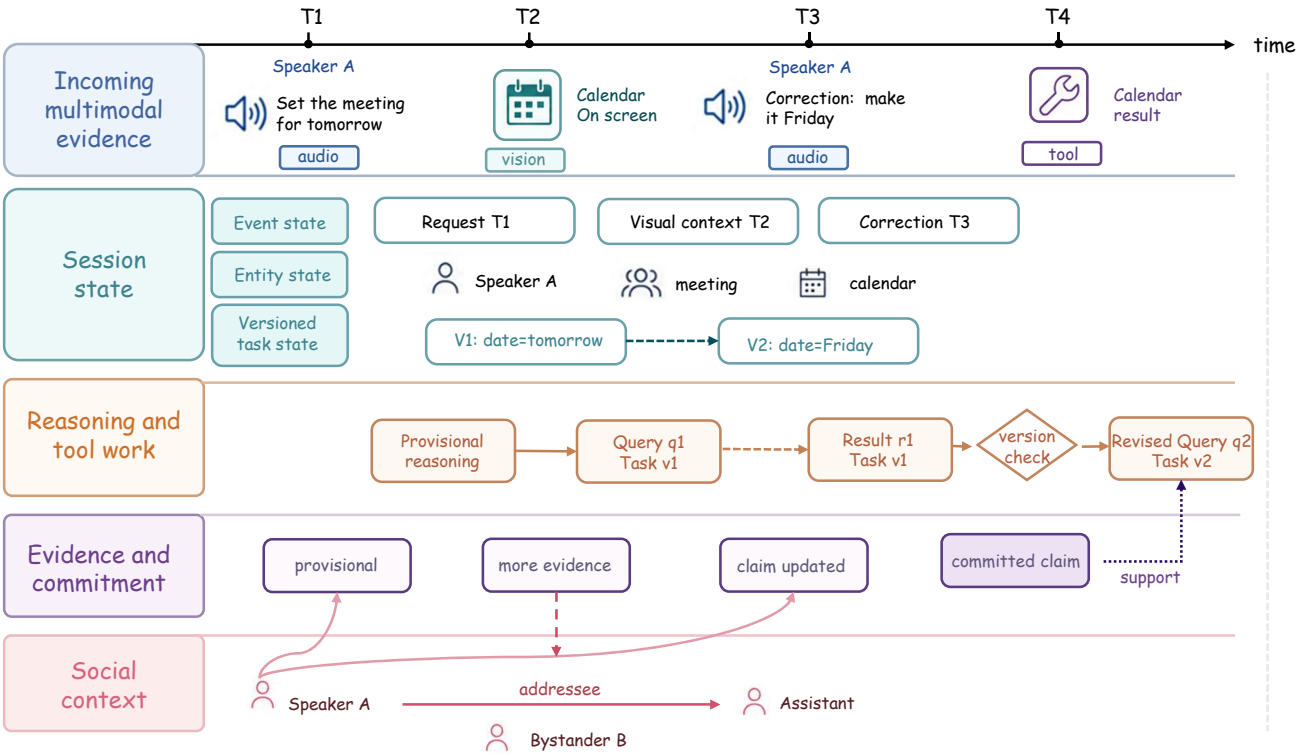

Figure 10: Maintaining a consistent understanding during a session. Multimodal evidence updates event, entity, and versioned task state. Reasoning and tool results proceed concurrently and retain their task version. After a correction, returned evidence is checked before it is incorporated into a response.

5.4. Factual Grounding and Correctness

A model may begin responding before all evidence needed for an answer is available. It must distinguish direct observations, retrieved information, memory, and inference, and identify which claims each source supports at the current time.

Evidence supporting a response. Before presenting a factual claim, the model should identify its supporting evidence and verify that it remains relevant to the current task [23, 87]. Visual evidence can be indexed by object [338], region [173], track [42], or event interval [275]. Acoustic evidence can be indexed by speaker [195], sound source [2], prosodic segment [229], or time range [101]. Retrieved evidence carries the query, source, return time, and task version. Memory-based evidence points back to the event, entity, or earlier task state from which it was recovered [23, 265]. This structure allows the model to answer two questions before presenting substantive content: what supports the claim, and whether that support is valid for the current task state.

The strength of evidence can change while the model is already interacting. A partial request may support an acknowledgment before enough information is available for a factual answer. A visible event may resolve a reference after the user's speech has begun, and a retrieval result may arrive after an initial spoken phrase. The system can keep an answer provisional and communicate its uncertainty until supporting evidence arrives.

Perceptual and external grounding. Streaming retrieval provides a clear example of external grounding. Stream RAG and MoshiRAG allow later response content to depend on evidence retrieved while speech

continues [11, 48]. Qwen3-Omni provides a complementary perceptual example through its dedicated audio captioning model, which is fine-tuned for detailed audio descriptions with reduced hallucination [302]. The sources differ, but in both cases the model should track the evidence supporting each claim and check its relevance to the current task.

This subsection defines the positive grounding capability. Section 7 measures content quality and the timing of useful evidence. Section 8.1 examines preservation of pretrained language capabilities, including the degradation reported by URO-Bench and analyses of the text-speech understanding gap [55, 304]; Section 8.2 examines premature claims, loss of grounding, and outdated evidence. Table 7 links these capabilities to evaluation and failure analysis.

5.5. Social and Pragmatic Understanding

A live assistant receives social information that is reduced or absent in a text transcript. Prosody can indicate uncertainty, urgency, affect, irony, or conversational stance. Face, gaze, body orientation, and speaker location help determine who produced an utterance and who is being addressed. The lexical form of a request, the speaker's role, and the preceding interaction also affect register, politeness, and intent. Social and pragmatic understanding concerns interpreting these cues to inform response decisions.

Paralinguistic state. Speech models that preserve prosody provide a basis for this competence. SPIRIT-LM interleaves speech and text within a shared token stream, while its expressive variant augments speech units with pitch and style information [199]. USDM uses paralinguistic speech information to support more natural spoken responses [133]. Prosodic cues inform estimates of speaker emotion, confidence, stance, and intent. They can change the interpretation of the same words and can influence later responses or memory updates.

Speaker and addressee. Multi-party interaction requires separate estimates of speaker identity and addressee. A person can be audible without addressing the assistant. A speaker can refer to another participant while looking at an object. Addressee recognition therefore combines linguistic context with gaze, orientation, and conversational history. Earlier multimodal work showed that gaze and utterance content can be combined for addressee recognition [188]. More recent work shows that the problem remains challenging in multimodal multi-party dialogue: on a triadic-dialogue benchmark, GPT-4o achieves addressee-recognition accuracy only marginally above chance [117].

Visual cues help distinguish speakers in multiparty settings. AV-Dialog combines audio and visual cues to track the target speaker in noisy multi-speaker dialogue [42]. The model can then track who is speaking, who is being addressed, and which visible entities are relevant to the current utterance. The later decision to take the floor, backchannel, or remain silent is an interaction behavior and is evaluated in Section 7.

Role, register, and pragmatic context. Social interpretation also considers participants' roles and relationships. The same instruction can carry different authority depending on who speaks, and the same words can convey a request, correction, joke, or refusal depending on prosody and context. Role-conditioned systems such as PersonaPlex demonstrate that duplex behavior can be explicitly conditioned on textual role prompts and speech-based voice prompts [230]. Explicit role prompts differ from inferring roles during an unconstrained interaction, where the relevant role may need to be inferred from multimodal context [117, 188].

This subsection focuses on interpreting social cues. Section 7 evaluates response behavior and timing, while Sections 8.3 and 9.2 discuss privacy, personalization, and multiparty coordination.

Table 7: Capability map for omni intelligence. The final two columns are cross-references to the dedicated measurement and trade-off analyses in later sections.

Capability	Definition	Key mechanisms	Representative systems	Changes required for native interaction	Measurement	Failure or trade-off
Temporal grounding	Event time, boundaries, duration, ordering, overlap, and relation to the current time	Temporal localization; explicit timestamps and temporal position encodings; shared multimodal time axis; event-state prediction	VTimeLLM; TimeChat; SyncLLM; MiniCPM-o 4.5; ROMA	Track events incrementally using only observations available so far	Sec. 7.1 timing and temporal quality	Sec. 8.2 premature or delayed responses and temporal misalignment
Memory and session state	Evolving event, entity, goal, correction, and task-version state	Working, episodic, and long-term memory; temporal and modality indexing; state versioning; cache-based, retrieved, or latent memory	ReKV; InteractiveOmni; EgoMem; FlexMem; LatentStream	Maintain revisable, attributed state over continuous audio-video sessions	Sec. 7.1 long-session quality	Sec. 8.2 evidence and state coordination
Reasoning and tool use	Reasoning and external computation using partially observed, revisable input	Inner-monologue or latent reasoning; background computation; asynchronous retrieval and tools; conditional and overlapping reasoning; task-version binding	Moshi; Qwen3-Omni; FLAIR; STITCH; VST; R3-Streaming; Stream RAG; MoshiRAG; DuplexOmni	Overlap computation with listening and speaking	Sec. 7.1 content, repair, and tool success	Sec. 8.1 capability preservation; Sec. 8.2 runtime coordination
Factual grounding	Claims supported by current perceptual, retrieved, remembered, or inferred evidence	Source tracking; links between claims and evidence; uncertainty estimates; evidence checks before responding	Stream RAG; MoshiRAG; Qwen3-Omni	Ensure claims are correct at the time they are made	Sec. 7.1 content quality and evidence timing	Sec. 8.1 factual capability preservation; Sec. 8.2 loss of grounding
Social and pragmatic understanding	Speaker emotion, intent, addressee, roles, relationships, and conversational context	Prosodic state; audiovisual speaker tracking; gaze and addressee inference; state conditioned on participants	SPIRIT-LM; USDM; AV-Dialog	Interpret multimodal cues from multiple participants	Sec. 7.1 timing and robustness; Sec. 7.2 multi-party coverage	Sec. 8.3 source use and privacy

5.6. Synthesis: Interdependent Capabilities

The capabilities above operate on shared state. Temporal grounding supplies the event structure used by memory and online reasoning. Memory keeps the entities, corrections, and task versions that later reasoning must recover. Reasoning and tools add inferred and external evidence. Factual grounding determines which

parts of that evidence support a claim at the current time. Social understanding assigns observations and requests to the correct participants and supplies role and intent information to the task state.

Continuous interaction requires these capabilities to share a consistent state as new input arrives. A completed prompt can be processed after its evidence and intent are fixed, whereas a native omni model must keep time, state, inference, evidence, and participant attribution consistent while the session continues to change. Sections 7 and 8 examine their evaluation and failure modes.

6. Realizing Omni Interactivity

Omni interactivity concerns response timing and behavior during an ongoing session. We examine response latency, turn-taking, interruption recovery, proactive assistance, adaptation across turns, and the serving and client mechanisms that support them.

These behaviors build on the representations of time, silence, overlap, and backchannels described in Section 3.3, together with the temporal grounding, session state, reasoning, and tool validity examined in Section 5. They depend on both the interaction policy and its execution across the model, server, and client. Section 7 covers their metrics and benchmarks, Section 8 analyzes the intelligence–interactivity trade-off, and Section 9 develops the open research agenda.

Figure 11 summarizes the resulting interaction loop. Continuous multimodal input updates the shared state, and the interaction policy selects silence, a backchannel, substantive speech, a proactive intervention, or a control action. Playback and user feedback inform the next decision. The model, server, and client must agree on input timing, playback progress, and which outputs have reached the user.

6.1. Latency and Responsiveness

End-to-end latency measures the delay from a relevant input event to output reaching the user [57, 346]. It includes capture and chunking, audio or visual encoding, inference, speech synthesis, network transport, buffering, and playback delays. Each stage contributes to the total latency, and a stall at any stage delays the observable response.

Cascaded systems serialize several stages [44, 57]. Voice activity detection closes the input segment, automatic speech recognition produces text, the language model generates a response, and text-to-speech converts that response into audio. Streaming components reduce individual delays, although module boundaries still introduce buffering and scheduling gaps. Native omni designs can overlap capture, encoding, prefill, decoding, and synthesis [61, 281, 301]. Section 4 examines their architectural details, while the latency budget records the resulting critical path.

Interruption latency also matters. An interruption must propagate through the client audio controller, server scheduler, generation state, and queued playback. Fast response onset combined with slow stopping still produces a rigid interaction. Systems should therefore specify timing requirements for both starting and stopping output, including input chunking, response onset, per-token generation, playback continuity, and stop propagation. Section 7 defines the corresponding measurements and reporting protocol.

Smaller audio chunks and more frequent video sampling can improve responsiveness but increase processing cost [72, 257]. Larger playback buffers tolerate network variation but delay stopping. Larger batches improve throughput but can increase queueing delay. These dependencies determine the deadlines that the serving stack in Section 6.6 must protect. Response onset and the arrival of task-relevant information can occur at different times, an issue examined as part of the intelligence–interactivity trade-off in Section 8. The system

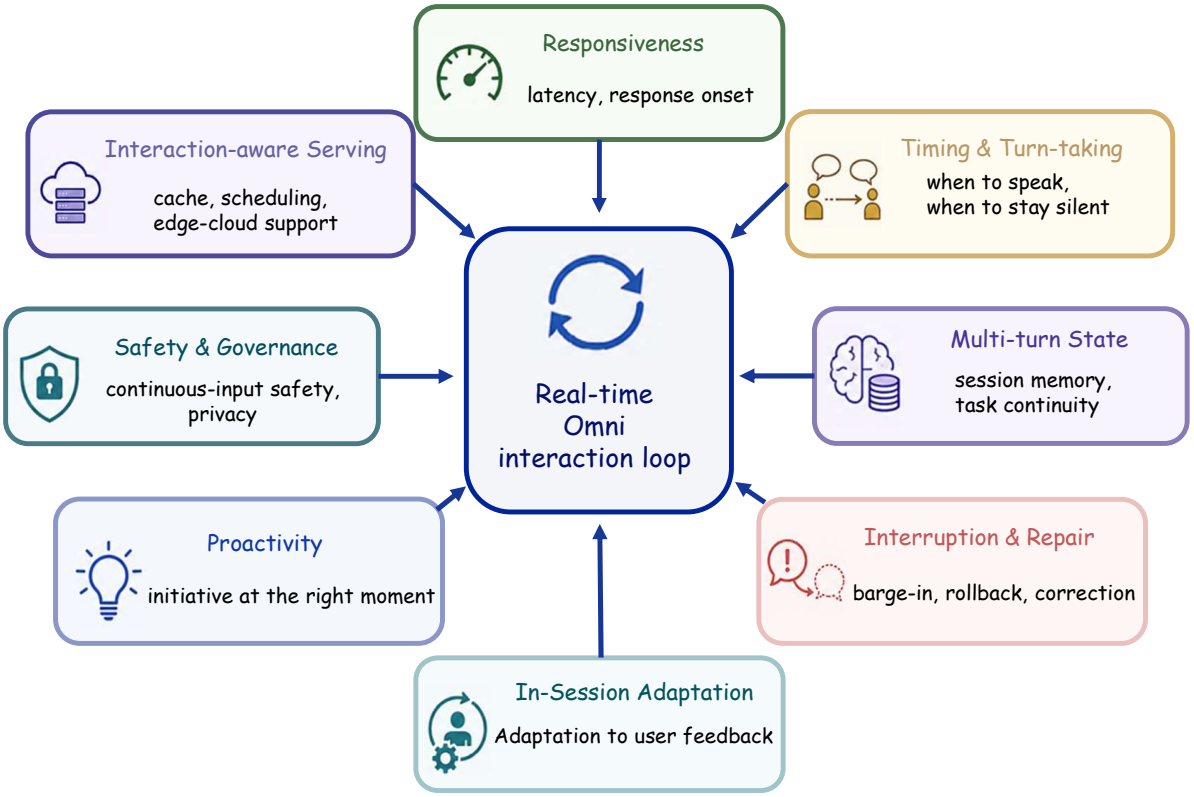

Figure 11: The omni interaction loop. Continuous audio, video, text, and system events update a shared state. The interaction policy controls response timing, floor management, repair, proactivity, and adaptation. Serving and client mechanisms execute each decision and return playback and user feedback to the loop.

should indicate whether it is waiting, acknowledging a request, seeking clarification, or providing an answer. A brief acknowledgment can precede a complete answer, with its delivery timed using turn-taking predictions. In a deployed dialogue robot, recent work prepares a context-aware prefatory response once user intent becomes predictable, while a separate Voice Activity Projection process determines when to deliver it [202]. This separates *content readiness* from *floor-taking readiness*.

6.2. Turn-Taking, Silence, and Backchannels

Turn-taking converts the time-aware representations described in Section 3.3 into an output policy. At each update, the model chooses among continued listening, silence, a backchannel, floor-taking, and yielding. Silence can allow the user to continue, indicate uncertainty, or delay a response until an event occurs. Backchannels acknowledge the user, while substantive speech begins the model's turn. Yielding releases the floor.

A policy can use acoustic, linguistic, visual, and task-state cues to select an action and its timing [70, 188, 214, 251]. Acoustic cues provide prosody, pause structure, overlap, and speech onset. Linguistic context indicates whether an utterance is complete, while visual context provides gaze, gesture, addressee, and ongoing activity. Task state indicates whether the model has enough information to proceed. Section 5 examines how these signals are interpreted and maintained in the interaction state.

Human conversation constrains how these cues should be mapped to actions. Long gaps can feel unresponsive, premature floor-taking cuts off the user, and excessive backchannels become distracting. A pause may indicate

Table 8: Interruption repair stages. Each stage changes a different part of the live system.

Stage	Required transition	Failure if omitted
Detect	Classify barge-in, backchannel, side speech, or noise	False stop or ignored correction
Yield	Duck speech and flush pending playback	Model talks over the user
Roll back	Align context with played content	References to undelivered output
Cancel	Stop work tied to the old task version	Stale evidence or side effects
Revise	Apply the corrected intent and constraints	Continued execution of the old request
Resume	Continue from the corrected interaction state	Repetition or lost task progress

hesitation or turn completion, so the appropriate response depends on linguistic, visual, and social context. Background speech may require silence even when the acoustic channel is active.

Current systems place turn-taking control in different components. Speech-native full-duplex models such as dGSLM and Moshi learn timing through parallel conversational streams [61, 198]. Freeze-Omni uses a chunk-level state predictor to control listening and response transitions [281]. SALMONN-omni represents speaking decisions with learned state tokens during generation [324]. BayLing-Duplex learns listen, speak, and stop control within a single autoregressive speech–language model using a small set of control tokens [76]. Analyses of interacting Moshi instances identify internal activations associated with upcoming turn changes [226]. A deployed system can combine learned timing predictions with explicit execution rules shared by the client and scheduler.

Section 3 describes the supporting representations, and Section 7 evaluates response timing, takeover rate, silence, and backchannel quality. The present subsection specifies the action policy that these representations support and the system executes.

6.3. Interruption and Repair

Interruptions can change the user's request while speech or background work continues. The system must stop pending output, update the context and task, cancel outdated work where possible, and resume consistently.

Table 8 separates this process into linked stages. Detection classifies incoming activity during model speech. When an interruption requires the system to yield, it returns the floor and stops pending playback. Rollback then aligns the model context with content delivered to the user, while cancellation stops retrieval, reasoning, and tools associated with a superseded task version. Revision applies the correction to the versioned state defined in Section 5, allowing resumption to proceed from the corrected interaction state.

