

MiniCPM-o 4.5: Towards Real-Time Full-Duplex Omni-Modal Interaction

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 Changlin Liu Hanyu Liu Ziyang Wang 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 Maoson Sun*
Zhiyuan Liu* Yuan Yao*

MiniCPM-o Team, OpenBMB

🌐 MiniCPM-o 4.5 Demo 🗨️ MiniCPM-o 4.5 Model 📄 MiniCPM-o 4.5 Code

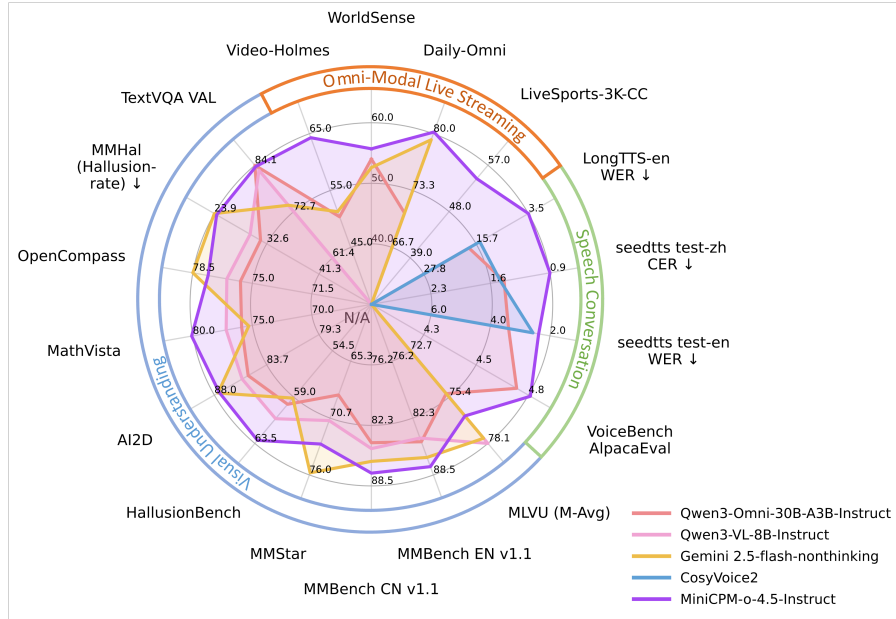


Figure 1: **Evaluation results on diverse capabilities.** MiniCPM-o 4.5 achieves state-of-the-art open-source vision-language performance at its scale, approaching Gemini 2.5 Flash. It also surpasses Qwen3-Omni-30B-A3B in omni-modal capabilities and speech generation quality.

Abstract

Recent progress in multimodal large language models (MLLMs) has brought AI capabilities from static offline data processing to real-time streaming interaction, yet they still remain far from human-level multimodal interaction. The key bottlenecks are no longer modality coverage or latency alone, but the interaction paradigm itself. First, perception and response are still separated into alternating phases, preventing models from incorporating new inputs for timely adjustment during generation. Second, most current models remain reactive, responding only to explicit user requests instead of acting proactively in the evolving multimodal environment. We present **MiniCPM-o 4.5**, our latest effort towards human-like multimodal interaction, which mitigates these gaps by real-time full-duplex omni-modal interaction.

*Corresponding authors.

It can see, listen, and speak simultaneously in real-time, while also exhibiting proactive behaviors such as issuing reminders or comments based on its continuous understanding of the live scene. The key technique behind MiniCPM-o 4.5 is **Omni-Flow**, a unified streaming framework that aligns omni-modal inputs and outputs along a shared temporal axis. This formulation converts conventional turn-based interaction into a full-duplex, time-aligned process, enabling simultaneous perception and response and allowing proactive behavior to arise within the same framework. With a total of 9B parameters, MiniCPM-o 4.5 approaches Gemini 2.5 Flash in vision-language capabilities, delivering state-of-the-art open-source performance at its scale. It also surpasses Qwen3-Omni-30B-A3B in omni-modal understanding and delivers better speech generation, with significantly higher computation efficiency. Driven by its efficient architecture design and inference optimization, the model can perform real-time full-duplex omni-modal interaction on edge devices with less than 12GB RAM cost. More importantly, MiniCPM-o 4.5 can be viewed as a representative example of a promising trend (Figure 2): Multimodal foundation models are shipping towards human-like interactive paradigms, poised to engage with the dynamic omni-modal world in the near future.