To support rollback and consistent resumption, the conversational record should distinguish generated, queued, played, and canceled content, using playback progress to track what has been delivered to the user. ChipChat implements this by tracking vocalized tokens and rolling the context back to the point reached by playback [159]. Training can also shape the conversational behavior surrounding these transitions. Multi-Faceted Interactivity Alignment uses separate rewards for pause handling, turn-taking, backchannels, and interruption, together with a response-quality reward [201]. By combining interaction and content objectives, it connects repair behavior to the alignment objectives discussed in Section 4.3.

Repair must also extend to tool operations, which may already have changed external state when an inter-

ruption arrives. Recovery therefore needs to account for what was executed and which results remain valid. The versioned task and evidence states in Section 5 help identify outdated work, while recovery may require cancellation, confirmation, or compensation. These requirements also inform the evaluation in Section 7, which examines whether repair succeeds and how much obsolete content is delivered before recovery.

Coordinating this recovery requires information from the client. Echo cancellation and near-end speech detection help identify incoming user speech, while audio ducking reduces playback and cancellation stops pending output. Playback timestamps then help the server distinguish content already delivered to the user from content that should be removed from the shared conversational record. Section 6.6 places these functions within the complete serving and client path.

6.4. Proactive Interaction

Proactive interaction occurs when the system initiates a contribution without a new explicit request. A visual event, acoustic cue, elapsed interval, tool result, or change in task state can trigger this decision [37, 242, 274, 305, 306]. Temporal grounding and event understanding examined in Section 5 determine what happened and when, while the proactive policy determines whether an intervention is appropriate at the current moment. The policy should consider task relevance, urgency, confidence, distraction, privacy, social context, and whether an action can be reversed, together with user permissions and preferences [139, 305, 306]. A safety warning may justify immediate speech, while a low-confidence suggestion can wait for a pause or appear as a quiet visual cue [104].

Streaming video systems provide early implementations of this policy. VideoLLM-online predicts a response or silence at each frame, MM-Duet allows either participant to contribute during video playback, and systems with separate perception, decision, and reaction components continue observing while selecting a response [37, 215, 284]. StreamBridge adds proactive activation to models trained for offline video understanding [274]. Proact-VL trains multimodal language models for continuous video perception, autonomous response triggering, and latency-aware commentary generation. Its Live Gaming Benchmark evaluates response timing and content selection in solo commentary, co-commentary, and user guidance [305]. These systems repeatedly decide whether to speak or continue observing as new input arrives, using model predictions to guide response timing.

When background results arrive, the model decides whether to use them immediately, wait, notify the user nonverbally, or discard them. It considers whether the result still matches the task and whether the user is currently speaking. Foreground–background architectures provide the concurrent computation, while the proactive policy controls how its results are presented [11, 48, 262].

Section 7 evaluates proactive response timing, while Sections 8 and 9 discuss decisions based on incomplete evidence, persistent interaction, privacy, and safety.

6.5. Adaptation Across Turns

Within-session adaptation changes response behavior based on user feedback. Subsequent turns should reflect preferences about response length, speaking pace, acknowledgment frequency, interruption tolerance, and initiative [136, 259]. A corrected answer demonstrates repair, while a sustained change in behavior demonstrates adaptation. The interaction layer can represent these preferences through policy variables such as response length, speech rate, pause duration, backchannel rate, confirmation threshold, and proactivity threshold [136, 259]. Each preference update should specify where it applies and when it expires. A task-specific correction can expire with the task, while a session preference can persist until the conversation

ends. Section 5 examines the memory and provenance needed to store these updates, and Section 9 considers persistent personalization across sessions.

Interaction-policy adaptation. Appropriate turn-taking depends on the task and participant roles. A tutor may interrupt proactively, whereas a counseling assistant may need to avoid interruption. Instruct-FD evaluates whether full-duplex systems can follow instructions about when to listen, backchannel, interrupt, or continue speaking, and reports a substantial gap between fluid and controllable full-duplex behavior [259]. DuplexGen shows that preferences over turn-taking vary across scenarios and can be learned from a small amount of preference supervision [136]. Together, these studies motivate interaction policies controlled by instructions and user preferences, with timing patterns learned from conversational data serving as an initial policy.

Preferences can be stored explicitly or inferred from recent interaction history. Explicit state supports inspection, correction, and deletion [45], while context-conditioned control infers preferences from recent feedback [144]. Both approaches can adapt behavior during a session without changing model weights. Section 4.3 examines offline alignment for timing, turn-taking, backchannels, and interruption.

Preference updates should rely on repeated or explicit feedback. A single interruption may correct content, signal urgency, or result from background noise, so it does not necessarily indicate a lasting preference [98, 144]. Repeated interruptions during long answers provide stronger evidence for shorter responses, and rejected proactive suggestions support a higher intervention threshold in the same context [98, 144]. Repeated corrections about an entity should increase confidence in the updated interpretation [45, 98].

Section 7 evaluates adaptation by comparing behavior before and after feedback under matched conditions. Within-session adaptation changes response timing, style, initiative, or confirmation behavior, while Section 9 considers learning, personalization, and consistency across sessions.

6.6. Full-Duplex Serving and Client Coordination

The behaviors above depend on a serving stack for long-lived, bidirectional sessions. Serving systems must coordinate incoming media, playback, state updates, and cancellation [112, 262, 281]. Audio and video arrive in frequent small chunks, speech decoding may continue as new input arrives, and an interruption can invalidate part of the current context. Tools and background models also return results on independent schedules [11, 48, 112].

Persistent streaming sessions provide the basic execution unit. Per-session encoder state, KV cache, interaction state, and playback state remain available across chunks and workers. Freeze-Omni separates user-specific caches from model execution so that a stateless worker pool can process active sessions [281]. Thinking Machines Lab appends 200 ms micro-turns to a persistent GPU-resident sequence and optimizes MoE kernels for the small input shapes produced by bidirectional serving [262]. Section 5 defines the semantic content of session state, while this subsection covers its storage, scheduling, and mutation.

Scheduling must meet speech-generation deadlines while processing new input. Incoming chunks produce frequent prefills that compete with ongoing decoding. Chunked prefill divides new context into bounded units, and continuous batching schedules work at iteration granularity [3, 4, 318]. Separating prefill and decoding across resources can reduce interference between them [208, 348]. Playback progress and user activity can also help interaction-aware schedulers prioritize sessions whose audio buffers are nearly empty and prepare for barge-in [346].

Cache management must limit memory use while retaining timestamps, task versions, corrections, and playback or output-status records needed for recovery. Paged allocation reduces fragmentation, while sliding

Table 9: System support for omni interactivity. Each mechanism covers a distinct part of the interaction loop.

Requirement	Representative mechanism	Failure when absent
Continuous media flow	Chunked prefill and deadline-aware scheduling	Audio stalls during new input
Bounded session state	Paged KV and selective retention	Memory growth or lost interaction state
Correct repair	Playback timestamps, queue flush, and rollback	References to unheard output
Valid background work	Task versions, validation, and cancellation	Stale results re-enter the session
Robust deployment	Edge control and adaptive media transport	Network delay breaks the dialogue loop

windows, attention sinks, selective eviction, and retrieval-backed caches limit active memory [63, 140, 296, 345]. Semantic relevance alone may not identify every state required for repair.

MoE backbones introduce additional latency at the small batch sizes common in interactive decoding. Expert-weight movement, routing, and cross-device communication can dominate each decoding step [59, 79, 84, 114, 124, 319]. Expert locality, prefetching, phase-specific parallelism, and kernels optimized for small inputs can reduce this cost [114, 262, 319]. The system should keep decoding time within its latency target under realistic concurrent workloads.

Time-critical functions can run on the client or edge device, while larger models, retrieval, and tools run in the cloud. Edge functions can include wake-word detection [233], VAD [154], echo cancellation [254], audio compression [60, 326], floor control [70, 251], and interruption detection [154, 281]. UAF consolidates several of these functions by predicting VAD, turn detection, speaker recognition, ASR, question answering, and interruption-related control signals as a streaming autoregressive sequence [154]. Network degradation may require reduced media rates, lower codec bitrates, delayed background work, or simpler local behavior [128, 135, 174, 218].

The client completes the control loop by capturing media, performing echo cancellation, reducing playback volume during an interruption, maintaining a bounded jitter buffer, and sending playback timestamps to the server. It also shows whether the system is listening, processing, speaking, muted, or waiting for confirmation. These signals support turn coordination and user control. Table 9 maps the main system mechanisms to the behavioral failures they prevent.

Omni interactivity requires timely responses, appropriate turn-taking, coordinated recovery, useful proactive assistance, adaptation to feedback, and serving mechanisms that support these behaviors. Section 7 measures these capabilities, Section 8 examines their relation to intelligence, and Section 9 identifies the remaining research problems.

7. Evaluating Intelligence and Interactivity

Real-time evaluation must assess response content together with timing, interruption handling, correction recovery, and consistency across turns. We review evaluation metrics in Section 7.1, existing benchmarks and their coverage in Section 7.2, and limitations of the available protocols in Section 7.3.

7.1. Evaluation Metrics

We organize evaluation into six dimensions: latency, timing quality, content quality, robustness, long-session quality, and repair. Figure 12 summarizes their representative metrics and evaluation settings.

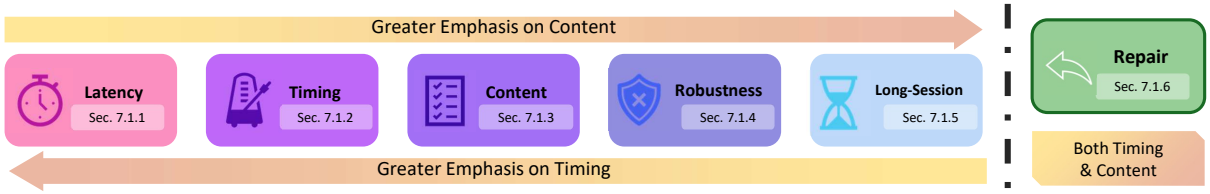

Figure 12: The six evaluation dimensions defined in Section 7.1 for native real-time omni models: latency, timing, content, robustness, long-session quality, and repair.

7.1.1. Latency

Latency captures several distinct delays, including the time to begin generating a response, deliver audible speech, and react to an interruption. Each measure depends on its starting event: first-token and first-audio latency may be measured from user-turn completion or from an annotated point when sufficient input has arrived. The following measures distinguish these stages of the interaction:

- **First-token latency**: elapsed time from declared input event to emission of the first text or semantic token.
- **First-audio latency**: elapsed time from the declared input event to the first audible output at the client, including synthesis, buffering, and playback delays.
- **Interruption-stop latency**: elapsed time from the onset of an interrupting user utterance to cessation of model speech at the client. State whether the start event is annotated speech onset or system detection of the interruption.
- **Response-after-interruption latency**: elapsed time from the interruption event to the start of the subsequent response, using the start and end events specified by the evaluation protocol. Full-Duplex-Bench v1 (FDB-v1) reports this as *Latency After Interruption* [160], which measures response resumption rather than how quickly the model stops its ongoing speech.
- **Turn-taking latency**: the gap between the end of the user's utterance and the onset of the model's response. FDB-v1 measures this as *response latency*, "the time (in seconds) between the end of the user's speech and the start of the model's response," computed only on turns where the model actually takes over the floor [160].

Observed latency depends on both the architecture and the conditions under which it runs. In cascaded VAD–ASR–LLM–TTS systems, delays accumulate across components and their buffers, with additional delays from network transmission and client playback. Interpreting reported values therefore requires information about the hardware, workload, number of concurrent sessions, and buffering configuration. FDB-v3 reports wall-clock latency for multi-step tool-use tasks [162], while SimulEval provides latency measures for simultaneous translation [179].

7.1.2. Timing Quality

Timing quality evaluates whether responses and backchannels occur at appropriate moments. TurnGPT predicts turn completion from text [69], while Voice Activity Projection (VAP) uses self-supervised learning to predict future speech activity from audio [70]. Follow-up work evaluates the effects of prosody [71] and fillers or pauses [125]. Skantze surveys turn-taking in spoken dialogue systems [251]. MuVAP extends prediction to multimodal, multiparty settings [214].

FDB-v1 uses **Takeover Rate (TOR)**, the fraction of probes in which the model begins a speaking turn, across pause handling, backchanneling, turn-taking, and interruption [160]. Its preferred direction depends on the scenario. A low takeover rate is desirable when a pause does not yield the user's turn; a high rate is desirable when the user yields and expects a response. Interruption results must identify who interrupts whom and the intended response behavior rather than assigning one preferred direction to all interruptions. For proactive video responses, ProactiveVideoQA evaluates response timing as well as content using PAUC [283].

7.1.3. Content Quality and Instruction Following

Content quality evaluates correctness, completeness, and instruction following. OmniBench measures static multimodal understanding and does not evaluate streaming interaction or response timing [156]. VoiceBench evaluates instruction following under acoustic and speaker variation [43]. Dynamic-SUPERB and its Phase-2 extension cover classification, regression, and sequence-generation tasks [110, 111]. Audio MultiChallenge evaluates instruction retention and recall of audio cues that are absent from transcripts [93]. SpokenWOZ reports task completion and dialogue-state tracking performance in spoken task-oriented dialogue [245].

7.1.4. Interaction Robustness

Robustness evaluates performance under overlapping speech, noise, interruptions, and side conversations. Overlap should be evaluated as a normal conversational condition as well as a source of interference; dGSLM models overlapping speech, laughter, and backchannels in two-channel dialogue [198]. FDB-v1.5 evaluates four overlap scenarios: user interruptions, user backchannels, speech directed to another person, and background speech. It measures response behavior, timing, and prosodic adaptation [164]. FDB-v1 also evaluates interruption handling [160]. VoiceBench varies speakers, acoustic conditions, and content [43]. IRAF addresses interfering speech by adjusting the contribution of incoming audio features; it is a method addressing a robustness failure, rather than a benchmark [347].

7.1.5. Long-Session Quality

Long-session evaluation covers consistency, entity and event tracking, task state, and memory over extended interaction. FDB-v2 uses an automated examiner with staged goals and two pacing conditions, Fast and Slow. Its tasks cover daily interaction, correction, entity tracking, and safety, and its results identify difficulties with overlapping speech, user corrections, and entity tracking [163]. Audio MultiChallenge adapts inference-memory and self-coherence evaluations to spoken multi-turn dialogue [93]. M3-DuplexBench evaluates multilingual dialogue with varying amounts of prior context [82]. SpokenWOZ measures task-state maintenance across spoken turns [245], while StreamingBench tests retention of earlier streamed information [165]. Multi-turn evaluation and continuous hours-long interaction should be distinguished when reporting session length. User simulation offers a complementary approach to long-session evaluation: a simulated user interacts with the system, and performance is assessed from the resulting interaction history [21].

7.1.6. Repair Behavior

Repair evaluation tests whether the model incorporates a correction and stops obsolete output or work. Audio MultiChallenge evaluates mid-utterance and previous-turn edits [93], while FDB-v2 tests correction in streaming multi-turn dialogue [163]. We propose reporting *correction-recovery rate*, the fraction of correction trials whose final response satisfies the revised request, and *stale-content rate*, the fraction that continue

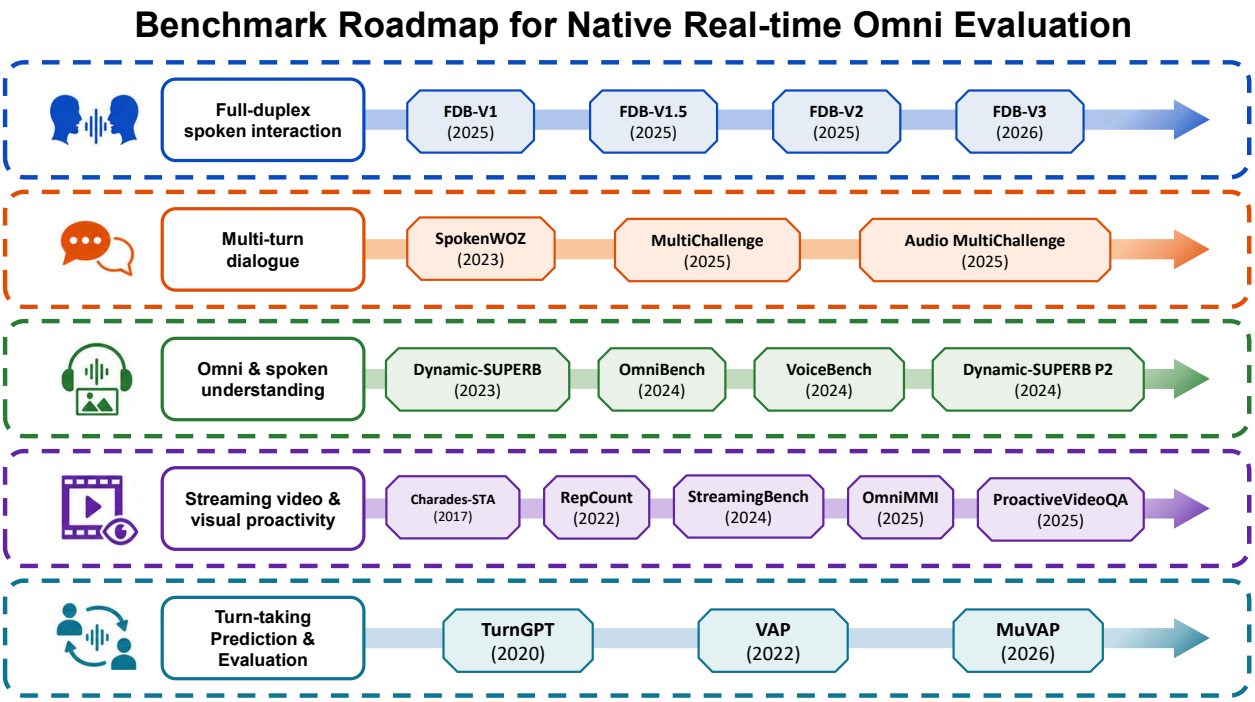

Figure 13: Representative evaluation resources by *first public release year*, grouped by benchmark family. Temporal-grounding and turn-taking evaluations precede benchmarks for end-to-end native interaction, while public full-duplex and online omnimodal benchmarks emerged mainly from 2025 onward.

using the superseded request after a specified recovery allowance. Reports should define the correction event, scoring rubric, recovery allowance, and trial denominator.

7.2. Comparative Benchmark Coverage

We group benchmarks by their primary focus on **Omni Intelligence** or **Omni Interactivity**, while comparing both groups using the six dimensions in Section 7.1. These groups describe a benchmark's primary purpose; individual benchmarks may assess capabilities from both groups.

Figure 13 presents the historical development of the benchmark families. The intelligence branch includes multi-turn dialogue evaluation, static omni-modal and spoken understanding, and streaming audiovisual and temporal understanding. The interactivity branch includes full-duplex spoken interaction, streaming and proactive omni interaction, and reusable tools and measures for evaluating turn-taking. Within each family, resources are ordered by first public release year.

Table 10 compares coverage across evaluation dimensions. Filled symbols indicate direct coverage through a dedicated task or separately reported metric. Partial symbols indicate limited or indirect coverage: a subset of tasks, or a capability affecting an aggregate score without a separate measurement. Empty symbols indicate absent coverage. These marks characterize benchmark scope rather than model performance.

Public reproducibility is treated separately from capability coverage. Benchmarks with released data, code, a hosted evaluator, or a public leaderboard support independent comparison. Publicly documented benchmarks whose release status cannot yet be independently confirmed are marked with ‡. Vendor and internal evaluations are marked with † and are not co-ranked with public benchmarks.

Table 10: Benchmarks grouped by whether they primarily assess Omni Intelligence or Omni Interactivity. Coverage is assessed across six dimensions from Section 7.1: **La**=Latency, **Ti**=Timing, **Co**=Content, **Ro**=Robustness, **Ls**=Long-session, and **Re**=Repair. Vendor/internal evaluations are reported separately and must not be co-ranked with public benchmarks. Coverage marks reflect the authors' assessment of each benchmark's stated scope.

Benchmark [ref]	Input modality	Evaluation setting	La	Ti	Co	Ro	Ls	Re
I. Omni Intelligence								
Evaluates the model's multimodal perception, semantic understanding, reasoning, instruction following, dialogue consistency, and long-context comprehension.								
I-A. Multi-turn dialogue understanding and consistency								
Audio MultiChallenge [93]	Speech (E2E)	Multi-turn dialogue	○	○	●	◑	●	●
MultiChallenge (text) [62]	Text	Multi-turn, turn-based	○	○	●	○	●	◑
SpokenWOZ [245]	Speech + text	Task-oriented dialogue	○	○	●	◑	●	○
I-B. Omni and spoken understanding								
VoiceBench [43]	Speech	Voice-assistant tasks	○	○	●	●	○	○
OmniBench [156]	Image + audio + text	Static tri-modal QA	○	○	●	○	○	○
Dynamic-SUPERB [110]	Speech + audio	Zero-shot instruction	○	○	●	◑	○	○
Dynamic-SUPERB Phase-2 [111]	Speech + audio	Zero-shot instruction	○	○	●	◑	○	○
I-C. Streaming visual intelligence								
StreamingBench [165]	Streaming video + audio	Streaming video QA	◑	◑	●	○	◑	○
Charades-STA [86]	Video + language	Temporal localization	○	◑	◑	○	○	○
Charades [246]	Video + language	Temporal understanding	○	◑	◑	○	○	○
RepCount (TransRAC) [107]	Video	Repetition counting	○	◑	●	○	○	○

II. Omni Interactivity								
Evaluates the model's real-time responsiveness, temporal coordination, turn-taking, interruption handling, proactivity, full-duplex communication, and recovery from interaction failures.								
Benchmark [ref]	**Input modality**	**Interaction setting**	**La**	**Ti**	**Co**	**Ro**	**Ls**	**Re**
II-A. Full-duplex spoken interaction								
Full-Duplex-Bench v1 [160]	Speech (duplex)	Single-turn probes	●	●	○	◑	○	○
Full-Duplex-Bench v1.5 [164]	Speech (duplex)	Overlapping speech	●	◑	○	●	○	○
Full-Duplex-Bench v2 [163]	Speech (duplex)	Multi-turn (Fast/Slow)	◑	●	◑	◑	●	●
Full-Duplex-Bench v3 [162]	Speech + tools	Multi-step tool use	●	●	◑	◑	◑	○
M3-DuplexBench [82]	Speech (duplex)	Multi-turn, multilingual	◑	●	◑	◑	●	◑
Instruct-FD [259]	Speech (duplex)	Instruction-conditioned turns	○	●	◑	◑	◑	◑
II-B. Streaming and proactive interaction								
OmniMMI [285]	Streaming video + audio	Streaming interaction	◑	●	●	○	◑	○
ProactiveVideoQA [283]	Streaming video	Proactive interaction	◑	●	◑	○	○	○
StreamArena [343]	Streaming video + audio	Long-horizon agentic	◑	◑	●	○	●	○
Proact-VL [305]	Streaming video	Proactive commentary	◑	●	◑	○	○	○
II-C. Turn-taking and interaction timing								
VAP [70]	Two-channel audio	Turn-taking modelling	○	●	○	○	○	○
TurnGPT [69]	Text	Turn-taking modelling	○	●	○	○	○	○
MuVAP [214]	Audio + video	Multiparty turn-taking	○	●	○	○	○	○
II-D. Vendor and internal interaction evaluations								
TimeSpeak [262]	Speech, reported	Interaction, reported	◑	●	◑	○	○	○
CueSpeak [262]	Speech, reported	Interaction, reported	◑	●	◑	○	○	○

7.2.1. Omni Intelligence Benchmarks

These benchmarks evaluate multimodal understanding, reasoning, event timing, and retention across turns. We distinguish three benchmark families.

Multi-turn dialogue understanding and consistency. This family evaluates whether model capabilities persist when relevant information and instructions are distributed across multiple dialogue turns. MultiChallenge measures instruction retention, inference memory, self-coherence, and related failures in text-based dialogue [62]. Audio MultiChallenge transfers these evaluation targets to end-to-end speech and introduces audio-specific challenges involving ambient and paralinguistic cues, as well as voice editing and spoken repair [93]. SpokenWOZ evaluates task-oriented spoken dialogue, including dialogue-state tracking, cross-turn slots, reasoning slots, and task completion [245]. ChildAgentEval compares agent capabilities with age-specific developmental measures without imposing real-time interaction constraints [241]. These benchmarks evaluate content quality and retention across turns, primarily under turn-based protocols without explicit real-time deadlines.

Static multimodal and spoken understanding. This family measures semantic understanding, reasoning, instruction following, and robustness over static or turn-complete input. OmniBench evaluates integrated image, audio, and text reasoning [33, 156]. VoiceBench evaluates knowledge, instruction following, safety, and robustness to variation in speaker, acoustic environment, and spoken content [43]. Dynamic-SUPERB and Dynamic-SUPERB Phase-2 provide broad instruction-conditioned evaluation across speech, music, and environmental sounds, covering classification, regression, and sequence-generation tasks [110, 111]. These benchmarks assess baseline capabilities that should be preserved when adding real-time interaction.

Streaming audiovisual and temporal intelligence. This family evaluates whether a model can understand events as a multimodal stream unfolds and associate observations with their temporal positions. StreamingBench evaluates real-time, omni-source, and contextual understanding at multiple points in a streaming video [165]. Charades-STA provides language-conditioned temporal localization, while Charades supplies temporally annotated activity understanding [86, 246]. RepCount and TransRAC provide fine-grained evaluation of repetitive-action counting and temporal structure [107]. Charades-STA, Charades, and RepCount assess event timing and repetitive-action counting. These static datasets assess what happened and when, but do not measure autonomous response timing.

7.2.2. Omni Interactivity Benchmarks

The omni interactivity branch evaluates when and how a model acts during an interaction. It focuses on latency, turn coordination, interruption handling, proactive behavior, and recovery from interaction failures.

Full-duplex spoken interaction. The Full-Duplex-Bench series evaluates simultaneous listening and speaking. FDB-v1 evaluates pause handling, backchanneling, turn taking, and interruption using controlled behavioral probes [160]. FDB-v1.5 extends the setting to overlapping speech, including user interruptions, listener backchannels, side conversations, and ambient speech [164]. FDB-v2 introduces an automated examiner, Fast and Slow pacing conditions, and multi-turn task families covering daily interaction, correction, entity tracking, and safety [163]. FDB-v3 adds multi-step tool use with real human audio and reports task accuracy, wall-clock latency, and turn-taking behavior [162]. HumDial-FDBench further contributes real dual-channel human conversations containing interruption, overlap, and feedback events [271]. M3-DuplexBench evaluates English and Japanese conversation and question answering with different amounts of dialogue history [82]. Instruct-FD evaluates whether systems follow explicit instructions about listening,

acknowledging, interrupting, and continuing speech [259]. Together, these resources provide the strongest existing coverage of latency, interaction timing, and overlap robustness, although their modality coverage remains primarily speech-based.

Streaming and proactive omni interaction. This family evaluates systems that operate over continuous audiovisual input and must decide whether and when to respond. OmniMMI combines streaming audiovisual understanding with proactive reasoning across multiple interaction tasks [285]. ProactiveVideoQA isolates the response-initiation decision and evaluates whether a model answers at an appropriate point during video playback [283]. Proact-VL evaluates proactive commentary over continuous game video, jointly scoring response timing and content across solo commentary, co-commentary, and user guidance [305]. StreamArena evaluates extended streaming interaction, including current perception, recall of earlier observations, proactive responses, and tool use [343]. OmniInteract combines online audiovisual input with measures of correctness, timing, interruption handling, invalid outputs, and context retention [176]. These benchmarks connect omni intelligence with interaction timing, but do not yet provide a complete standard for bidirectional spoken overlap, semantic repair, tool use, and long-session operation under a shared protocol.

Turn-taking prediction and evaluation. VAP, TurnGPT, and MuVAP provide reusable tasks and measures for assessing conversational timing within broader evaluations. VAP predicts future voice activity from two-channel audio and supports evaluation of turn shifts and backchannels [70]. TurnGPT predicts turn completion from textual dialogue context [69]. MuVAP extends future-activity projection to multimodal, multiparty settings using audio and visual speaker information [214]. These tasks and measures can be incorporated into broader omni-interaction evaluations.

Vendor and internal evaluations. TimeSpeak and CueSpeak are reported as internal Thinking Machines evaluations [262]. TimeSpeak examines time-triggered speech initiation, while CueSpeak evaluates responses triggered by semantic cues during ongoing speech. These vendor-reported evaluations are presented separately because the survey does not identify public datasets and protocols supporting independent reproduction.

7.3. Limitations of Existing Benchmarks

Current benchmarks cover different aspects of content quality, timing, and interaction behavior, but evaluating these capabilities together remains a challenge. The following proposals concern joint evaluation under common episodes and resource budgets, including where existing benchmarks provide partial coverage:

7.3.1. Joint Intelligence–Interactivity Evaluation

Existing static benchmarks can measure whether an answer is correct, while full-duplex benchmarks can measure whether a model yields, interrupts, or takes the floor. Recent benchmarks such as OmniMMI, ProactiveVideoQA, and OmniInteract begin to combine streaming understanding with online response decisions [176, 283, 285], but a joint protocol should evaluate continuous audio–video reasoning, bidirectional spoken overlap, semantic correction, and end-to-end serving latency using the same episodes and resource budget. One way to narrow the gap between offline replay and closed-loop evaluation is to augment recorded episodes with synthetic observations of plausible future states [30]. User simulation could also support closed-loop evaluation by generating user actions in response to the model's outputs, allowing assessment of how the model adapts to changes in user goals and knowledge [22]. A future protocol should evaluate the same system in at least four conditions: offline content evaluation, causal streaming perception, proactive response generation, and full-duplex interruption. Results should be reported as a Pareto frontier.

Proposed measurements: Evaluate content quality and interactive behavior together under the same episodes and resource budgets, reporting event-level timing and computational cost alongside capability scores. These measurements show how response quality changes under real-time constraints. Report accuracy at specified deadlines and define the input event and output endpoint for response onset and useful-answer time. A useful answer should satisfy a task-specific rubric; report failures to produce one separately. Define accuracy after interruption using the revised request, and replay timestamped episodes without access to future input.

7.3.2. Long-session Continuous Audio and Video

StreamingBench and OmniMMI evaluate bounded video streams and timed probes [165, 285]; FDB-v2 and Audio MultiChallenge evaluate multiple turns [93, 163]. These protocols are substantially shorter and more controlled than persistent meetings, tutoring sessions, assistive agents, embodied monitoring, and long-running task execution. StreamArena begins to probe hour-scale agentic streaming interaction, but remains video-centric rather than bidirectionally spoken [343]. Long-session evaluation must separate recent-scene perception from long-range memory, cap computation and memory growth, and ensure that the model remains an active participant throughout the session rather than answering only at the end.

Candidate metrics: entity and event retrieval as a function of elapsed time; contradiction rate; task-state recovery after long gaps; accuracy at fixed memory and compute budgets; latency drift over session duration. State the memory budget, compute limit, and session duration for each result.

7.3.3. Multi-party and Addressee-aware Full-duplex Interaction

Most end-to-end interaction benchmarks remain dyadic. MuVAP evaluates multiparty next-speaker prediction but is a predictive primitive rather than an interactive-agent benchmark [214]. HumDial-FDBench provides realistic full-duplex speech but does not establish a broad evaluation suite for three-or-more-speaker coordination [271].

Proposed measurements: Evaluate three or more participants with annotated speaker identity, addressee, and expected responder. Report addressee accuracy, responses to bystanders, speaker-conditioned interruption rates, and recovery from attribution errors. Measure response allocation as each participant's response rate among eligible addressed turns, accounting for differences in opportunities to respond.

7.3.4. Multilingual and Code-switching Full Duplex

Public full-duplex evaluation remains dominated by English, with limited coverage of other languages and little standardized evaluation of code-switching, low-resource languages, and language-dependent timing. M3-DuplexBench begins to address this by evaluating multi-turn full-duplex dialogue in English and Japanese [82], but broader language coverage and code-switching remain open. Static accent or multilingual understanding does not establish multilingual turn-taking competence.

Proposed measurements: Compare content accuracy and response timing by language. Score code-switch recognition against annotated language-switch events, and score recovery by whether the system follows the revised language instruction within a stated interval. Specify backchannel appropriateness in each language.

7.3.5. Repair, Cancellation, and Tool Use During Continuous Input

Audio MultiChallenge evaluates voice editing, FDB-v2 evaluates correction, and FDB-v3 evaluates disfluent multi-step tool use [93, 162, 163]. What remains missing is their combination: a user revises the request

while the model is speaking, a background tool or retrieval process is already running, and the system must cancel, update, or safely reuse partial work without emitting stale results.

Candidate metrics: correction-recovery rate and stale-content rate as defined above; cancellation success among cancellable operations; unnecessary tool execution and its cost; and elapsed time from the correction to the revised action. State the denominator and recovery deadline for each measure.

7.3.6. Controllable Interaction Policy

Existing benchmarks largely ask whether interaction is *natural*. They rarely ask whether it is *appropriate* for a specified role, or whether it can be changed on instruction. Instruct-FD evaluates instruction-conditioned turn management and finds a large gap between fluid and controllable behavior [259], while DuplexGen shows that the target turn-taking policy is itself scenario dependent [136].

Proposed measurements: instruction adherence for listening, acknowledging, interrupting, and continuing speech. Specify when the instruction changes and measure the delay until compliant behavior begins, using a stated observation interval and scenario-specific scoring rubric.

7.3.7. Safety, Privacy, and Deployment Robustness Under Continuous Input

Turn-based safety evaluation does not cover bystander speech, visible private information, environmental leakage, cross-channel prompt injection, packet loss, delayed frames, or safety actions that must proactively interrupt the user. A system may also appear safe only because network delay prevents it from responding in time.

Proposed measurements: define privacy leakage as unauthorized disclosure of specified private information captured in audio or video. Identify prohibited actions, required interventions, and intervention deadlines, then report missed and false interventions, disclosure rates, and errors under overlapping speech, controlled packet loss, and frame delay.

8. The Intelligence–Interactivity Trade-off

A central question is whether real-time speech and multimodal capabilities can be added while preserving the reasoning, factual knowledge, grounding, and tool use learned during language pretraining. Evaluations examine how speech and interaction objectives affect these capabilities [55, 270, 304].

The trade-off also concerns reliable use of those capabilities under real-time constraints and changing evidence. Reasoning, retrieval, grounding, and tool execution may remain incomplete after speech begins, while users may interrupt or revise the task [47, 48, 112]. Figure 14 distinguishes capability preservation during adaptation from response, evidence, and state coordination during runtime inference.

8.1. Capability Preservation under Real-Time Omni Adaptation

Evaluations identify capability losses in some models after speech adaptation. URO-Bench reports that several end-to-end spoken dialogue models lag behind their underlying language-model backbones in instruction following and exhibit catastrophic forgetting [304]. Related analysis attributes the text–speech understanding gap to two mechanisms: forgetting of textual capabilities during speech adaptation and misalignment between speech and text representations [55]. Parameter-importance analysis suggests that speech fine-tuning can alter parameters associated with textual reasoning in different layers [270]. The

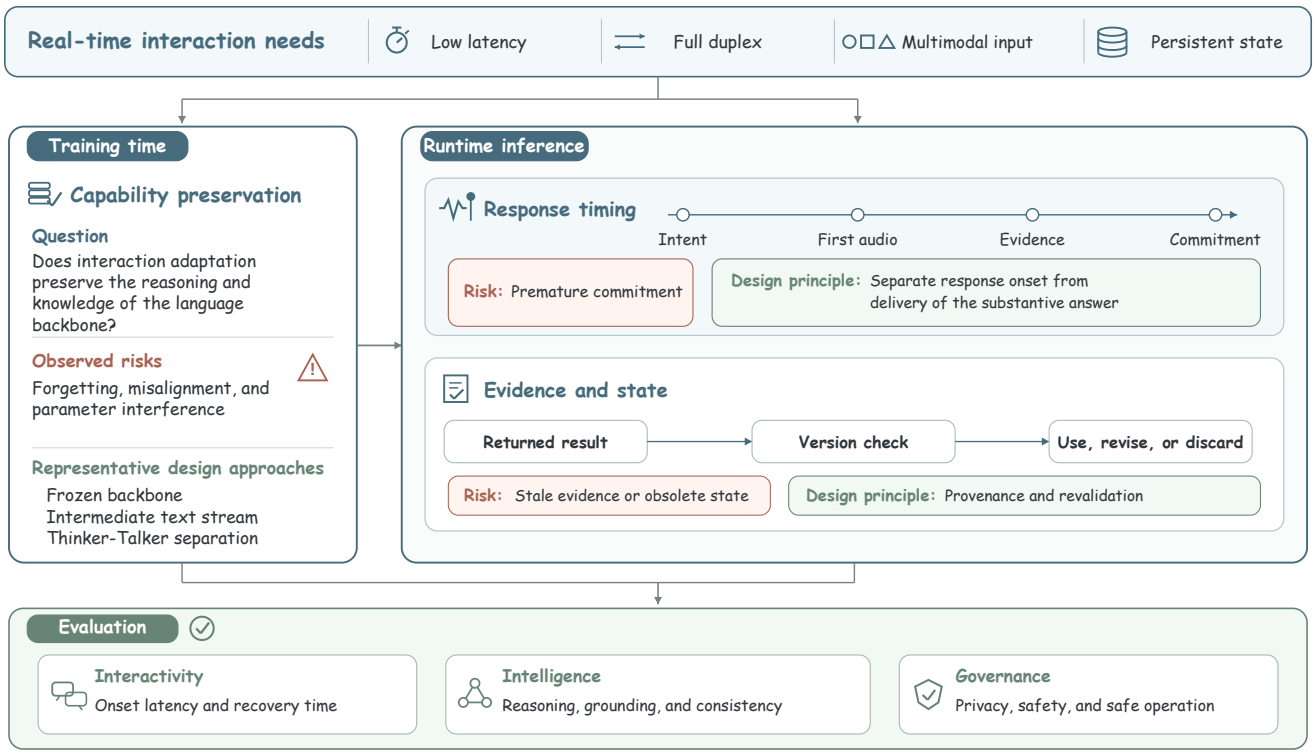

Figure 14: Trade-off between intelligence and interactivity in native real-time omni systems. Real-time interaction requires preserving pretrained capabilities during adaptation and coordinating responses, evidence, and state updates during inference. The diagram summarizes associated risks and representative design approaches.

effect depends on training data, optimization, modality interfaces, and whether the language backbone is updated. These findings do not establish that capability loss is inevitable.