Figure 2: **Evolution of AI interaction paradigms.** AI interaction have progressed from text-only to multimodal understanding and omni live streaming. MiniCPM-o 4.5 advances this trajectory toward more human-like full-duplex interaction by enabling simultaneous perception and response.

1 Introduction

Progress in multimodal large language models (MLLMs) has enabled increasingly rich interaction over images, speech, video, and text, bringing AI systems closer to more natural forms of communication [1, 2, 3, 4] (Figure 2). The main challenge towards human-like interaction now is no longer modality coverage or response latency alone, but the underlying interaction paradigm. In current models, perception and response are still confined to alternating phases, making it difficult to continuously incorporate newly arriving information for timely adjustment during generation, as shown in Figure 3. Moreover, model behaviors remain strictly request-driven, rather than being proactively initiated from the evolving multimodal environment.

Tackling this challenge requires moving beyond turn-based passive response generation to continuous and proactive interaction. First, perception and response should remain continuously coupled in token-level over time, so that listening, watching, speaking, and writing can proceed in parallel instead of being forced into a serialized pipeline. Second, interaction should be more context-driven rather than purely reactive. Instead of waiting for explicit user triggers, a more human-like model should be able to initiate appropriate behaviors from ongoing context, such as delivering real-time scene description or offering reminders. This is particularly important in long-horizon assistance and ambient interaction.

We present MiniCPM-o 4.5, our latest effort towards human-like multimodal interaction. It can see, listen, and speak simultaneously in real-time, while also exhibiting proactive behaviors such as issuing reminders or comments based on its continuous understanding of the live scene. The key technique behind this model is Omni-Flow, a unified streaming framework that aligns multimodal inputs and outputs along a shared temporal axis. Rather than treating interaction as a sequence of distinct turns, Omni-Flow formulates interaction as a continuous full-duplex process, in which perception and response unfold in parallel and proactive behaviors can emerge from ongoing context within the same interaction loop. To fully exploit the rich omni-modal knowledge during training, MiniCPM-o 4.5 is built on an end-to-end multimodal architecture featuring token-level continuous connections. We also devise a time-aligned interleaving speech generation strategy, ensuring output speech is tightly aligned with the concurrent environment context.

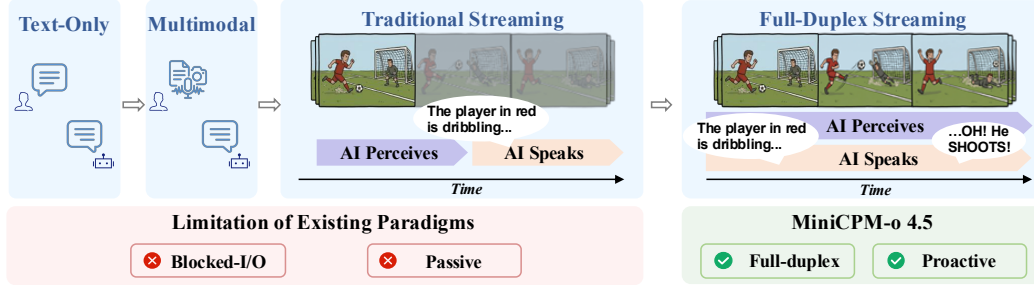


Figure 3: **From turn-based interaction to full-duplex streaming.** Existing interaction paradigms separate perception and response as alternating phases, leading to blocked information flow and passive behavior. In contrast, MiniCPM-o 4.5 continuously perceives incoming multimodal streams while speaking, allowing the model to update its response in real time and act proactively.