A recurring architectural response is to separate interactive speech generation from the representations and computation used for language reasoning. Moshi retains an intermediate linguistic stream within its speech–text foundation model, keeping semantic processing explicit during real-time dialogue [61]. Freeze-Omni freezes the pretrained LLM backbone and attaches separate speech input and output modules [281]. Qwen3-Omni extends this separation through a Thinker–Talker architecture, in which multimodal understanding and reasoning are separated from streaming speech generation [302]. JoyAI-Talker combines a Thinker–Talker architecture with unified speech–text joint training that is designed to mitigate degradation in language understanding and reasoning performance while retaining performance on textual reasoning, STEM tasks, and logic, adding state-driven duplex control and expressive paralinguistic generation [20]. BayLing-Duplex adds listen, speak, and stop control tokens to a speech-language model and reports retained response quality [76].

These designs aim to preserve pretrained reasoning and factual knowledge during speech adaptation. Reported results are consistent with this principle, but they do not yet isolate the effect of individual design choices. Freeze-Omni reports spoken-input performance close to its frozen backbone's text-input performance, while Qwen3-Omni reports performance comparable to same-sized single-modal models across text, image, audio, and video tasks [281, 302]. In contrast, broader analyses of speech-adapted LLMs identify textual forgetting, cross-modal misalignment, and disruption of text-critical parameters [55, 270]. Because these studies use different backbones, training data, model scales, and evaluation settings, controlled comparisons

are still needed to determine when backbone freezing, constrained adaptation, or explicit reasoning–speech separation best preserves the capabilities of the original language model.

Interaction-policy optimization can also reduce response quality unless its objectives explicitly preserve semantic content. Multi-Faceted Interactivity Alignment uses separate rewards for pause handling, turn-taking, backchanneling, and user interruption, together with a response-quality reward [201]. Capability losses can therefore occur both when adding speech during adaptation and when optimizing interaction timing without preserving response quality.

8.2. Runtime Trade-offs

8.2.1. Response Timing, Reasoning, and Grounding

A model can start responding before reasoning, retrieval, or tool execution is complete. We distinguish *response-onset latency* from *useful-answer time*, when task-relevant information first reaches the user, using the reporting requirements in Section 7.1. Systems that overlap speech with reasoning, retrieval, or tool execution illustrate why these endpoints differ [11, 47, 48, 112].

Early speech is not necessarily a failure. Brief acknowledgments and backchannels can preserve conversational flow while the system continues to reason, and full-duplex benchmarks distinguish these behaviors from interruptions, side conversations, and ambient speech [164]. The problem arises when early output implies that the answer is already grounded, states unsupported content, or commits to an interpretation that the user later corrects. Full-Duplex-Bench-v2 directly evaluates correction and entity-tracking tasks and reports persistent difficulties under overlapping multi-turn interaction [163]. In tool-augmented settings, a query issued before the user's request is sufficiently clear may retrieve irrelevant information; recent work formalizes this issue as tool-intent stabilization [11, 83]. Systems should distinguish early acknowledgments from factual answers and validate later claims or tool results against the current task.

8.2.2. Evidence and State Coordination

Observations and returned results differ in source, timing, reliability, and relevance. These include user and background speech, video, partial transcripts, retrieval outputs, tool results, and memory. This creates three recurring risks: source confusion, in which incidental input is treated as user-directed evidence; temporal misalignment, in which information is attached to the wrong moment; and stale evidence, in which a result associated with an earlier task state is used after the user's intent has changed. Static multimodal evaluation shows that irrelevant modalities can distort model decisions, while full-duplex and streaming benchmarks expose additional difficulties in overlap handling and online visual understanding [28, 164, 165]. Recent omni systems accordingly treat temporal alignment and synchronized multimodal processing as central architectural requirements [56, 263]. Reliable interaction therefore requires evidence to remain associated with its source, time, and current task relevance.

Longer sessions require consistent tracking of people, objects, events, goals, corrections, tool calls, and superseded tasks. Long-context and memory benchmarks show that relevant information may be overlooked and that sustained interaction requires temporal reasoning, knowledge updates, and consistency across sessions [172, 181, 291]. Recent spoken and audio-visual benchmarks explicitly extend evaluation to correction handling, entity tracking, and multimodal multi-turn memory [163, 264]. Memory should retain each observation's source, modality, and time, support consistent entity identification, and allow task updates. Retrieved evidence must be checked for relevance to the current task version [63, 291]. Together, these runtime risks show that interactivity depends on coordinating when evidence becomes valid and which version of the

Table 11: How real-time interaction requirements affect capability preservation and reliable operation. Each interactivity requirement links to the capabilities or coordination mechanisms it depends on and the failures that may occur.

Interactivity requirement	Capability or coordination requirement	Typical manifestation
Speech and omni adaptation	Preserving reasoning and textual capabilities [55, 304]	Capability degradation
Interaction-timing alignment	Preserving response content while optimizing timing [201]	Reduced response quality
Immediate response	Completing deliberation and grounding before commitment [48, 112]	Premature or unsupported answers
Continuous multimodal input	Selecting source-, time-, and task-relevant evidence [164, 165]	Source confusion, temporal misalignment, or stale evidence
Persistent interaction	Maintaining revisable entity and task states [181, 291]	Contradictory or obsolete state
Continuous sensing	Use captured information only as permitted [206, 334]	Unauthorized disclosure or use of sensitive data

interaction it belongs to. Table 11 summarizes these recurring forms of the intelligence–interactivity trade-off.

8.3. Safety, Privacy, and Evaluation Implications

Continuous interaction requires rules for capturing, processing, transmitting, and storing information that may not be addressed to the system, including bystander speech and private visual content. Audio and audio–visual safety benchmarks show that unsafe intent can be expressed through speech and distributed across modality combinations, while selective-hearing evaluation reveals substantial leakage of incidental bystander speech [206, 209, 253, 334]. Permission checks must apply to each information-handling operation and to any action it triggers. These boundaries may require different decisions for the same input: sensitive bystander speech may need to be detected locally for redaction without being transmitted to an external model or written into persistent memory. Privacy risk also extends to the model's internal state. Analysis of full-duplex systems such as SALM-Duplex and Moshi finds that speaker identity can become easier to infer from hidden states as a conversation progresses. Streaming anonymization can reduce this exposure while adding delay or reducing task performance [139]. Privacy in continuous interaction therefore concerns not only stored media but also what persistent internal representations reveal about the user.

Continuous sensing can improve contextual understanding, but it also requires the system to track where information comes from, whether it remains relevant, and how it may be used, including permission checks. Evaluation therefore needs to assess both whether observations and inferences are used for authorized purposes and whether the system intervenes at appropriate times. These requirements for handling information and actions must also be evaluated.

Existing benchmarks remain fragmented: full-duplex evaluations focus on turn-taking and spoken repair, streaming benchmarks evaluate online perception, long-term memory benchmarks test retrieval and consistency across extended histories, and multimodal safety benchmarks assess jailbreak, refusal, or privacy risks [163–165, 181, 206, 291, 334]. First-token latency cannot reveal when useful information becomes available; final-answer accuracy cannot reveal whether stale or misattributed evidence was used; and final refusal cannot reveal whether sensitive information was already transmitted or stored. Evaluate response timing, reasoning quality, evidence validity, state consistency, privacy exposure, and tool or memory actions

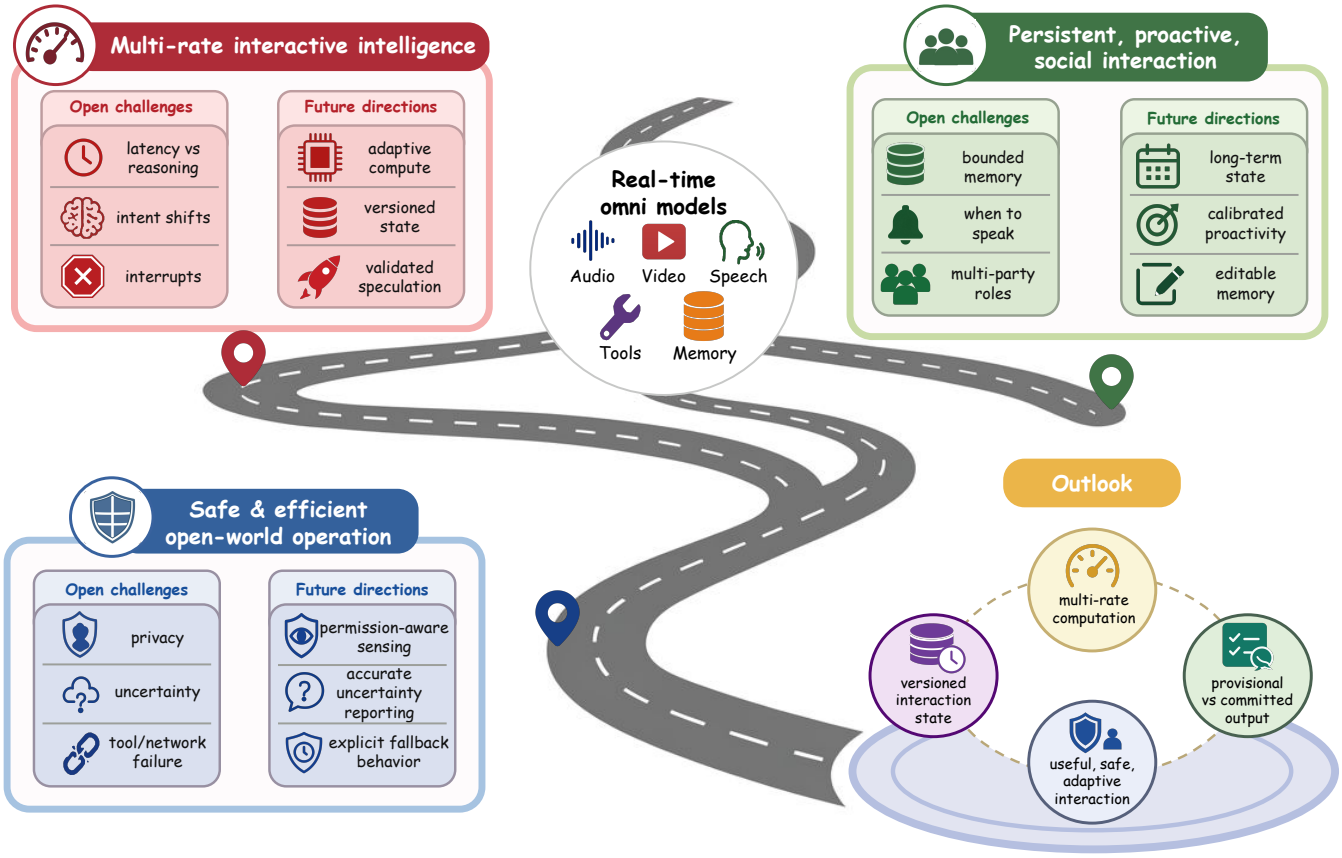

Figure 15: Open problems and outlook. Three directions for native real-time omni models: coordinating interaction and reasoning across different computational rates; supporting persistent, proactive interaction that respects participants' roles, intentions, and preferences; and ensuring safe and efficient operation in open-world settings. Each direction connects current challenges with corresponding research priorities.

together. Event-level records can show their trade-offs more clearly than a single aggregate score.

9. Design Principles and Open Problems

Native real-time omni models must coordinate perception, reasoning, generation, and action as new information arrives. Remaining challenges concern maintaining these capabilities jointly within latency and resource limits [56, 112, 176, 340]. We organize the research agenda around computation at different time scales; persistent, proactive, and social interaction; and safe, efficient operation under changing deployment conditions. Figure 15 summarizes these directions.

9.1. Multi-Rate Interactive Intelligence

Immediate interaction and extended reasoning have different timing requirements. Turn-taking, interruption handling, and short acknowledgments require prompt responses, whereas retrieval, verification, planning, and tool execution may take longer. A system can coordinate multiple models or internal streams, each with different compute, context, and output constraints [11, 48, 112, 292, 340].

Recent streaming systems suggest that this coordination is not a single architectural pattern but a family of

mechanisms. *Asynchronous parallelism* runs foreground interaction and background computation concurrently, as in StreamMind, which separates latency-critical interaction from persistent multimodal recall and longer-horizon processing [343]. *Conditional computation* invokes deeper reasoning only when it is needed: R3-Streaming compresses memory, assesses response readiness, and routes only difficult queries toward stronger reasoning [170]. *Perception–reasoning overlap* runs reasoning while new evidence arrives, as in Video Streaming Thinking, which processes incoming clips during video playback [96]. *Multi-rate state evolution* consolidates memory on a slower timescale than immediate perception, as in LatentStream's progressive latent memory [219]. *Decoupled interaction timing* separates content preparation from floor-taking, as in contextual preface generation paired with a Voice Activity Projection onset decision [202]. These mechanisms must share a consistent interaction state even when their inputs come from different task versions.

Background work may begin from an incomplete request and finish after the task changes. Associating each operation with its originating task version allows returned results to be checked against the current request, outdated work to be canceled where possible, and reusable intermediate results to be retained. Speculative retrieval and tool use can reduce delay when the relevant intent stabilizes sufficiently early [83, 103]. Recovery also depends on what the operation has done: computation can often be discarded or repeated, whereas an external action that cannot be undone may require a compensating action. Coordinating this background work with ongoing interaction creates additional scheduling demands. Frequent input updates, persistent state, asynchronous results, and cancellation must be handled alongside decoding. Disaggregated execution [348], latency-aware serving [4], and interaction-aware scheduling [346] provide mechanisms for allocating computation. When omni systems generate visual output, component-level acceleration can complement scheduling: rectified-flow auto-coding and optimized flow inference reduce generation cost [77, 78], although integrating these methods with real-time interaction scheduling remains open. Efficient scheduling must also account for the current task: even a quickly computed result may no longer be useful if it arrives after it is needed, relies on outdated evidence, or has been invalidated by a correction.

9.2. Persistent, Proactive, and Social Interaction

Persistent memory should distinguish current and superseded information, retain unresolved tasks, and record the evidence supporting stored conclusions. Streaming attention, compression, sparse prefill, and cache retrieval reduce computational growth [63, 126, 193, 296]. Memory must also separate observations from inferences, preserve source and time, and support bounded storage, revision, and deletion [181, 291].

Proactive responses should use current observations, task requirements, and user instructions. The system must decide whether an event warrants a response, whether the user is available, and whether waiting would be more helpful. Recent work has begun to model proactive response from streaming sensory context and to evaluate both response content and timing [56, 176, 283, 306]. Proact-VL treats response triggering as part of the interactive policy over an ongoing video stream and evaluates timing and content jointly [305], while Instruct-FD and DuplexGen indicate that the appropriate proactive and turn-taking policy is role- and scenario-dependent and may need to be set by instruction rather than learned as a single fixed prior [136, 259]. When intent is uncertain, the model should retain alternative interpretations and defer consequential actions until sufficient evidence is available. Evaluation should distinguish helpful interventions, missed opportunities, unnecessary interruptions, and appropriate decisions to wait.

These decisions become harder in multi-party and personalized settings. Speaker identity, addressee, conversational floor, and authority are related but distinct. A person may be speaking near the system without addressing it, and the person addressing the system may not be authorized to trigger an action. Existing full-duplex data and turn-taking models only begin to capture overlapping speakers, repair, and floor com-

petition [214, 271]. Personalization should distinguish temporary behavior from lasting preferences and allow users to inspect, correct, or delete stored information. Systems also need explicit boundaries between shared and private participant context. Systems such as Ex-Omni-2D generate synchronized visual output alongside personalized speech [341], introducing additional questions about attention and social timing. The central challenge is to maintain one coherent interaction while preserving the distinct roles, permissions, and histories of its participants.

9.3. Safe and Efficient Operation in Open-World Settings

Continuous perception requires safety checks throughout the interaction, including for incoming audio and video. Audio and video may contain bystander speech, private screens, or environmental details that are available to the model but not permissible to use. Systems must enforce permissions for capture, processing, transmission, storage, and action. Selective-hearing results show that attending to the intended speaker does not by itself prevent leakage from surrounding speech [334]. Analogous leakage arises from internal state: speaker identity can remain recoverable from the hidden states of full-duplex models and grow more linkable across the first conversational turns, and streaming anonymization trades latency and utility to reduce it [139]. A specific open problem is preserving useful context without retaining unnecessary biometric, identity, or other sensitive attributes. Permission checks also apply to visual evidence, retrieved records, and tool outputs. Safety must also govern action timing. A provisional interpretation may be sufficient to start local computation, but not to send a message, disclose information, or control a device. SafeDreamer uses a learned world model to predict rewards and safety costs for candidate action sequences and incorporates cost constraints into action selection [113]. Greater uncertainty or more consequential actions may require additional confirmation or oversight, including monitoring for high-impact operations [14].

Reliable systems need explicit fallback behavior when network, compute, sensors, or tools degrade. Real-time behavior depends on device memory, network delay, server load, sensor quality, and the availability of external tools. When these conditions deteriorate, the system may need to reduce visual sampling, postpone background reasoning, route computation between edge and cloud, or fall back to a more limited interaction mode. Edge-capable omni models and interaction-aware serving show that such choices can be incorporated into system design [56, 346], but fallback behavior should communicate missing information or delays and preserve interaction state. Dropping a modality should lower confidence where that evidence matters, delaying a tool should not be hidden behind an unsupported answer, and a connectivity failure should not erase the interaction state. The open problem is to maintain causal input processing, accurate uncertainty reporting, interruption handling, and action safeguards when full capability is temporarily unavailable.

Training data and evaluations should include these failure conditions. Natural interaction data should include overlap, hesitation, correction, cancellation, silence, and cases in which a proactive response or tool call should be withheld [93, 271, 349]. Evaluations should replay timestamped streams without access to future input and compare systems under controlled resource and network conditions. They should record when intent becomes clear, when useful information is delivered, which evidence is used, and how tool and memory state change. The protocol should define how intent clarity and useful information are annotated. Recent benchmarks increasingly cover disfluent tool use and online audio-visual interaction [162, 176], but they do not yet combine long-session state, multi-party authority, privacy boundaries, and deployment stress in one setting. The resulting systems should be compared on a joint frontier of answer quality, informative latency, repair, cost, and safety rather than reduced to a single aggregate score.

9.4. Outlook

Both integrated and modular architectures can support computation at different time scales. Beyond these architectures, an emerging direction is to pair generative omni models with compact, JEV-like decision modules for streaming and latency-critical control. PlayJev provides an early example by mapping raw visual observations and candidate actions directly to real-time action probabilities, although extending this approach from GUI games to open-ended multimodal interaction remains an open problem [203].

The key requirements are consistent task versions, revisable intermediate reasoning, memory with traceable sources, appropriate decisions to wait, and actions that remain valid after interruptions. Progress should be measured by preserved reasoning and grounding during continuous input, consistent state across sessions and participants, and adaptive computation with accurate uncertainty reporting and privacy protection. Evaluate usefulness, time to a useful answer, recovery from errors, resource cost, and safe action together. The goal is a system that remains useful and correct as evidence and user intent change during its response.

10. Conclusion

Real-time interaction changes the conditions under which multimodal intelligence must operate. A model must reason from incomplete input, decide when to respond, and remain receptive to evidence that may change its interpretation or invalidate an ongoing action. This survey examined how these demands connect multimodal representations, model architectures, training objectives, persistent state, and serving systems. Together, they motivate Interactive Omni Intelligence: the ability to sustain understanding, reasoning, and generation as the world and the conversation continue to unfold. Our conceptual framework comprises two coupled components: Omni Intelligence and Omni Interactivity. The Intelligence–Interactivity Trade-off characterizes how the demands of timely, continuous interaction constrain the multimodal capabilities a system can effectively deploy. Together, the framework, design analysis, and evaluation agenda provide a roadmap for native real-time Interactive Omni Intelligence.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

[2] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2019. doi: 10.1109/JSTSP.2018.2885636. URL https://arxiv.org/abs/1807.00129.

[3] Amey Agrawal, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, and Ramachandran Ramjee. Sarathi: Efficient llm inference by piggybacking decodes with chunked prefills. *arXiv preprint arXiv:2308.16369*, 2023. URL https://arxiv.org/abs/2308.16369.

[4] Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav Gulavani, Alexey Tumanov, and Ramachandran Ramjee. Taming Throughput-Latency tradeoff in LLM inference with Sarathi-Serve. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 117–134, Santa Clara, CA, July 2024. USENIX Association. ISBN 978-1-939133-40-3. URL https://www.usenix.org/conference/osdi24/presentation/agrawal.

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736, 2022. URL https://doi.org/10.52202/068431-1723.

[6] Alibaba Cloud. Qwen3.5-Omni-Realtime: Real-time omni-modal interaction. Alibaba Cloud Model Studio Documentation, 2026. URL https://docs.modelstudio.console.alibabacloud.com/en/model-studio/realtime.

[7] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-tts: A family of high-quality versatile speech generation models, 2024. URL https://arxiv.org/abs/2406.02430.

[8] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520/.

[9] Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. On the landscape of spoken language models: A comprehensive survey. *Transactions on Machine Learning Research*, 2025. URL https://mlanthology.org/tmlr/2025/arora2025tmlr-landscape/.

[10] Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. Talking turns: Benchmarking audio foundation models on turn-taking dynamics. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/82f68b38747c406672f7f9f6bab86775-Abstract-Conference.html.

[11] Siddhant Arora, Haidar Khan, Kai Sun, Xin Luna Dong, Sajal Choudhary, Seungwhan Moon, Xinyuan Zhang, Adithya Sagar, Surya Teja Appini, Kaushik Patnaik, Sanat Sharma, Shinji Watanabe, Anuj Kumar, Ahmed Aly, Yue Liu, Florian Metze, and Zhaojiang Lin. Stream RAG: Instant and accurate spoken dialogue systems with streaming tool usage. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2510.02044.

[12] Siddhant Arora, Jinchuan Tian, Jiatong Shi, Hayato Futami, Yosuke Kashiwagi, Emiru Tsunoo, and Shinji Watanabe. Optimizing conversational quality in spoken dialogue systems with reinforcement learning from AI feedback. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 41049–41066, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.2040. URL https://aclanthology.org/2026.findings-acl.2040/.

[13] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/25f7be9694d7b32d5cc670927b8091e1-Abstract-Conference.html.

[14] Chloe Autio, Reva Schwartz, Jesse Dunietz, Shomik Jain, Martin Stanley, Elham Tabassi, Patrick Hall, and Kamie Roberts. Artificial intelligence risk management framework: Generative artificial intelligence profile. NIST AI 600-1, National Institute of Standards and Technology, Gaithersburg, MD, July 2024. URL https://doi.org/10.6028/NIST.AI.600-1.

[15] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_rylwJxrYDS.html.

[16] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

[17] Detao Bai, Shimin Yao, Weixuan Chen, Chengen Lai, Yuanming Li, Zhiheng Ma, and Xihan Wei. OmniEncoder: See, hear, and feel continuous motion like humans with one encoder. *arXiv preprint arXiv:2605.01506*, 2026. URL https://arxiv.org/abs/2605.01506.

[18] He Bai, Tatiana Likhomanenko, Ruixiang Zhang, Zijin Gu, Zakaria Aldeneh, and Navdeep Jaitly. dMel: Speech tokenization made simple. *arXiv preprint arXiv:2407.15835*, 2024. URL https://arxiv.org/abs/2407.15835.

[19] Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, et al. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*, 2024. URL https://arxiv.org/abs/2405.16640.

[20] Yinhao Bai, Jinming Chen, Yafeng Chen, Wei Deng, Boya Dong, Nan Duan, Yu Gu, Weisheng Han, Yankun Huang, Ming Ke, Hao Li, Jingdong Li, Xiangyu Liang, Ning Liu, Yuan Liu, Ji Miao, Jiaqi Wang, Qi Wang, Wenchao Wang, Yuxuan Wang, Zhenfang Wang, Zhangyu Xiao, Chao Xue, Hongfei Xue, Fan Yu, Tianyi Zhang, Yuan Zhang, Yuqi Zhang, and Lin Zhu. Joyai-talker: Full-duplex speech interactive large model built for empathetic voice agents, 2026. URL https://arxiv.org/abs/2608.01119.

[21] Krisztian Balog and ChengXiang Zhai. User simulation for evaluating information access systems. *Foundations and Trends in Information Retrieval*, 18(1–2):1–261, 2024. doi: 10.1561/1500000098. URL https://doi.org/10.1561/1500000098.

[22] Krisztian Balog and ChengXiang Zhai. The indispensable role of user simulation in the pursuit of AGI. *Communications of the ACM*, 69(5):45–48, 2026. doi: 10.1145/3768170. URL https://doi.org/10.1145/3768170.

[23] Pratyay Banerjee, Masud Moshtaghi, Shivashankar Subramanian, Amita Misra, and Ankit Chadha. APEX-MEM: Agentic semi-structured memory with temporal reasoning for long-term conversational AI. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16470–16489, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.749. URL https://aclanthology.org/2026.acl-long.749/.

[24] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

[25] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409. URL https://doi.org/10.1109/TASLP.2023.3288409.

[26] ByteDance Seed. SeedRealtime audio-visual full-duplex LLM released: Toward omni-modal natural interaction. ByteDance Seed Blog, August 2026.

[27] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.807. URL https://aclanthology.org/2024.findings-acl.807/.

[28] Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models, 2026. URL https://arxiv.org/abs/2505.19616.

[29] Wei Cao, Chang Luo, Biao Zhang, Matthias Nießner, and Jiapeng Tang. Motion2VecSets: 4d latent vector set diffusion for non-rigid shape reconstruction and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20496–20506, June 2024. doi: 10.1109/CVPR52733.2024.01937. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cao_Motion2VecSets_4D_Latent_Vector_Set_Diffusion_for_Non-rigid_Shape_Reconstruction_CVPR_2024_paper.html.

[30] Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In Joseph Lim, Shuran Song, and Hae-Won Park, editors, *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 4709–4722. PMLR, 2025. URL https://proceedings.mlr.press/v305/cao25a.html.

[31] Wei Cao, Hao Zhang, Fengrui Tian, Yulun Wu, Yingying Li, Shenlong Wang, Ning Yu, and Yaoyao Liu. FreeOrbit4D: Training-free arbitrary camera redirection for monocular videos via foreground-complete 4D reconstruction. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2026. doi: 10.1145/3799902.3811122. URL https://doi.org/10.1145/3799902.3811122.

[32] Xu Cao. Enhancing video vision language model with hippocampal sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 33682–33692, 2026.

[33] Xu Cao, Yifan Shen, Bolin Lai, Wenqian Ye, Yunsheng Ma, Joerg Heintz, Jintai Chen, Meihuan Huang, Jianguo Cao, Aidong Zhang, and James M. Rehg. What is the visual cognition gap between humans and multimodal LLMs? In *Proceedings of the Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=78lTuD6wiO.

[34] Galo Castillo-López, Gael de Chalendar, and Nasredine Semmar. A survey of recent advances on turn-taking modeling in spoken dialogue systems. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 254–271, Bilbao, Spain, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-248-0. URL https://aclanthology.org/2025.iwsds-1.27/.

[35] Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid?, 2026. URL https://arxiv.org/abs/2605.06527.

[36] Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, page 3670–3674. ISCA, August 2021. doi: 10.21437/interspeech.2021-1965. URL http://dx.doi.org/10.21437/Interspeech.2021-1965.

[37] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. VideoLLM-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18407–18418, June 2024. doi: 10.1109/CVPR52733.2024.01742. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_VideoLLM-online_Online_Video_Large_Language_Model_for_Streaming_Video_CVPR_2024_paper.html.

[38] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16 (6):1505–1518, October 2022. ISSN 1941-0484. doi: 10.1109/jstsp.2022.3188113. URL http://dx.doi.org/10.1109/JSTSP.2022.3188113.

[39] Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text-to-speech synthesizers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 33:705–718, 2025. doi: 10.1109/TASLPRO.2025.3530270. URL https://doi.org/10.1109/TASLPRO.2025.3530270.

[40] Sanyuan Chen, Min-Jae Hwang, Sho Inoue, Anna Sun, Bokai Yu, David Kant, Dongmin Hyun, Dorian Desblancs, Gregory Antonovsky, Oleg Repin, Peng-Jen Chen, Xutai Ma, Zehai Tu, Juan Pino, and Wei-Ning Hsu. Alignment-free text-audiobox for voice dubbing and full-duplex dialogue synthesis, 2026. URL https://arxiv.org/abs/2609.03992.

[41] Tao Chen, Kun Zhang, Qiong Wu, Xiao Chen, Chao Chang, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Scaling the long video understanding of multimodal large language models via visual memory mechanism, 2026. URL https://arxiv.org/abs/2603.29252.

[42] Tuochao Chen, Bandhav Veluri, Hongyu Gong, and Shyamnath Gollakota. AV-dialog: Spoken dialogue models with audio-visual input. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 42208–42225, San Diego, California, United States, July 2026. Association for

Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1954. URL https://aclanthology.org/2026.acl-long.1954/.

[43] Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398, 2026. doi: 10.1162/tacl.a.628. URL https://aclanthology.org/2026.tacl-1.18/.

[44] Yuxuan Chen and Haoyuan Yu. From turn-taking to synchronous dialogue: A survey of full-duplex spoken language models. *arXiv preprint arXiv:2509.14515*, 2025. URL https://arxiv.org/abs/2509.14515.

[45] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *ECAI 2025*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pages 2993–3000. IOS Press, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

[46] Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. SHANKS: Simultaneous hearing and thinking for spoken language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8951–8972, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.404. URL https://aclanthology.org/2026.acl-long.404/.

[47] Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. STITCH: Simultaneous thinking and talking with chunked reasoning for spoken language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=5Z1eMhCeTb.

[48] Chung-Ming Chien, Manu Orsini, Eugene Kharitonov, Neil Zeghidour, Karen Livescu, and Alexandre Défossez. MoshiRAG: Asynchronous knowledge retrieval for full-duplex speech language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/66336.

[49] Ju-Chieh Chou, Chung-Ming Chien, Wei-Ning Hsu, Karen Livescu, Arun Babu, Alexis Conneau, Alexei Baevski, and Michael Auli. Toward joint language modeling for speech units and text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6582–6593, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.438. URL https://aclanthology.org/2023.findings-emnlp.438/.

[50] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

[51] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech 2018*, page 1086–1090. ISCA, Sept 2018. doi: 10.21437/interspeech.2018-1929. URL http://dx.doi.org/10.21437/Interspeech.2018-1929.

[52] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech

pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE, 2021. doi: 10.1109/ASRU51503.2021.9688253. URL https://arxiv.org/abs/2108.06209.

[53] Christopher Cieri, David Miller, and Kevin Walker. The fisher corpus: a resource for the next generations of speech-to-text. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL https://aclanthology.org/L04-1500/.

[54] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 47704–47720. Curran Associates, Inc., 2023. doi: 10.52202/075280-2066. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/94b472a1842cd7c56dcb125fb2765fbd-Abstract-Conference.html.

[55] Santiago Cuervo, Skyler Seto, Maureen de Seyssel, Richard He Bai, Zijin Gu, Tatiana Likhomanenko, Navdeep Jaitly, and Zakaria Aldeneh. Closing the gap between text and speech understanding in LLMs. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=dDHnO3Vhyj.

[56] Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction, 2026. URL https://arxiv.org/abs/2604.27393.

[57] Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and Irwin King. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970, Vienna, Austria, July 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.acl-long.682/.

[58] Wenqian Cui, Lei Zhu, Xiao-Hui Li, Zhihan Guo, Haoli Bai, Lu Hou, and Irwin King. TurnGuide: Enhancing meaningful full duplex spoken interactions via dynamic turn-level text-speech interleaving. In *Interspeech*, 2026. URL https://arxiv.org/abs/2508.07375.

[59] DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. URL https://arxiv.org/abs/2412.19437.

[60] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=ivCd8z8zR2.

[61] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *CoRR*, abs/2410.00037, 2024. doi: 10.48550/ARXIV.2410.00037. URL https://doi.org/10.48550/arXiv.2410.00037.

[62] Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18632–18702, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

[63] Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, ZhongTao, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 42115–42127, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/67a9b444cbcd647572c88194619f72d5-Paper-Conference.pdf.

[64] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 52545–52567, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5e2217482fa75556f1970be809acd3f8-Abstract-Conference.html.

[65] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.183. URL https://aclanthology.org/2023.emnlp-main.183/.

[66] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[67] Xuan Du, Qiangyu Yan, Wenshuo Li, Borui Jiang, Changming Xiao, Han Shu, and Xinghao Chen. Thinking-while-speaking: A controlled, interleaved reasoning method for real-time speech generation. *arXiv preprint arXiv:2605.20946*, 2026. URL https://arxiv.org/abs/2605.20946.

[68] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. Cosyvoice 2: Scalable streaming speech synthesis with large language models, 2024. URL https://arxiv.org/abs/2412.10117.

[69] Erik Ekstedt and Gabriel Skantze. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.268. URL https://aclanthology.org/2020.findings-emnlp.268/.

[70] Erik Ekstedt and Gabriel Skantze. Voice activity projection: Self-supervised learning of turn-taking events. In *Interspeech 2022*, pages 5190–5194, 2022. doi: 10.21437/Interspeech.2022-10955. URL https://www.isca-archive.org/interspeech_2022/ekstedt22_interspeech.html.

[71] Erik Ekstedt and Gabriel Skantze. How much does prosody help turn-taking? investigations using voice activity projection models. In Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek, editors, *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 541–551, Edinburgh, UK, September 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.sigdial-1.51. URL https://aclanthology.org/2022.sigdial-1.51/.

[72] Nikolaos Ellinas, Georgios Vamvoukakis, Konstantinos Markopoulos, Aimilios Chalamandaris, Georgia Maniati, Panos Kakoulidis, Spyros Raptis, June Sig Sung, Hyoungmin Park, and Pirros Tsiakoulis. High quality streaming speech synthesis with low, sentence-length-independent latency. In *Proceedings of Interspeech 2020*, pages 2022–2026, 2020. doi: 10.21437/Interspeech.2020-2464. URL https://www.isca-archive.org/interspeech_2020/ellinas20_interspeech.html.

[73] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):1–11, July 2018. ISSN 1557-7368. doi: 10.1145/3197517.3201357. URL http://dx.doi.org/10.1145/3197517.3201357.

[74] Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. LLaMA-Omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PYmrUQmMEw.

[75] Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. LLaMA-Omni 2: LLM-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18617–18629, Vienna, Austria, July 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.acl-long.912/.

[76] Qingkai Fang, Shoutao Guo, and Yang Feng. Bayling-duplex: Native full-duplex speech dialogue with a single autoregressive llm, 2026. URL https://arxiv.org/abs/2606.14528.

[77] Sen Fang, Yalin Feng, Yanxin Zhang, and Dimitris N. Metaxas. Rac: Rectified flow auto coder, 2026. URL https://arxiv.org/abs/2603.05925.

[78] Sen Fang, Hongbin Zhong, Yalin Feng, Yanxin Zhang, and Dimitris N. Metaxas. Streamflow: Theory, algorithm, and implementation for high-efficiency rectified flow generation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=pvaviwbsry.

[79] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://jmlr.org/papers/v23/21-0998.html.

[80] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, Haoyu Cao, Di Yin, Long Ma, Xiawu Zheng, Rongrong Ji, Yunsheng Wu, Ran He, Caifeng Shan, and Xing Sun. VITA: Towards open-source interactive omni multimodal LLM, 2024. URL https://arxiv.org/abs/2408.05211.

[81] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/

085713-2529. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6ce9d51dded7dac82b3b4d3dbb1d73bc-Abstract-Conference.html.

[82] Ryo Fukuda, Atsushi Ando, Hiroki Kanagawa, Takatomo Kano, Marc Delcroix, Naohiro Tawara, and Yuya Chiba. M3-duplexbench: A multi-turn, multilingual, multidomain benchmark for full-duplex spoken dialogue models, 2026. URL https://arxiv.org/abs/2607.29125.

[83] Elroy Galbraith. When does streaming tool use help? characterizing tool-intent stabilization in streaming retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2606.20113.

[84] Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. MegaBlocks: Efficient sparse training with mixture-of-experts. In *Proceedings of Machine Learning and Systems*, volume 5, pages 288–304, 2023. URL https://proceedings.mlsys.org/paper_files/paper/2023/hash/5a54f79333768effe7e8927bcccffe40-Abstract-mlsys2023.html.

[85] Daniel Galvez, Greg Diamos, Juan Torres, Keith Achorn, Juan Cerón, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/202cb962ac59075b964b07152d234b70-Abstract-round1.html.

[86] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5277–5285. IEEE, October 2017. doi: 10.1109/ICCV.2017.563. URL https://doi.org/10.1109/ICCV.2017.563.

[87] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

[88] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. URL https://arxiv.org/abs/2312.10997.

[89] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. doi: 10.48550/arXiv.2312.11805. URL https://arxiv.org/abs/2312.11805.

[90] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE, 1992. doi: 10.1109/ICASSP.1992.225858. URL https://doi.org/10.1109/ICASSP.1992.225858.

[91] Jingyao Gong. MiniMind-O technical report: An open small-scale speech-native omni model. *arXiv preprint arXiv:2605.03937*, 2026. URL https://arxiv.org/abs/2605.03937. Code: https://github.com/jingyaogong/minimind-o.

[92] Google. Gemini live. Google Gemini product page, August 2024. URL https://gemini.google/overview/gemini-live/. Gemini Live began rolling out on August 13, 2024.

[93] Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, and Yunzhong He. Audio MultiChallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35740–35770, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1654. URL https://aclanthology.org/2026.acl-long.1654/.

[94] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

[95] Alex Graves. Sequence transduction with recurrent neural networks. In *ICML 2012 Workshop on Representation Learning*, 2012. URL https://arxiv.org/abs/1211.3711. arXiv:1211.3711.

[96] Yiran Guan, Liang Yin, Dingkang Liang, Jianzhong Ju, Zhenbo Luo, Jian Luan, Yuliang Liu, and Xiang Bai. Video streaming thinking: Videollms can watch and think simultaneously, 2026. URL https://arxiv.org/abs/2603.12262.

[97] Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. Recent advances in discrete speech tokens: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4184–4204, April 2026. doi: 10.1109/TPAMI.2025.3643619. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3643619.

[98] Yuren Hao, Shuhaib Mehri, ChengXiang Zhai, and Dilek Hakkani-Tür. User preference modeling for conversational LLM agents: Weak rewards from retrieval-augmented interaction, 2026. URL https://arxiv.org/abs/2603.20939.