For better compatibility with existing infrastructure and applications, MiniCPM-o 4.5 also supports traditional turn-based interaction and can be flexibly switched between the full-duplex omni-modal streaming mode and the traditional usage mode (like MiniCPM-o 2.6 and MiniCPM-V 4.5, with upgraded performance). Extensive evaluation shows that the model achieves leading vision-language and omni-modal capabilities. With a total of 9B parameters, it approaches Gemini 2.5 Flash in vision-language capabilities, delivering state-of-the-art open-source performance at its scale. It surpasses Qwen3-Omni-30B-A3B in omni-modal understanding and also delivers higher quality speech generation. Taking advantage of its end-to-end continuous connections, MiniCPM-o 4.5 can accept multimodal system prompts that contain both text and reference audio, thus supporting advanced speech generation capabilities such as voice cloning. Moreover, MiniCPM-o 4.5 retains the strong visual strengths of the MiniCPM family, including robust OCR, low hallucination, and multilingual support.

Our contributions are three-fold: (1) We present **MiniCPM-o 4.5 9B**, the first full-duplex omni-modal LLM. It can run efficiently on edge devices with less than 12GB RAM. (2) Extensive evaluations show that MiniCPM-o 4.5 approaches Gemini 2.5 Flash in vision-language capabilities and achieves state-of-the-art open-source performance at its scale. It also surpasses Qwen3-Omni-30B-A3B in omni-modal understanding and speech generation quality, with significantly higher computational efficiency. (3) We identify continuous full-duplex and proactive multimodal interaction as a key step toward more human-like interactive intelligence, and propose the Omni-Flow framework, which aligns multimodal inputs and outputs along a shared temporal axis for full-duplex interaction modeling.

2 End-to-End Omni-Modal Architecture

MiniCPM-o 4.5 is built on an end-to-end omni-modal architecture that supports both full-duplex interaction under Omni-Flow and conventional turn-based inference. As illustrated in Figure 4, it comprises three main components: (1) **multimodal encoders** that process visual and audio inputs in an streaming manner; (2) an **LLM backbone** that performs omni-modal understanding and text generation; and (3) **speech decoders**, including an interleaved speech token decoder that autoregressively generates discrete speech tokens and a streaming flow-matching decoder that converts speech tokens into audio waveforms. All learnable components—from multimodal encoders through the LLM backbone to the speech token decoder, totaling approximately 9B parameters—are differentially connected in token-level, enabling end-to-end gradient propagation and joint optimization across modalities during training. Detailed architectural configurations are provided in Appendix A.

Visual Encoding. MiniCPM-o 4.5 adopts the LLaVA-UHD [5] image partitioning strategy to encode any aspect high-resolution images and improve compression rate with a resampler module [1]. We adopt a max resolution of 448×448 for the full-duplex streaming mode and otherwise 2240×2240 . Specifically, each image is first divided into slices, and each slice is then encoded into 1024 tokens by a SigLIP ViT [6] (0.4B) and compressed into 64 tokens by the resampler module. This yields a $16 \times$ token compression ratio, which is higher than the common $4 \times$ compression [7, 3, 4], enabling substantially more efficient visual processing.