[99] Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. Textually pretrained speech language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c859b99b5d717c9035e79d43dfd69435-Abstract-Conference.html.

[100] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE, 2024. doi: 10.1109/SLT61566.2024.10832365. URL https://arxiv.org/abs/2407.05361.

[101] Shawn Hershey, Daniel P. W. Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R. Channing Moore, and Manoj Plakal. The benefit of temporally-strong labels in audio event classification. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9414579. URL https://arxiv.org/abs/2105.07031.

[102] Meghana Holla and Ismini Lourentzou. Commonsense for zero-shot natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2166–2174, 2024. URL https://ojs.aaai.org/index.php/AAAI/article/view/27989.

[103] Coleman Hooper, Minwoo Kang, Suhong Moon, Nicholas Lee, et al. Speculative interaction agents: Building real-time agents with asynchronous I/O and speculative tool calling. *arXiv preprint arXiv:2605.13360*, 2026. URL https://arxiv.org/abs/2605.13360.

[104] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 159–166. Association for Computing Machinery, 1999. doi: 10.1145/302979.303030. URL https://doi.org/10.1145/302979.303030.

[105] Chi-Yuan Hsiao, Ke-Han Lu, Yu-Kuan Fu, Guan-Ting Lin, Hsiao-Tsung Hung, and Hung yi Lee. Aspirin: Action space projection for interactivity-optimized reinforcement learning in full-duplex speech language models, 2026. URL https://arxiv.org/abs/2604.10065.

[106] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291. URL https://doi.org/10.1109/TASLP.2021.3122291.

[107] Huazhang Hu, Sixun Dong, Yiqun Zhao, Dongze Lian, Zhengxin Li, and Shenghua Gao. Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18991–19000, 2022. doi: 10.1109/CVPR52688.2022.01843. URL https://openaccess.thecvf.com/content/CVPR2022/html/Hu_TransRAC_Encoding_Multi-Scale_Temporal_Correlation_With_Transformers_for_Repetitive_Action_CVPR_2022_paper.html.

[108] Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. HiAgent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32779–32798, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1575. URL https://aclanthology.org/2025.acl-long.1575/.

[109] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14271–14280, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VTimeLLM_Empower_LLM_to_Grasp_Video_Moments_CVPR_2024_paper.html.

[110] Chien-Yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung-Yi Lee. Dynamic-SUPERB: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 12136–12140, 2024. doi: 10.1109/ICASSP48485.2024.10448257. URL https://doi.org/10.1109/ICASSP48485.2024.10448257.

[111] Chien-yu Huang, Wei-Chih Chen, Shu-Wen Yang, Andy T. Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, William Chen, Chih-Kai Yang, Xuanjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih-Heng Wang, Chun-Yi Kuan, Ke-Han Lu, Kai-Wei Chang, Fabian Alejandro Ritter Gutierrez, and et al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=s7lzZpAW7T.

[112] Muye Huang, Lingling Zhang, Xingyu Yu, Lei Shi, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He, and Jun Liu. Duplexomni: Real-time listening, seeing, thinking, and speaking for full-duplex interaction, 2026. URL https://arxiv.org/abs/2606.09186.

[113] Weidong Huang, Jiaming Ji, Chunhe Xia, Borong Zhang, and Yaodong Yang. SafeDreamer: Safe reinforcement learning with world models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ece182f93af26c64187ba3f7dfd4309a-Abstract-Conference.html.

[114] Ranggi Hwang, Jianyu Wei, Shijie Cao, Changho Hwang, Xiaohu Tang, Ting Cao, and Mao Yang. Pre-Gated MoE: An algorithm-system co-design for fast and scalable mixture-of-expert inference. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture*, pages 1018–1031, 2024. doi: 10.1109/ISCA59077.2024.00078. URL https://doi.org/10.1109/ISCA59077.2024.00078.

[115] Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. Multilingual turn-taking prediction using voice activity projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11873–11883, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1036/.

[116] Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. Real-time and continuous turn-taking prediction using voice activity projection. *arXiv preprint arXiv:2401.04868*, 2024. doi: 10.48550/arXiv.2401.04868. URL https://arxiv.org/abs/2401.04868.

[117] Koji Inoue, Divesh Lala, Mikey Elmers, Keiko Ochi, and Tatsuya Kawahara. An LLM benchmark for addressee recognition in multi-modal multi-party dialogue. In Maria Ines Torres, Yuki Matsuda, Zoraida Callejas, Arantza del Pozo, and Luis Fernando D'Haro, editors, *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 330–334, Bilbao, Spain, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-248-0. URL https://aclanthology.org/2025.iwsds-1.36/.

[118] Koji Inoue, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. Yeah, un, oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7171–7181, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.367. URL https://aclanthology.org/2025.naacl-long.367/.

[119] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International*

Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

[120] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs and outputs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=fILj7WpI-g.

[121] Dhruv Jain, Harshit Shukla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. VoiceAgentBench: Are voice assistants ready for agentic tasks? *arXiv preprint arXiv:2510.07978*, 2025. doi: 10.48550/arXiv.2510.07978. URL https://arxiv.org/abs/2510.07978.

[122] Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. WavChat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*, 2024. doi: 10.48550/arXiv.2411.13577. URL https://arxiv.org/abs/2411.13577.

[123] Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. WavTokenizer: An efficient acoustic discrete codec tokenizer for audio language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ea1f5f0878d43ff4fb8bf64ef4a2326c-Abstract-Conference.html.

[124] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. doi: 10.48550/arXiv.2401.04088. URL https://arxiv.org/abs/2401.04088.

[125] Bing'er Jiang, Erik Ekstedt, and Gabriel Skantze. What makes a good pause? investigating the turn-holding effects of fillers. In Radek Skarnitzl and Jan Volín, editors, *Proceedings of the 20th International Congress of Phonetic Sciences – ICPhS 2023*, pages 3512–3516, Prague, Czech Republic, 2023. International Phonetic Association. URL https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/828.pdf.

[126] Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H Abdi, Dongsheng Li, Chin-Yew Lin, et al. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *Advances in Neural Information Processing Systems*, 37:52481–52515, 2024. doi: 10.52202/079017-1663. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5dfbe6f5671e82c76841ba687a8a9ecb-Abstract-Conference.html.

[127] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

[128] H. Jin and Y. Wu. Ce-collm: Efficient and adaptive large language models through cloud-edge collaboration. In *Proceedings of the IEEE International Conference on Web Services*, 2025. doi: 10.1109/ICWS67624.2025.00046. URL https://ieeexplore.ieee.org/document/11169709/.

[129] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22605–22623. PMLR, 2024. URL https://proceedings.mlr.press/v235/ju24b.html.

[130] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 7669–7673. IEEE, May 2020. doi: 10.1109/icassp40776.2020.9052942. URL http://dx.doi.org/10.1109/ICASSP40776.2020.9052942.

[131] Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. LibriHeavy: A 50,000 hours ASR corpus with punctuation casing and context. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10991–10995. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447120. URL https://arxiv.org/abs/2309.08105.

[132] Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.593. URL https://aclanthology.org/2022.acl-long.593/.

[133] Heeseung Kim, Soonshin Seo, Kyeongseok Jeong, Ohsung Kwon, Soyoon Kim, Jungwhan Kim, Jaehong Lee, Eunwoo Song, Myungwoo Oh, Jung-Woo Ha, Sungroh Yoon, and Kang Min Yoo. Paralinguistics-aware speech-empowered large language models for natural conversation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 131072–131103. Curran Associates, Inc., 2024. doi: 10.52202/079017-4166. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ecfa5340fd896e5314bc5e132b5dd5ca-Paper-Conference.pdf.

[134] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR, 2021. URL https://proceedings.mlr.press/v139/kim21f.html.

[135] Sehoon Kim, Karttikeya Mangalam, Suhong Moon, Jitendra Malik, Michael W. Mahoney, Amir Gholami, and Kurt Keutzer. Speculative decoding with big little decoder. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-1705. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7b97adeafa1c51cf65263459ca9d0d7c-Abstract-Conference.html.

[136] Takyoung Kim, Kang wook Kim, Sang Hoon Woo, Julia Hirschberg, Gunhee Kim, and Dilek Hakkani-Tür. Duplexgen: Adaptive synthesis of human-ai turn-taking dialogues, 2026. URL https://arxiv.org/abs/2607.26178.

[137] Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus. In *Interspeech 2023*, pages 5496–5500, 2023. doi: 10.21437/Interspeech.2023-1584. URL https://www.isca-archive.org/interspeech_2023/koizumi23_interspeech.html.

[138] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2023. doi: 10.52202/075280-1214. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/58d0e78cf042af5876e12661087bea12-Abstract-Conference.html.

[139] Nikita Kuzmin, Tao Zhong, Jiajun Deng, Yingke Zhu, Tristan Tsoi, Tianxiang Cao, Simon Lui, Kong Aik Lee, and Eng Siong Chng. Privacy-preserving end-to-end full-duplex speech dialogue models, 2026. URL https://arxiv.org/abs/2603.08179.

[140] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*. Association for Computing Machinery, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

[141] Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. Parler-tts. https://github.com/huggingface/parler-tts, 2024.

[142] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021. doi: 10.1162/tacl_a_00430. URL https://aclanthology.org/2021.tacl-1.79/.

[143] Md Tahmid Rahman Laskar, Xue-Yong Fu, Seyyed Saeed Sarfjoo, Quinten McNamara, Jonas Robertson, and Shashi Bhushan TN. From text to voice: A reproducible and verifiable framework for evaluating tool calling LLM agents. *arXiv preprint arXiv:2605.15104*, 2026. doi: 10.48550/arXiv.2605.15104. URL https://arxiv.org/abs/2605.15104.

[144] Allison Lau, Younwoo Choi, Vahid Balazadeh Meresht, Keertana Chidambaram, Vasilis Syrgkanis, and Rahul G. Krishnan. Personalized adaptation via in-context preference learning, 2024. URL https://arxiv.org/abs/2410.14001.

[145] Gael Le Lan, Varun Nagaraja, Ernie Chang, David Kant, Zhaoheng Ni, Yangyang Shi, Forrest Iandola, and Vikas Chandra. Stack-and-delay: A new codebook pattern for music generation. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 796–800. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447392. URL https://arxiv.org/abs/2309.08804.

[146] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202

of *Proceedings of Machine Learning Research*, pages 19274–19286. PMLR, 2023. URL https://proceedings.mlr.press/v202/leviathan23a.html.

[147] Stephen C. Levinson. Turn-taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1):6–14, 2016. doi: 10.1016/j.tics.2015.10.010. URL https://doi.org/10.1016/j.tics.2015.10.010.

[148] Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6:731, 2015. doi: 10.3389/fpsyg.2015.00731. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.00731/full.

[149] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[150] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026. doi: 10.48550/arXiv.2602.01541. URL https://arxiv.org/abs/2602.01541.

[151] Jisen Li, Bingxuan Li, Nanyi Jiang, Xuying Ning, Xiyao Wang, Yifan Shen, Heng Wang, Yuqing Jian, Xiaoxia Wu, Ben Athiwaratkun, et al. Multi-turn agentic scientific literature search via workflow induction. *arXiv preprint arXiv:2607.00597*, 2026. doi: 10.48550/arXiv.2607.00597. URL https://arxiv.org/abs/2607.00597.

[152] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

[153] Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. YODAS: YouTube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023. doi: 10.1109/ASRU57964.2023.10389689. URL https://arxiv.org/abs/2406.00899.

[154] Yadong Li, Guoxin Wu, Haiping Hou, and Biye Li. Uaf: A unified audio front-end llm for full-duplex speech interaction, 2026. URL https://arxiv.org/abs/2604.19221.

[155] Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *Computer Vision – ECCV 2024*, volume 15104 of *Lecture Notes in Computer Science*, pages 323–340. Springer, 2024. doi: 10.1007/978-3-031-72952-2_19. URL https://link.springer.com/chapter/10.1007/978-3-031-72952-2_19.

[156] Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan Wen, Yanghai Wang, Shihao Li, Zhaoxiang Zhang, Ruibo Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. OmniBench: Towards the future of universal omni-language models. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, volume 38,

2025. doi: 10.52202/085713-1036. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2c7c4a12a9dcf7dace9896d08154c705-Abstract-Datasets_and_Benchmarks_Track.html.

[157] Yuxin Li, Donghang Wu, Guan-Ting Lin, Hung-yi Lee, Chengwei Qin, Zhehuai Chen, and Chen Chen. Decoupling conversational dynamics in full-duplex spoken models through reinforcement learning. *arXiv preprint arXiv:2607.07148*, 2026. doi: 10.48550/arXiv.2607.07148. URL https://arxiv.org/abs/2607.07148.

[158] Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis, 2024. URL https://arxiv.org/abs/2411.01156.

[159] Tatiana Likhomanenko, Luke Carlson, Richard He Bai, Zijin Gu, Han Tran, Zakaria Aldeneh, Yizhe Zhang, Ruixiang Zhang, Huangjie Zheng, and Navdeep Jaitly. ChipChat: Low-latency cascaded conversational agent in MLX. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–4. IEEE, 2025. doi: 10.1109/ASRU65441.2025.11434783. URL https://machinelearning.apple.com/research/chipchat.

[160] Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung-Yi Lee. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop*, pages 1–8. IEEE, 2025. doi: 10.1109/ASRU65441.2025.11433838. URL https://doi.org/10.1109/ASRU65441.2025.11433838.

[161] Guan-Ting Lin, Prashanth Gurunath Shivakumar, Aditya Gourav, Yile Gu, Ankur Gandhe, Hung-yi Lee, and Ivan Bulyko. Align-SLM: Textless spoken language models with reinforcement learning from AI feedback. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20395–20411, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.997. URL https://aclanthology.org/2025.acl-long.997/.

[162] Guan-Ting Lin, Chen Chen, Zhehuai Chen, and Hung-yi Lee. Full-duplex-bench-v3: Benchmarking tool use for full-duplex voice agents under real-world disfluency. *CoRR*, abs/2604.04847, 2026. doi: 10.48550/ARXIV.2604.04847. URL https://doi.org/10.48550/arXiv.2604.04847.

[163] Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe, and Hung-yi Lee. Full-duplex-bench-v2: A multi-turn evaluation framework for duplex dialogue systems with an automated examiner. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 27–36, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-short.4. URL https://aclanthology.org/2026.acl-short.4/.

[164] Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, Shinji Watanabe, and Hung-Yi Lee. Full-duplex-bench v1.5: Evaluating overlap handling for full-duplex speech models. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 19447–19451. IEEE, 2026. doi: 10.1109/ICASSP55912.2026.11463576. URL https://doi.org/10.1109/ICASSP55912.2026.11463576.

[165] Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the gap for MLLMs to achieve streaming video understanding. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 12147–12151. IEEE, 2026. doi: 10.1109/ICASSP55912.2026.11463959. URL https://doi.org/10.1109/ICASSP55912.2026.11463959.

[166] Ting-En Lin, Yuchuan Wu, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. Duplex conversation: Towards human-like interaction in spoken dialogue systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pages 3299–3308. Association for Computing Machinery, 2022. doi: 10.1145/3534678.3539209. URL https://doi.org/10.1145/3534678.3539209.

[167] Yueqian Lin, Zhengmian Hu, Jayakumar Subramanian, Qinsi Wang, Nikos Vlassis, Hai Helen Li, and Yiran Chen. AsyncVoice agent: Real-time explanation for LLM planning and reasoning. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–4. IEEE, 2025. doi: 10.1109/ASRU65441.2025.11434747. URL https://arxiv.org/abs/2510.16156.

[168] Linguistic Data Consortium. CALLHOME Collection in Six Languages. LINGUIST List 8.1209, 1997. URL https://linguistlist.org/issues/8/1209/. Language-specific telephone speech, transcript, and lexicon releases for American English, Egyptian Arabic, German, Japanese, Mandarin Chinese, and Spanish.

[169] Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D. Plumbley. SemantiCodec: An ultra low bitrate semantic audio codec for general sound. *IEEE Journal of Selected Topics in Signal Processing*, 18(8):1448–1461, December 2024. doi: 10.1109/JSTSP.2024.3506286. URL https://arxiv.org/abs/2405.00233.

[170] Jinming Liu, Jianguo Huang, Zhaoyang Jia, Jiahao Li, Xiaoyi Zhang, Zongyu Guo, Bin Li, Wenjun Zeng, Yan Lu, and Xin Jin. An efficient streaming video understanding framework with agentic control, 2026. URL https://arxiv.org/abs/2605.17921.

[171] Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. *arXiv preprint arXiv:2602.19163*, 2026. URL https://arxiv.org/abs/2602.19163.

[172] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

[173] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pages 38–55. Springer, 2024. doi: 10.1007/978-3-031-72970-6_3. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6319_ECCV_2024_paper.php.

[174] Z. Liu, C. Zhao, F. Iandola, C. Lai, Y. Tian, I. Fedorov, Y. Xiong, et al. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. In *Proceedings of the International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/liu24ce.html.

[175] Yuxuan Lou, Kai Yang, and Yang You. MoST: Mixing speech and text with modality-aware mixture of experts. *arXiv preprint arXiv:2601.10272*, 2026. URL https://arxiv.org/abs/2601.10272.

[176] Xudong Lu, Xueying Li, Annan Wang, Yang Bo, Jinpeng Chen, Zengliang Li, Nianzu Yang, Rui Liu, Xue Yang, Jingwen Hou, et al. Omniinteract: Benchmarking real-world streaming interaction for real-time omnimodal assistants. *arXiv preprint arXiv:2605.26485*, 2026. doi: 10.48550/arXiv.2605.26485. URL https://arxiv.org/abs/2605.26485.

[177] Linhan Ma, Dake Guo, Kun Song, Yuepeng Jiang, Shuai Wang, Liumeng Xue, Weiming Xu, Huan Zhao, Binbin Zhang, and Lei Xie. WenetSpeech4TTS: A 12,800-hour mandarin TTS corpus for large speech generation model benchmark. In *Interspeech 2024*, pages 1840–1844, 2024. doi: 10.21437/Interspeech.2024-2343. URL https://www.isca-archive.org/interspeech_2024/ma24d_interspeech.html.

[178] Mingbo Ma, Baigong Zheng, Kaibo Liu, Renjie Zheng, Hairong Liu, Kainan Peng, Kenneth Church, and Liang Huang. Incremental text-to-speech synthesis with prefix-to-prefix framework. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3886–3896, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.346. URL https://aclanthology.org/2020.findings-emnlp.346/.

[179] Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. SIMULEVAL: An evaluation toolkit for simultaneous translation. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.19. URL https://aclanthology.org/2020.emnlp-demos.19/.

[180] Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. Language model can listen while speaking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24831–24839, 2025. doi: 10.1609/aaai.v39i23.34665. URL https://ojs.aaai.org/index.php/AAAI/article/view/34665.

[181] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

[182] Gallil Maimon, Michael Hassid, Amit Roth, and Yossi Adi. Scaling analysis of interleaved speech-text language models. In *The 2nd Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=IXwgE8hyJs.

[183] Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. VoxtLM: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 13326–13330. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447112. URL https://doi.org/10.1109/ICASSP48485.2024.10447112.

[184] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and*

Language Processing, 32:3339–3354, 2024. ISSN 2329-9304. doi: 10.1109/taslp.2024.3419446. URL http://dx.doi.org/10.1109/TASLP.2024.3419446.

[185] Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, Helen M. Meng, and Furu Wei. Autoregressive speech synthesis without vector quantization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1287–1300, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.65. URL https://aclanthology.org/2025.acl-long.65/.

[186] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. Augmented language models: A survey. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=jh7wH2AzKK.

[187] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. doi: 10.1109/ICCV.2019.00272. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Miech_HowTo100M_Learning_a_Text-Video_Embedding_by_Watching_Hundred_Million_Narrated_ICCV_2019_paper.html.

[188] Thao Le Minh, Nobuyuki Shimizu, Takashi Miyazaki, and Koichi Shinoda. Deep learning based multimodal addressee recognition in visual scenes with utterances. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1546–1553. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/214. URL https://doi.org/10.24963/ijcai.2018/214.

[189] Kentaro Mitsui, Koh Mitsuda, Toshiaki Wakatsuki, Yukiya Hono, and Kei Sawada. PSLM: Parallel generation of text and speech with LLMs for low-latency spoken dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2692–2700, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.151. URL https://aclanthology.org/2024.findings-emnlp.151/.

[190] Abdelrahman Mohamed, Hung yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210, 2022. doi: 10.1109/JSTSP.2022.3207050. URL https://doi.org/10.1109/JSTSP.2022.3207050.

[191] Pooneh Mousavi, Gallil Maimon, Adel Moumen, Darius Petermann, Jiatong Shi, et al. Discrete audio tokens: More than a survey! *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=eqNchtvc6v.

[192] Pooneh Mousavi, Jarod Duret, Darius Petermann, Artem Ploujnikov, Luca Della Libera, Anastasia Kuznetsova, Cem Subakan, and Mirco Ravanelli. DASB: Discrete audio and speech benchmark. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=vGWrp0NjaE.

[193] Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 2024. doi: 10.48550/arXiv.2404.07143. URL https://arxiv.org/abs/2404.07143.

[194] Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, R. J. Skerry-Ryan, and Michele Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered LLM. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e393677793767624f2821cec8bdd02f1-Abstract-Conference.html.

[195] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620, 2017. doi: 10.21437/Interspeech.2017-950. URL https://www.isca-archive.org/interspeech_2017/nagrani17_interspeech.html.

[196] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. doi: 10.48550/arXiv.2112.09332. URL https://arxiv.org/abs/2112.09332.

[197] Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. DuplexChat: Constructing speaker-separated full-duplex dialogue speech at scale for spoken dialogue language modeling. *arXiv preprint arXiv:2607.04941*, 2026. doi: 10.48550/arXiv.2607.04941. URL https://arxiv.org/abs/2607.04941.

[198] Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266, 2023. doi: 10.1162/tacl_a_00545. URL https://aclanthology.org/2023.tacl-1.15/.

[199] Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. SpiRit-LM: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52, 2025. doi: 10.1162/tacl_a_00728. URL https://aclanthology.org/2025.tacl-1.2/.

[200] Xuying Ning, Katherine Tieu, Dongqi Fu, Tianxin Wei, Zihao Li, Yuanchen Bei, Jiaru Zou, Mengting Ai, Zhining Liu, Ting-Wei Li, et al. Code as agent harness. *arXiv preprint arXiv:2605.18747*, 2026. doi: 10.48550/arXiv.2605.18747. URL https://arxiv.org/abs/2605.18747.

[201] Atsumoto Ohashi, Neil Zeghidour, Alexandre Défossez, and Eugene Kharitonov. Multi-faceted interactivity alignment in full-duplex speech models, 2026. URL https://arxiv.org/abs/2606.11167.

[202] Yuki Okafuji, Koji Inoue, and Yoshiki Ohira. Low-latency turn-taking via context-aware preface generation in a real-world dialogue robot, 2026. URL https://arxiv.org/abs/2607.23204.

[203] OmniJev. PlayJev: A multimodal JEV-like model for small games. GitHub repository, 2026. URL https://github.com/OmniJev/PlayJev. Accessed September 23, 2026.

[204] OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. doi: 10.48550/arXiv.2410.21276. URL https://arxiv.org/abs/2410.21276.

[205] OpenAI. GPT-Realtime-2 model. https://developers.openai.com/api/docs/models/gpt-realtime-2, 2026. OpenAI API Documentation, accessed September 7, 2026.

[206] Leyi Pan, Zheyu Fu, Yunpeng Zhai, Shuchang Tao, Sheng Guan, Shiyu Huang, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Felix Henry, Lijie Wen, and Aiwei Liu. Omni-SafetyBench: A benchmark for safety evaluation of audio-visual large language models. *arXiv preprint arXiv:2508.07173*, 2025. doi: 10.48550/arXiv.2508.07173. URL https://arxiv.org/abs/2508.07173.

[207] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://doi.org/10.1109/ICASSP.2015.7178964.

[208] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. Splitwise: Efficient generative LLM inference using phase splitting. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture*, pages 118–132. IEEE, 2024. doi: 10.1109/ISCA59077.2024.00019. URL https://doi.org/10.1109/ISCA59077.2024.00019.

[209] Zifan Peng, Yule Liu, Zhen Sun, Mingchen Li, Zeren Luo, Jingyi Zheng, Wenhan Dong, Xinlei He, Xuechao Wang, Yingjie Xue, Shengmin Xu, and Xinyi Huang. JALMBench: Benchmarking jailbreak vulnerabilities in audio language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DJkQ236C8B.

[210] Mathis Pink, Qinyuan Wu, Vy Ai Vo, Javier Turek, Jianing Mu, Alexander Huth, and Mariya Toneva. Position: Episodic memory is the missing piece for long-term LLM agents, 2025. URL https://arxiv.org/abs/2502.06975.

[211] Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 325–351, 2024. doi: 10.1109/TASLP.2023.3328283. URL https://doi.org/10.1109/TASLP.2023.3328283.

[212] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, page 2757–2761. ISCA, October 2020. doi: 10.21437/interspeech.2020-2826. URL http://dx.doi.org/10.21437/Interspeech.2020-2826.

[213] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, volume 36, pages 42748–42761, 2023. doi: 10.52202/075280-1852. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8540fba4abdc7f9f7a7b1cc6cd60e409-Abstract.html.

[214] Haotian Qi and Gabriel Skantze. Muvap: Multimodal multiparty voice activity projection for turn-taking prediction in the wild. *arXiv preprint arXiv:2606.16731*, 2026. doi: 10.48550/arXiv.2606.16731. URL https://arxiv.org/abs/2606.16731.

[215] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video LLMs with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR), pages 24045–24055, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Qian_Dispider_Enabling_Video_LLMs_with_Active_Real-Time_Interaction_via_Disentangled_CVPR_2025_paper.html.

[216] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/28e50ee5b72e90b50e7196fde8ea260e-Abstract-Conference.html.

[217] Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343, 2025. doi: 10.1007/s11704-024-40678-2. URL https://doi.org/10.1007/s11704-024-40678-2.

[218] Guanqiao Qu, Qiyuan Chen, Wei Wei, Zheng Lin, Xianhao Chen, and Kaibin Huang. Mobile edge intelligence for large language models: A contemporary survey. *IEEE Communications Surveys & Tutorials*, 27(6):3820–3860, 2025. doi: 10.1109/COMST.2025.3527641. URL https://arxiv.org/abs/2407.18921.

[219] Hongyu Qu, Guangming Yao, Ling Xing, Xiaobin Hu, Rongxing Ding, Guibin Zhang, Fan Zhang, Yi Yuan, Xiangbo Shu, and Shuicheng Yan. Beyond retrieval: Progressive latent memory evolution for streaming video understanding, 2026. URL https://arxiv.org/abs/2609.04131.

[220] Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026. doi: 10.48550/arXiv.2604.15804. URL https://arxiv.org/abs/2604.15804.

[221] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[222] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

[223] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

[224] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ren_TimeChat_A_Time-sensitive_Multimodal_Large_Language_Model_for_Long_Video_CVPR_2024_paper.html.

[225] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=piLPYqxtWuA.

[226] Pablo Riera, Pablo Brusco, Cristina Kuo, Marcelo Sancinetti, and S. R. K. Branavan. Synchronization and turn-taking in full-duplex speech dialogue models, 2026. URL https://arxiv.org/abs/2605.20356.

[227] Matthew Roddy and Naomi Harte. Neural generation of dialogue response timings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2442–2452, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.221. URL https://aclanthology.org/2020.acl-main.221/.

[228] Matthew Roddy, Gabriel Skantze, and Naomi Harte. Multimodal continuous turn-taking prediction using multiscale RNNs. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 186–190. Association for Computing Machinery, 2018. doi: 10.1145/3242969.3242997. URL https://doi.org/10.1145/3242969.3242997.

[229] Nathan Roll, Calbert Graham, and Simon Todd. PSST! prosodic speech segmentation with transformers. In *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 476–487, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.31. URL https://aclanthology.org/2023.conll-1.31/.

[230] Rajarshi Roy, Jonathan Raiman, Sang-Gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. PersonaPlex: Voice and role control for full duplex conversational speech models. In *ICASSP 2026 – 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16137–16141. IEEE, 2026. doi: 10.1109/ICASSP55912.2026.11463413. URL https://doi.org/10.1109/ICASSP55912.2026.11463413.

[231] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974. doi: 10.2307/412243. URL https://doi.org/10.2307/412243.

[232] Takaaki Saeki, Shinnosuke Takamichi, and Hiroshi Saruwatari. Low-latency incremental text-to-speech synthesis with distilled context prediction network. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 749–756. IEEE, 2021. doi: 10.1109/ASRU51503.2021.9687904. URL https://doi.org/10.1109/ASRU51503.2021.9687904.

[233] Tara N. Sainath and Carolina Parada. Convolutional neural networks for small-footprint keyword spotting. In *Proceedings of Interspeech 2015*, pages 1478–1482, 2015. doi: 10.21437/Interspeech.2015-352. URL https://doi.org/10.21437/Interspeech.2015-352.

[234] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, et al. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-2997. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

[235] Frank Seide, Yangyang Shi, Morrie Doulaty, Yashesh Gaur, Junteng Jia, and Chunyang Wu. Speech ReaLLM – real-time speech recognition with multimodal language models by teaching the flow of time. In *Interspeech 2024*, pages 1900–1904, 2024. doi: 10.21437/Interspeech.2024-571. URL https://www.isca-archive.org/interspeech_2024/seide24_interspeech.html.

[236] Yuan Shangguan, Kate Knister, Yanzhang He, Ian McGraw, and Françoise Beaufays. Analyzing the quality and stability of a streaming end-to-end on-device speech recognizer. In *Proceedings of Interspeech 2020*, pages 591–595, 2020. doi: 10.21437/Interspeech.2020-1194. URL https://www.isca-archive.org/interspeech_2020/shangguan20_interspeech.html.

[237] Hang Shao, Heting Gao, Yunhang Shen, Jiawei Chen, Zuwei Long, Dong Yang, Ke Li, and Xing Sun. DeepOmni: Towards seamless and smart speech interaction with adaptive modality-specific MoE. *arXiv preprint arXiv:2506.21864*, 2025. URL https://arxiv.org/abs/2506.21864.

[238] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in VLMs. In *Advances in Neural Information Processing Systems*, volume 38, pages 20406–20437, 2025. doi: 10.52202/085713-0606. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1a17a06de88cf77f25cda0da91615a54-Abstract-Conference.html.

[239] Yifan Shen, Pei Tian, Xinzhuo Li, Bowen Fang, Shujun Xia, Bingxuan Li, Ana Jojic, Wenming Ye, Xu Cao, James Matthew Rehg, and Ismini Lourentzou. CogniRoute: Learning to route social evidence in omni-modal models. *arXiv preprint arXiv:2606.20970*, 2026. doi: 10.48550/arXiv.2606.20970. URL https://arxiv.org/abs/2606.20970.

[240] Yifan Shen, Jian Xu, Boyi Li, Yuner Zhang, Tianjiao Yu, Bingxuan Li, Houze Yang, Rushi Wang, and Xu Cao. ChronoVision: Temporal reasoning via latent state reconstruction. *arXiv preprint arXiv:2608.05631*, 2026. doi: 10.48550/arXiv.2608.05631. URL https://arxiv.org/abs/2608.05631.

[241] Yifan Shen, Jiawen Zhang, Jian Xu, Junho Kim, Ismini Lourentzou, Xu Cao, and Meihuan Huang. Evaluating cognitive age alignment in interactive AI agents. *arXiv preprint arXiv:2605.17894*, 2026. doi: 10.48550/arXiv.2605.17894. URL https://arxiv.org/abs/2605.17894.

[242] Ying Shen, Daniel Bis, Cynthia Lu, and Ismini Lourentzou. Elba: Learning by asking for embodied visual navigation and task completion. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 5177–5186, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Shen_ELBA_Learning_by_Asking_for_Embodied_Visual_Navigation_and_Task_WACV_2025_paper.html.

[243] Yi-Jen Shih, Desh Raj, Chunyang Wu, Wei Zhou, SK Bong, Yashesh Gaur, Jay Mahadeokar, Ozlem Kalinli, and Mike Seltzer. Can speech llms think while listening?, 2025. URL https://arxiv.org/abs/2510.07497.

[244] Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. *arXiv preprint arXiv:2504.07951*, 2025. doi: 10.1109/ICCV51701.2025.00009. URL https://openaccess.thecvf.com/content/ICCV2025/html/Shukor_Scaling_Laws_for_Native_Multimodal_Models_ICCV_2025_paper.html.

[245] Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 39088–39118. Curran

Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7b16688a2b053a1b01474ab5c78ce662-Paper-Datasets_and_Benchmarks.pdf.

[246] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision – ECCV 2016*, volume 9905 of *Lecture Notes in Computer Science*, pages 510–526. Springer, 2016. doi: 10.1007/978-3-319-46448-0_31. URL https://doi.org/10.1007/978-3-319-46448-0_31.

[247] Yash Raj Singh. CrossOracle: An agentic framework for real-time expert personas with parallelized acoustic synthesis. In *Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 254–257, Atlanta, Georgia, USA, August 2026. Association for Computational Linguistics. URL https://aclanthology.org/2026.sigdial-1.18/.

[248] Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=vY9nzQmQBw.

[249] Hubert Siuzdak, Florian Grötschla, and Luca A. Lanzendörfer. SNAC: Multi-scale neural audio codec. In *Audio Imagination: NeurIPS 2024 Workshop on AI-Driven Speech, Music, and Sound Generation*, 2024. URL https://openreview.net/forum?id=PFBF5ctj4X.

[250] Gabriel Skantze. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230, Saarbrücken, Germany, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5527. URL https://aclanthology.org/W17-5527/.

[251] Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178, 2021. ISSN 0885-2308. doi: 10.1016/j.csl.2020.101178. URL https://www.sciencedirect.com/science/article/pii/S088523082030111X.

[252] Gabriel Skantze and Bahar Irfan. Applying general turn-taking models to conversational human-robot interaction. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 859–868. IEEE, 2025. doi: 10.1109/HRI61500.2025.10973958. URL https://arxiv.org/abs/2501.08946.

[253] Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Guangxian Ouyang, Zhenhao Chen, and Xiuying Chen. Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27294–27308, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1259. URL https://aclanthology.org/2026.acl-long.1259/.

[254] Kusha Sridhar, Ross Cutler, Ando Saabas, Tanel Pärnamaa, Markus Loide, Hannes Gamper, Sebastian Braun, Robert Aichner, and Sriram Srinivasan. ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9413457. URL https://doi.org/10.1109/ICASSP39728.2021.9413457.

[255] Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, et al. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106

(26):10587–10592, 2009. doi: 10.1073/pnas.0903616106. URL https://www.pnas.org/doi/10.1073/pnas.0903616106.

[256] Jiapeng Tang, Wei Cao, Biao Zhang, Chang Luo, Yaoyao Liu, and Matthias Nießner. Motion2VecSets: Non-rigid shape reconstruction and tracking with 4D latent set diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):9519–9536, August 2026. doi: 10.1109/TPAMI.2026.3680779. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2026.3680779.

[257] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29118–29128, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tang_Adaptive_Keyframe_Sampling_for_Long_Video_Understanding_CVPR_2025_paper.html.

[258] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2): 1355–1376, 2026. doi: 10.1109/TCSVT.2025.3566695. URL https://doi.org/10.1109/TCSVT.2025.3566695.

[259] Yuzhi Tang, Wentao Ma, Xiling Zhao, Ahmad Salimi, Sepehr Harfi Moridani, Dongming Shen, Jixuan Wang, Abdulrahman Abdulrazzag, Murdock Aubry, Yu-Hua Chen, Daniel Lee, Jaewon Lee, Jonah Mackey, Silin Meng, Nicholas Stranges, Chenxu Xiong, Hao Yu, Yi Zhu, Mu Li, and Alex Smola. Instruct-fd: Can your full-duplex speech system follow turn-taking instructions?, 2026. URL https://arxiv.org/abs/2607.20460.

[260] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. GitHub repository, 2023. URL https://github.com/tatsu-lab/stanford_alpaca.

[261] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5.

[262] Thinking Machines Lab. Interaction models: A scalable approach to human–ai collaboration. Thinking Machines Lab, Connectionism, May 2026. URL https://thinkingmachines.ai/blog/interaction-models/. Accessed September 20, 2026.

[263] Xueyun Tian, Wei Li, Bingbing Xu, Heng Dong, Yuanzhuo Wang, and Huawei Shen. ROMA: Real-time omni-multimodal assistant with interactive streaming understanding. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 23018–23039, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1153. URL https://aclanthology.org/2026.findings-acl.1153/.

[264] Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, Xueheng Li, Lumin Li, Chenxu Guo, Jiasheng Zhou, Jiandong Chen, Xianye Wu, Jiahao Wang, Silei Wu, Lei Chen, Hanming Deng, Yuxuan Song, Dinghao Zhou, Guiping Zhong, Ken Zheng, Shiyin Kang, and Lewei Lu. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue, 2025. URL https://arxiv.org/abs/2510.13747.

[265] Hung Pham Van, Nguyen Manh Hieu, Khang Pham Tran Tuan, Nam Le Hai, Linh Ngo Van, Nguyen Thi Ngoc Diep, and Trung Le. MemORAI: Memory organization and retrieval via adaptive graph intelligence for LLM conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 28235–28253, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1408. URL https://aclanthology.org/2026.findings-acl.1408/.

[266] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. doi: 10.48550/arXiv.1807.03748. URL https://arxiv.org/abs/1807.03748.

[267] Shikhar Vashishth, Harman Singh, Shikhar Bharadwaj, Sriram Ganapathy, Chulayuth Asawaroengchai, Kartik Audhkhasi, Andrew Rosenberg, Ankur Bapna, and Bhuvana Ramabhadran. STAB: Speech tokenizer assessment benchmark. *arXiv preprint arXiv:2409.02384*, 2024. doi: 10.48550/arXiv.2409.02384. URL https://arxiv.org/abs/2409.02384.

[268] Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21390–21402, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1192. URL https://aclanthology.org/2024.emnlp-main.1192/.

[269] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.80. URL https://aclanthology.org/2021.acl-long.80/.

[270] Chao Wang, Rui-Chen Zheng, Yang Ai, and Zhen-Hua Ling. Understanding textual capability degradation in speech LLMs via parameter importance analysis. In *ICASSP 2026 – 2026 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 16282–16286. IEEE, 2026. doi: 10.1109/ICASSP55912.2026.11461684. URL https://doi.org/10.1109/ICASSP55912.2026.11461684.