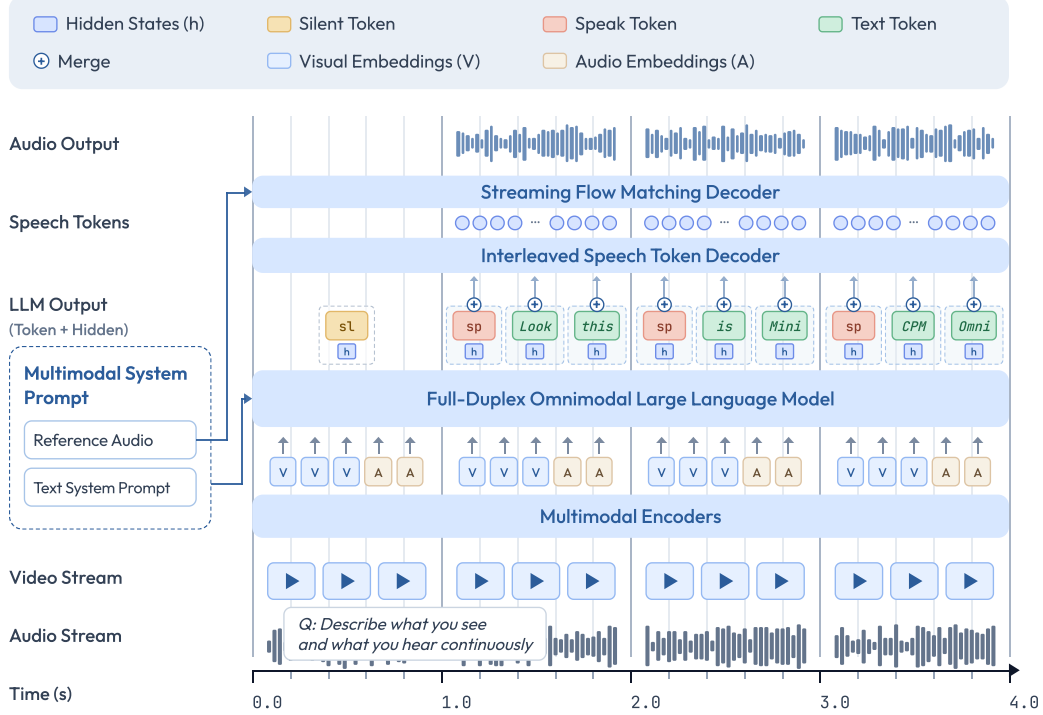


Figure 4: **End-to-end omni-modal architecture of MiniCPM-o 4.5.** Modality encoders, the LLM backbone, and speech decoders are connected through token-level hidden states in an end-to-end trainable architecture, with multimodal input and output streams aligned on a shared millisecond-level timeline for full-duplex streaming interaction.

Audio Encoding. A Whisper Medium [8] encoder (0.3B) encodes input audio in a chunk-based streaming fashion [9], producing 50 feature tokens per second. We then use a two-layer MLP projector to conduct a $5\times$ temporal compression, resulting in 10 audio tokens per second for the LLM backbone, reducing the token budget.

Text Decoding. The LLM backbone (Qwen3-8B [10]) generates text outputs and hidden states for speech generation. Since the LLM backbone only generate tokens in text domain, it requires just 3-4 decoding steps per second (i.e., human speech speed) during real-time full-duplex interaction. When backbones are instead required to directly generate speech tokens (typically about 25 tokens per second), as in recent works [11, 12], the efficiency can be significantly impeded, and the core language capabilities also tend to degrade [13, 14]. Our design avoids this by delegating speech token production to lightweight speech decoders described below.

Speech Token Generation. Speech generation demands not only correct pronunciation but also prosody and style shaped by context and instructions. We address this by leveraging the contextual understanding capability of the LLM backbone. For each text token passed to the lightweight Llama speech token decoder ($\sim 0.3B$), we sum its LLM backbone hidden states (reshaped by an MLP layer) and its speech decoder for further S3 [15] token generation. With prosodic decisions pre-encoded by the LLM backbone, the small speech decoder can devote its capacity to speech modeling. Moreover, input text tokens and output speech tokens are interleaved in a time-aligned manner to ensure output speech tightly couples with the concurrent environment context as detailed in Section 3.4.

Waveform Synthesis. A streaming flow-matching decoder [16, 12] converts generated S3 speech tokens into audio waveforms, based on the reference audio in the multimodal system prompt.

3 Omni-Flow

In existing interaction paradigms, perception and response are confined to alternating phases, resulting in the blocked I/O and passive responding problem as illustrated in Figure 3. To enable models to perceive and speak simultaneously, we propose the Omni-Flow framework that coordinates omni-modal input and output streams with a shared temporal axis. Inspired by the time-division multiplexing technique, Omni-Flow partitions the continuous interaction into fine-grained time windows of duration t . Within each window, the model incorporates newly arrived signals while producing the next output, converting conventional turn-taking into a stream of time-local updates as shown in Figure 4. As t becomes sufficiently small, perception and response become tightly coupled in time, naturally approximating full-duplex behavior.