[271] Chengyou Wang, Hongfei Xue, Guojian Li, Zhixian Zhao, Shuiyuan Wang, Shuai Wang, Xin Xu, Hui Bu, and Lei Xie. Full-duplex interaction in spoken dialogue systems: A comprehensive study from the icassp 2026 humdial challenge, 2026. URL https://arxiv.org/abs/2604.21406.

[272] Dingdong Wang, Mingyu Cui, Dongchao Yang, Xueyuan Chen, and Helen Meng. A comparative study of discrete speech tokens for semantic-related tasks with large language models. *arXiv preprint arXiv:2411.08742*, 2024. doi: 10.48550/arXiv.2411.08742. URL https://arxiv.org/abs/2411.08742.

[273] Dingdong Wang, Junan Li, Mingyu Cui, Dongchao Yang, Xueyuan Chen, and Helen M. Meng. Speech discrete tokens or continuous features? a comparative analysis for spoken language understanding in SpeechLLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24913–24924, Suzhou, China, 2025. Association for Computational Linguistics. doi:

10.18653/v1/2025.emnlp-main.1266. URL https://aclanthology.org/2025.emnlp-main.1266/.

[274] Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. StreamBridge: Turning your offline video large language model into a proactive streaming assistant. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4406. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bf6939f9058a391c47014731b2486e2a-Abstract-Conference.html.

[275] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-VideoLLM: Sharpening fine-grained temporal grounding in video large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 959–975, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.50. URL https://aclanthology.org/2025.findings-emnlp.50/.

[276] Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. Turn-taking and backchannel prediction with acoustic and large language model fusion. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 12121–12125. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447196. URL https://doi.org/10.1109/ICASSP48485.2024.10447196.

[277] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024. doi: 10.1007/s11704-024-40231-1. URL https://doi.org/10.1007/s11704-024-40231-1.

[278] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. doi: 10.48550/arXiv.2409.12191. URL https://arxiv.org/abs/2409.12191.

[279] Qichao Wang, Ziqiao Meng, Wenqian Cui, Yifei Zhang, Pengcheng Wu, Bingzhe Wu, Irwin King, Liang Chen, and Peilin Zhao. NTPP: Generative speech language modeling for dual-channel spoken dialogue via next-token-pair prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 63912–63928. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25by.html.

[280] Tongyan Wang et al. Finemola: Towards fine-grained motion-language alignment from clip-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5510–5519, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/HuMoGen/html/Wang_FineMoLA_Towards_Fine-Grained_Motion-Language_Alignment_from_Clip-Level_Supervision_CVPRW_2026_paper.html.

[281] Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of*

Machine Learning Research, pages 63345–63354. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25aw.html.

[282] Yifei Wang, Yueqi Wang, Zhenrui Yue, Huimin Zeng, Yong Wang, Ismini Lourentzou, Zhengzhong Tu, Xiangxiang Chu, and Julian McAuley. Fasa: Frequency-aware sparse attention. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e839f274b3d3d682a0a82e5129d2cc4a-Abstract-Conference.html.

[283] Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. Proactivevideoqa: A comprehensive benchmark evaluating proactive interactions in video large language models. *CoRR*, abs/2507.09313, 2025. doi: 10.48550/ARXIV.2507.09313. URL https://doi.org/10.48550/arXiv.2507.09313.

[284] Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. VideoLLM knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6338–6359, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.336. URL https://aclanthology.org/2025.findings-emnlp.336/.

[285] Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. OmniMMI: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18925–18935, 2025. doi: 10.1109/CVPR52734.2025.01763. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_OmniMMI_A_Comprehensive_Multi-modal_Interaction_Benchmark_in_Streaming_Video_Contexts_CVPR_2025_paper.html.

[286] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[287] Tianxin Wei, Ting-Wei Li, Zhining Liu, Xuying Ning, Ze Yang, Jiaru Zou, Zhichen Zeng, Ruizhong Qiu, Xiao Lin, Dongqi Fu, Zihao Li, Mengting Ai, Duo Zhou, Wenxuan Bao, Yunzhe Li, Gaotang Li, Cheng Qian, Yu Wang, Xiangru Tang, Yin Xiao, Liri Fang, Hui Liu, Xianfeng Tang, Yuji Zhang, Chi Wang, Jiaxuan You, Heng Ji, Hanghang Tong, and Jingrui He. A survey of agentic reasoning for large language models: Towards recursively self-improving and collective agents. *arXiv preprint arXiv:2601.12538*, 2026. doi: 10.48550/arXiv.2601.12538. URL https://arxiv.org/abs/2601.12538.

[288] Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. VideoRoPE: What makes for good video rotary position embedding? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 66118–66136. PMLR, 2025. URL https://proceedings.mlr.press/v267/wei25h.html.

[289] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, et al. SUPERB: Speech processing universal performance benchmark. In *Interspeech*, 2021. doi: 10.

21437/Interspeech.2021-1775. URL https://www.isca-archive.org/interspeech_2021/yang21c_interspeech.html.

[290] Jason D. Williams, Antoine Raux, and Matthew Henderson. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7:4–33, April 2016. doi: 10.5087/dad.2016.301. URL https://aclanthology.org/2016.dnd-7.5/.

[291] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

[292] Donghang Wu, Tianyu Zhang, Yuxin Li, Hexin Liu, Chen Chen, Eng Siong Chng, and Yoshua Bengio. The silent thought: Modeling internal cognition in full-duplex spoken dialogue models via latent reasoning, 2026. URL https://arxiv.org/abs/2603.17837.

[293] Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kaiwei Chang, Jiawei Du, Ke-Han Lu, Alexander H. Liu, Ho-Lam Chung, Yuan-Kuei Wu, Dongchao Yang, Songxiang Liu, Yi-Chiao Wu, Xu Tan, James Glass, Shinji Watanabe, and Hung-Yi Lee. Codec-SUPERB @ SLT 2024: A lightweight benchmark for neural audio codec models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 570–577. IEEE, 2024. doi: 10.1109/SLT61566.2024.10832364. URL https://arxiv.org/abs/2409.14085.

[294] Haibin Wu, Naoyuki Kanda, Sefik Emre Eskimez, and Jinyu Li. TS3-Codec: Transformer-based simple streaming single codec. In *Interspeech 2025*, pages 604–608, 2025. doi: 10.21437/Interspeech.2025-921. URL https://www.isca-archive.org/interspeech_2025/wu25f_interspeech.html.

[295] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53366–53397. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24e.html.

[296] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pages 21875–21895, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

[297] Hanke Xie, Dake Guo, Chengyou Wang, Yue Li, Wenjie Tian, Xinfa Zhu, Xinsheng Wang, Xiulin Li, Guanqiong Miao, Bo Liu, and Lei Xie. DialoSpeech: Dual-speaker dialogue generation with LLM and flow matching. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–6. IEEE, 2025. doi: 10.1109/APSIPAASC65261.2025.11249327. URL https://doi.org/10.1109/APSIPAASC65261.2025.11249327.

[298] Zhifei Xie and Changqiao Wu. Mini-Omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*, 2024. doi: 10.48550/arXiv.2408.16725. URL https://arxiv.org/abs/2408.16725.

[299] Zhifei Xie and Changqiao Wu. Mini-Omni2: Towards open-source GPT-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*, 2024. doi: 10.48550/arXiv.2410.11190. URL https://arxiv.org/abs/2410.11190.

[300] Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. BigCodec: Pushing the limits of low-bitrate neural speech codec. *arXiv preprint arXiv:2409.05377*, 2024. doi: 10.48550/arXiv.2409.05377. URL https://arxiv.org/abs/2409.05377.

[301] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. doi: 10.48550/arXiv.2503.20215. URL https://arxiv.org/abs/2503.20215.

[302] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. doi: 10.48550/arXiv.2509.17765. URL https://arxiv.org/abs/2509.17765.

[303] Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. StreamingVLM: Real-time understanding for infinite video streams. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gVbPWbA97s.

[304] Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. URO-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17211–17242, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.933. URL https://aclanthology.org/2025.findings-emnlp.933/.

[305] Weicai Yan, Yuhong Dai, Qi Ran, Haodong Li, Wang Lin, Tao Jin, Xing Xie, Hao Liao, and Jianxun Lian. Proact-vl: A proactive videollm for real-time ai companions. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=k9PKgV0L4C.

[306] Bufang Yang, Lilin Xu, Liekang Zeng, Kaiwei Liu, Siyang Jiang, Wenrui Lu, Hongkai Chen, Xiaofan Jiang, Guoliang Xing, and Zhenyu Yan. ContextAgent: Context-aware proactive LLM agents with open-world sensory perceptions. In *Advances in Neural Information Processing Systems*, volume 38, pages 185637–185671, 2025. doi: 10.52202/085713-5582. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4e5cd2079f5a6cf5176818add841531-Abstract-Conference.html.

[307] Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. HiFi-Codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023. doi: 10.48550/arXiv.2305.02765. URL https://arxiv.org/abs/2305.02765.

[308] Yifan Yang, Shujie Liu, Jinyu Li, Hui Wang, Lingwei Meng, et al. Interleaved speech-text language models for simple streaming text-to-speech synthesis. *arXiv preprint arXiv:2412.16102*, 2024. doi: 10.48550/arXiv.2412.16102. URL https://arxiv.org/abs/2412.16102.

[309] Yifan Yang, Zheshu Song, Jianheng Zhuo, Mingyu Cui, Jinpeng Li, Bo Yang, Yexing Du, Ziyang Ma, Xunying Liu, Ziyuan Wang, Ke Li, Shuai Fan, Kai Yu, Wei-Qiang Zhang, Guoguo Chen, and Xie Chen. GigaSpeech 2: An evolving, large-scale and multi-domain ASR corpus for low-resource languages with automated crawling, transcription and refinement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2673–2686. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.135. URL https://aclanthology.org/2025.acl-long.135/.

[310] Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. LiveStar: Live streaming assistant for real-world online video understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.

52202/085713-1050. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2ce4f0b8e24c45318352068603153590-Abstract-Conference.html.

[311] Zhenyu Yang, Kairui Zhang, Qi Liu, Tiancheng Liu, Long Ying, Dizhan Xue, Qibin Hou, Shengsheng Qian, and Changsheng Xu. Towards online interactors: A comprehensive survey on streaming video understanding. *preprints202606.1674.v1*, 2026. doi: 10.20944/preprints202606.1674.v1. URL https://www.preprints.org/manuscript/202606.1674.

[312] Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, Lingpeng Kong, Qi Liu, Yuanxing Zhang, and Xu Sun. TimeChat-Online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10807–10816. ACM, 2025. doi: 10.1145/3746027.3754839. URL https://doi.org/10.1145/3746027.3754839.

[313] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

[314] Yiqun Yao, Naitong Yu, Xiang Li, Xin Jiang, Xuezhi Fang, Wenjia Ma, Xuying Meng, Jing Li, Aixin Sun, and Yequan Wang. EgoMem: Lifelong memory agent for full-duplex omnimodal models. *arXiv preprint arXiv:2509.11914*, 2025. doi: 10.48550/arXiv.2509.11914. URL https://arxiv.org/abs/2509.11914.

[315] Junliang Ye, Zhengyi Wang, Ruowen Zhao, Shenghao Xie, and Jun Zhu. ShapeLLM-Omni: A native multimodal LLM for 3d generation and understanding. In *Advances in Neural Information Processing Systems*, volume 38, pages 99057–99090, 2025. doi: 10.52202/085713-2983. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/80b3224c35ad857b951b8bfccfd32fa5-Abstract-Conference.html.

[316] Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, Yike Guo, and Wei Xue. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25697–25705, 2025. doi: 10.1609/aaai.v39i24.34761. URL https://ojs.aaai.org/index.php/AAAI/article/view/34761.

[317] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 12 2024. ISSN 2095-5138. doi: 10.1093/nsr/nwae403. URL https://doi.org/10.1093/nsr/nwae403.

[318] Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. Orca: A distributed serving system for transformer-based generative models. In *Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation*, pages 521–538, Carlsbad, CA, 2022. USENIX Association. ISBN 978-1-939133-28-1. URL https://www.usenix.org/conference/osdi22/presentation/yu.

[319] Hanfei Yu, Xingqi Cui, Hong Zhang, Hao Wang, and Hao Wang. Taming latency-memory trade-off in MoE-based LLM serving via fine-grained expert offloading. In *Proceedings of the 21st European Conference on Computer Systems*, pages 176–191. Association for Computing Machinery, 2026. doi: 10.1145/3767295.3769319. URL https://doi.org/10.1145/3767295.3769319.

[320] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Yuanzhe Liu, and Ismini Lourentzou. CoRe3D: Collaborative reasoning as a foundation for 3d intelligence. *arXiv preprint arXiv:2512.12768*, 2025. doi: 10.48550/arXiv.2512.12768. URL https://arxiv.org/abs/2512.12768.

[321] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Onkar Susladkar, Yuanzhe Liu, Xiaona Zhou, and Ismini Lourentzou. Elsa3d: Elastic semantic anchoring for unified 3d understanding and generation, 2026. URL https://arxiv.org/abs/2607.06565.

[322] Tianjiao Yu, Xinzhuo Li, Muntasir Wahed, Jerry Xiong, Yifan Shen, Ying Shen, and Ismini Lourentzou. Dreampartgen: Semantically grounded part-level 3d generation via collaborative latent denoising, 2026. URL https://arxiv.org/abs/2603.19216.

[323] Tianjiao Yu, Vedant Shah, Muntasir Wahed, Ying Shen, Kiet A. Nguyen, and Ismini Lourentzou. Part^{2}GS: Part-aware modeling of articulated objects using 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18913–18923, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yu_Part2GS_Part-aware_Modeling_of_Articulated_Objects_using_3D_Gaussian_Splatting_CVPR_2026_paper.html.

[324] Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. SALMONN-omni: A codec-free LLM for full-duplex speech understanding and generation. *arXiv preprint arXiv:2411.18138*, 2024. doi: 10.48550/arXiv.2411.18138. URL https://arxiv.org/abs/2411.18138.

[325] Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. SALMONN-omni: A standalone speech LLM without codec injection for full-duplex conversation. In *Advances in Neural Information Processing Systems*, volume 38, pages 27885–27911, 2025. doi: 10.52202/085713-0829. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/233aee920dab065709145371b5900b8f-Abstract-Conference.html.

[326] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022. doi: 10.1109/TASLP.2021.3129994. URL https://doi.org/10.1109/TASLP.2021.3129994.

[327] Neil Zeghidour, Eugene Kharitonov, Manu Orsini, Václav Volhejn, Gabriel de Marmiesse, Edouard Grave, Patrick Pérez, Laurent Mazaré, and Alexandre Défossez. Streaming sequence-to-sequence learning with delayed streams modeling. *arXiv preprint arXiv:2509.08753*, 2025. doi: 10.48550/arXiv.2509.08753. URL https://arxiv.org/abs/2509.08753.

[328] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In *Interspeech 2019*, pages 1526–1530, 2019. doi: 10.21437/Interspeech.2019-2441. URL https://www.isca-archive.org/interspeech_2019/zen19_interspeech.html.

[329] Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot, 2024. URL https://arxiv.org/abs/2412.02612.

[330] Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Shengmin Jiang, Yuxiao Dong, and Jie Tang. Scaling speech-text pre-training with synthetic interleaved data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7b5ae891000049b91b3b62de596b1560-Abstract-Conference.html.

[331] ChengXiang Zhai. Towards a general intelligent information agent: Framework, models, and evaluation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2780–2789. ACM, 2026. doi: 10.1145/3805712.3808543. URL https://doi.org/10.1145/3805712.3808543.

[332] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

[333] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9637–9662, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.521. URL https://aclanthology.org/2024.acl-long.521/.

[334] Xiao Zhan, Guangzhi Sun, Jose Such, and Phil Woodland. Protecting bystander privacy via selective hearing in audio LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15180–15192, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.693. URL https://aclanthology.org/2026.acl-long.693/.

[335] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. WenetSpeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9746682. URL https://arxiv.org/abs/2110.03370.

[336] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.1055. URL https://aclanthology.org/2023.findings-emnlp.1055/.

[337] Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechAlign: Aligning speech generation to human preferences. In *Advances in Neural Information Processing Systems*, volume 37, pages 50343–50360, 2024. doi: 10.52202/079017-1594. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a016da670821af25f151f523a2e563f-Abstract-Conference.html.

[338] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3mRwyG5one.

[339] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient real-time understanding for long video streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21059–21069, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Flash-VStream_Efficient_Real-Time_Understanding_for_Long_Video_Streams_ICCV_2025_paper.html.

[340] Haoyang Zhang, Jun Chen, Donghang Wu, Yuxin Li, Yuxin Zhang, Xiangyu Tony Zhang, Che Liu, Qingjian Lin, Yizhou Peng, Hexin Liu, et al. Duplexsla: A full-duplex spoken language model with synchronized speech, language, and action. *arXiv preprint arXiv:2605.20755*, 2026. doi: 10.48550/arXiv.2605.20755. URL https://arxiv.org/abs/2605.20755.

[341] Haoyu Zhang, Zhipeng Li, Xiaoying Tang, Tianshu Yu, and Yiwen Guo. Ex-omni-2d: Expressive omni-modal dialogue models with native visual presence, 2026. URL https://arxiv.org/abs/2608.10720.

[342] Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chao-Hong Tan, Zhihao Du, and ShiLiang Zhang. OmniFlatten: An end-to-end GPT model for seamless voice conversation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14570–14580, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.709. URL https://aclanthology.org/2025.acl-long.709/.

[343] Xichen Zhang, Guankai Li, Yinghao Zhu, Shijian Wang, Sitong Wu, Shaozuo Yu, Meng Chu, Yuan Lu, and Jiaya Jia. Streamarena: Toward continuous, interactive, and long-horizon agentic streaming video understanding, 2026. URL https://arxiv.org/abs/2608.05703.

[344] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified speech tokenizer for speech language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=AF9Q8Vip84.

[345] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-1506. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract.html.

[346] Xiangyu Zhi, Peiqi Yin, Sheng Guan, Chenguang Zheng, James Cheng, and Xiao Yan. LiveServe: Interaction-aware serving for real-time omni-modal LLMs. *arXiv preprint arXiv:2606.22983*, 2026. doi: 10.48550/arXiv.2606.22983. URL https://arxiv.org/abs/2606.22983.

[347] Tao Zhong, Jiajun Deng, Nikita Kuzmin, Yinke Zhu, Tianxiang Cao, Tristan Tsoi, Zhili Tan, Simon Lui, and Xunying Liu. IRAF: Interference-resilient adaptive fusion for noise-robust end-to-end full-duplex spoken dialogue systems. *arXiv preprint arXiv:2606.06559*, 2026. doi: 10.48550/arXiv.2606.06559. URL https://arxiv.org/abs/2606.06559.

[348] Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. DistServe: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation*, pages 193–210, Santa Clara, CA, July 2024. USENIX Association. ISBN 978-1-939133-40-3. URL https://www.usenix.org/conference/osdi24/presentation/zhong-yinmin.

[349] Zhitong Zhou, Qingqing Zhang, Lei Luo, Jiechen Liu, and Ruohua Zhou. Open-source full-duplex conversational datasets for natural and interactive speech synthesis. *Data Intelligence*, 8(1):309–316, 2026. doi: 10.3724/2096-7004.di.2026.00d2. URL https://doi.org/10.3724/2096-7004.di.2026.00d2.