3.1 Time-Aligned Streams

We identify three time-aligned streams in the interaction: **env-visual**, which carries live visual observations of the environment; **env-audio**, which carries the acoustic scene, including user speech when present; **out-stream**, which represents the assistant’s text and speech outputs. Under this view, user requests are no longer treated as a privileged conversational role, but instead become part of the continuously observed world state, entering primarily through env-audio. Likewise, the model does not rely on explicit requests as the trigger before responding. Instead, the out-stream evolves coupled to ongoing perception. The model is therefore situated in an always-on multimodal environment, where it must determine not only *what* to output, but also *whether* and *when* to output on its own.

3.2 Unified Serialization

Given these streams, we organize them into a unified sequence that can be passed to a standard causal language model. For the k_{th} time chunk, inputs from env-visual and env-audio are encoded into visual token sequence $\mathbf{v}^k$ and audio token sequence $\mathbf{a}^k$, while updates in out-stream are represented as an output token sequence $\mathbf{o}^k$. When no output should be produced, $\mathbf{o}^k$ contains only a special [listen] token. We group these time-aligned tokens into $\mathbf{g}_k = [\mathbf{v}^k; \mathbf{a}^k; \mathbf{o}^k]$, and serialize the interaction by concatenating consecutive groups into a single sequence. Within each chunk, the model first processes newly arrived perceptual tokens and then generates output tokens, so that every output is conditioned on the most recent observation. Reducing the chunk size t increases the rate at which the model refreshes its perception, keeping it more closely aligned with the evolving environment. Since the model determines whether to output in each time window, it naturally supports proactive behavior and reduces the reliance on external VAD [17] modules.

3.3 Design Tradeoffs

Omni-Flow introduces several design choices that directly affect the stability and responsiveness of the model. We therefore conduct ablations along three dimensions: temporal granularity, boundary explicitness, and control formulation. Temporal granularity specifies the duration of each time chunk (1.0 s, 0.2 s, or 0.1 s). Boundary explicitness specifies whether consecutive groups are separated by explicit special tokens or not. Control formulation specifies how the model decides whether to speak: in the Listen-Speak (LS) formulation, the model first predicts a binary listen/speak control token before content generation; in the Listen-Text (LT) formulation, the model directly predicts either [listen] or normal text tokens in a shared output space. Results are shown in Table 1.

Table 1: Ablation of full-duplex design choices.

Chunk Size	Boundary	Control	AdvBench	AlpacaEval	IFEval	SDQA	MMLU
1.0 s	Explicit	LS	0.98	3.56	0.29	0.36	0.65
1.0 s	Explicit	LT	0.92	3.60	0.24	0.35	0.56
1.0 s	Implicit	LT	0.96	3.31	0.22	0.28	0.45
0.2 s	Explicit	LS	0.81	1.22	0.10	0.09	0.45
0.1 s	Explicit	LS	0.67	2.40	0.10	0.13	0.32

Temporal granularity governs the central latency-capacity tradeoff. Reducing the chunk size improves temporal responsiveness, but also leaves less modeling budget within each chunk for control

and generation. When chunks become too short, the model no longer has sufficient information for each time window to make stable decisions and produce coherent outputs, leading to substantial degradation. In our setting, a chunk size of 1.0 s provides the best balance.

Boundary explicitness is consistently beneficial. Explicitly marking the boundary between groups performs better. This suggests that distinguishing newly observed inputs from newly generated outputs is a nontrivial problem, and making this structure explicit can reduce the burden on the model.

Separating interaction control from content generation leads to more stable modeling. LS outperforms LT, indicating that deciding *whether* to speak should be decoupled from deciding *what* to say, and entangling both in a single prediction step makes full-duplex interaction harder to learn.

3.4 Time-Aligned Interleaving for Timely Speech Generation

Omni-Flow represents model outputs as a stream that evolves together with incoming inputs. However, maintaining temporal alignment between the spoken output and the latest observed context remains nontrivial. The difficulty comes from the mismatch between text generation time and speech playback time: if the text generated within an m -second interval takes much longer than m seconds to vocalize, the speech stream will progressively lag behind the model’s evolving state. As a result, the audio heard at a given moment may correspond to text generated much earlier, making the response temporally stale with respect to the ongoing interaction. This issue is further complicated by the fact that the vocalization duration of each text token is variable and context-dependent.

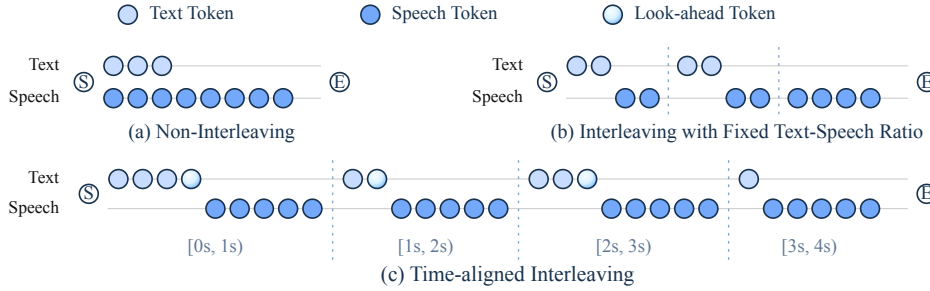


Figure 5: **Comparison of streaming speech generation strategies.** Existing methods either (a) maintain a large text lead or (b) rely on a fixed text-speech ratio, making the spoken content lag behind the evolving environment. We propose Time-Aligned Interleaving (TAIL), which adaptively interleaves text and speech so that the text generated in each time chunk corresponds to approximately the same duration of speech playback.

Existing streaming speech generation methods [11, 7, 18, 16] typically adopt one of two strategies shown in Figure 5 (a) and (b). Some methods first generate a relatively long span of text and then synthesize speech from it. Others interleave text and speech using a fixed text-to-speech token ratio. While both strategies can produce high-quality speech, they do not explicitly align the generated speech with the interaction timeline. The former allows text to run far ahead of playback, while the latter assumes a nearly fixed correspondence between text tokens and speech duration. In full-duplex interaction, both designs can cause the model to keep speaking content that is stale and not aligned with the concurrent environment.

To address this, we propose **Time-Aligned Interleaving (TAIL)**, a chunk-wise speech generation strategy that adaptively controls how much text to generate at each step. Rather than matching each chunk independently to a fixed speech duration, TAIL considers the accumulated playback progress over the entire interaction. At the k_{th} chunk, the model adjusts the amount of text to generate so that, after vocalizing the newly generated content, the speech stream approaches the current time boundary kt . If previous chunks have already introduced a slight playback delay, the model can adaptively generate fewer text tokens in the current chunk to let speech catch up. In this way, TAIL keeps the spoken response close to the model’s latest state instead of allowing text to run far ahead of audio.

We construct TAIL supervision from full-duplex streaming training data by collecting the start and end times of each text token. Tokens whose start times fall into $[(k-1)t, kt)$, together with their

corresponding speech tokens, are assigned to the k_{th} Omni-Flow chunk. This format teaches the model to learn a history-dependent interleaving pattern, where the number of text tokens in each chunk can vary according to the accumulated playback alignment.

Look Ahead Speech Generation. Speech generation may still require a limited future text context. For example, the pronunciation of “the” depends on the following word, as in “the apple” versus “the car”. TAIL therefore uses a bounded look-ahead mechanism: the speech tokens of the last few text tokens in chunk k are deferred to chunk $k + 1$, while the remaining tokens are spoken in chunk k . This provides local context for pronunciation and prosody without letting the text stream run substantially ahead of playback. As a result, TAIL preserves the time-aligned structure of Omni-Flow while enabling continuous and timely speech generation.

4 Data

4.1 Speech Data

We collect large-scale natural speech data for broad capability coverage and high-quality dialog data for controllable natural speech generation.

Large-scale Natural Speech Data. We process millions of hours of unlabeled speech data collected from diverse sources through a pipeline integrating multiple open-source components [19, 20, 21, 22, 23], yielding training sets for zero-shot TTS, ASR, and multi-turn multi-speaker dialogue. This diverse corpus encompasses a broad range of different speakers, accents, and conversational patterns.

Spoken Dialog Data. We first use a text-based LLM to generate colloquial, instruction-following dialogue from diverse seed queries. A subset of these dialogues is then re-recorded by professional voice actors under studio conditions. In the recording sessions, voice actors deliver in a conversational style rather than reading scripts verbatim, balancing structured content with improvised expression while varying emotion, speaking rate, and emphasis under a consistent vocal identity. The resulting corpus covers instruction-following TTS, question answering, and multi-turn natural dialogue.

4.2 Vision-Language Data

We introduce the vision-language data of MiniCPM-o 4.5 in this section. Building upon the data system of MiniCPM-V 4.5, we further expand the scale and improve the quality to cover broader task types and real-world scenarios.

High-Quality Knowledge and Alignment Data. We update the generator model used in the CapsFusion [24] pipeline to synthesize more informative image captions, and further refine our filtering process by improving image-text relevance estimation.

Complex Document and OCR Data. To better utilize document knowledge, we extend the unified document knowledge and OCR learning approach of MiniCPM-V 4.5 with a relevance-aware masking strategy. Specifically, instead of randomly masking text regions, we prioritize regions that are more relevant to figures and charts in document images. This encourages the model to focus more on visually grounded content, while reducing the proportion of training cases that can be solved primarily from textual context alone.

Real-World Scenarios Data. Capturing the nuances of practical user interactions is a core focus of our data curation. We introduce more natural and diverse query patterns. We significantly improve the depth and readability of model responses by rewriting short, direct-answer samples into detailed, chain-of-thought-style rationales. In addition, a reward-model-based filtering pipeline is applied to ensure overall data quality and alignment with human preferences.

Dense Video Perception Data. To strengthen the model’s video perception and cross-frame reasoning abilities, we construct a dense video captioning dataset which provides continuous, fine-grained descriptions of temporal events, human actions, and complex scene transitions.

Text-only Data. We also incorporate high-quality text-only instruction data from the MiniCPM 4.1 [25] post-training data set to maintain robust linguistic capabilities.

4.3 Omni-Modal Full-Duplex Data

Our omni-modal full-duplex data includes both large-scale web data and a smaller set of high-quality instruction samples. Each training sample contains the full visual input, audio input, output text and output speech, where each piece of information is tagged with a time index.

Large-scale Web Audio-video Data. We collect a large scale of web audio-video data to provide broad coverage of real-world full-duplex scenarios. Segments dominated by single-speaker speech or that have weak audio-visual relevance are filtered out. To further improve quality, we apply OCR-based subtitle removal [26], talking-head detection [27], and filtering over ASR-derived transcripts, reducing misleading shortcuts and low-information or noisy segments.

Full-Duplex Task Data. To support target full-duplex capabilities that require more precise interaction, we manually construct multiple scenarios and annotate corresponding instruction-following data. Based on these high-quality task samples, MiniCPM-o 4.5 supports advanced capabilities like continuous scene description and proactive reminding.

5 Training

In this section, we present the overall training pipeline for MiniCPM-o 4.5. One of the key challenges in advancing omni-modal capabilities is to retain the fundamental advantages of individual modalities while supporting efficient and seamless generalization across modalities. To this end, we design a carefully staged pipeline to progressively integrate speech into the multimodal system in a smooth and stable manner. Based on a pretraining checkpoint of MiniCPM-V 4.5. The pipeline first conducts speech pretraining to establish foundational audio understanding and speech generation capabilities. We then perform joint pretraining to construct unified cross-modal representations. Supervised fine-tuning is further employed to enable natural instruction following and high-quality interactions across text, speech, image, and video. Finally, we apply reinforcement learning to further improve reasoning abilities and mitigate hallucinations.

5.1 Speech Pretraining

MiniCPM-o 4.5 is initialized with a pretrained Whisper encoder and the pretraining checkpoint of MiniCPM-V 4.5, together with randomly initialized speech-related modules, including an audio projector, an LLM-to-speech projector, and a speech decoder. To preserve the backbone’s visual and linguistic capabilities, we freeze the pretrained components and update only newly added modules. This stage aligns Whisper features with the LLM hidden space and trains the speech decoder to transform LLM backbone hidden states into semantically and prosodically grounded speech tokens.

5.2 Joint Pretraining

In the second stage, we unfreeze all parameters and conduct joint pretraining on a balanced mixture of vision-language, speech, and omni-modal data. To stabilize optimization, we assign different modality combinations to different data-parallel ranks, ensuring a fixed data ratio at every training step. Besides conventional turn-based samples, the mixture includes proactive and full-duplex interaction data, where text tokens are aligned with speech and visual signals on a shared timeline. Trained with a unified next-token prediction objective, the model acquires real-time omni-modal interaction capabilities while maintaining its foundational visual understanding.

5.3 Joint Supervised Fine-Tuning

The joint supervised fine-tuning stage activates omni-modal capabilities and strengthens instruction following. It consists of two phases: large-scale instruction tuning for broad capability adaptation, followed by high-quality human-annotated tuning for fine-grained behavioral refinement. To enable flexible quality-efficiency trade-offs during inference, we augment omni-modal data with varying resolutions and frame rates, randomly setting the maximum frame resolution to 0.2–0.4 megapixels and sampling the frame rate uniformly from 1–5 FPS.

5.4 Reinforcement Learning

We further improve MiniCPM-o 4.5 with reinforcement learning. We first apply GRPO [28] to enhance reasoning and instruction following, using answer accuracy together with auxiliary rewards such as format reward. For accuracy rewards, we combine rule-based verification with an efficient judge model [29] to improve the recall of correct responses.

To improve token efficiency, we introduce a smooth length reward adapted from Kimi-K1.5 [30]:

$$r_{\text{len}}(i) = \begin{cases} s_i, & r_i = 1, \\ \min(0, s_i), & r_i = 0, \end{cases} \quad s_i = \left(0.5 - \frac{\ell_i - \ell_{\min}}{\ell_{\max} - \ell_{\min}}\right) \times \min\left(1, \frac{\ell_{\max} - \ell_{\min}}{\tau}\right). \quad (1)$$

Here, r_i is the correctness indicator, and $\ell_i, \ell_{\min}, \ell_{\max}$ are computed over responses to the same prompt. The $\min(0, s_i)$ term avoids rewarding short incorrect responses, and τ downscales the reward when length differences are small. We also include a general reward model to improve answer quality and suppress unintended code-mixing. For convergence efficiency, we do not include the length reward for the first 480 training steps.

Finally, we apply RLAI-F-V [31] to reduce hallucinations in visual scenarios. We find that hallucination mitigation learned from image-text data transfers effectively to omni-modal full-duplex interaction, reducing hallucinations in streaming settings as well.

6 Evaluation

In this section, we comprehensively evaluate MiniCPM-o 4.5 and other baseline models.

6.1 Modalities and Domains

We evaluate MiniCPM-o 4.5 across four modality capability groups: vision-language understanding, speech understanding and generation, text capability, and omni-modal streaming interaction. Vision-language understanding is further divided into five representative domains: STEM and general multimodal reasoning, document and OCR understanding, multi-image reasoning, hallucination, and video understanding. Speech evaluation covers both speech understanding and speech generation. Text evaluation measures whether the model preserves the language capabilities of its LLM backbone after omni-modal training. Omni-modal and streaming interaction evaluation covers both turn-based omni-modal understanding and full-duplex streaming interaction.

Vision-Language Understanding. We evaluate vision-language understanding across five representative domains. (1) *STEM and general multimodal reasoning.* For general vision-language comprehension, we include OpenCompass [32], MMBench V1.1 [33], MMVet [34], and MMStar [35], which cover diverse multimodal tasks. For STEM-oriented reasoning, we include MMMU [36], MathVista [37], and A12D [38], covering scientific knowledge, mathematical reasoning, and diagram understanding. We further include MMT-Bench [39] and MM-IFEval [40] to assess multitask generalization and multimodal instruction following. (2) *Document and OCR understanding.* This domain evaluates the ability to recognize, extract, and reason over text in visually rich documents and scene images. We use OCRBench [41], TextVQA [42], DocVQA [43], and OmniDocBench [44], which require joint modeling of textual content, visual layout, and document structure. (3) *Multi-image understanding.* This domain measures the ability to aggregate and compare information across multiple images. We adopt Mantis-Eval [45], MUIRBench [46], and MMSI-Bench [47], which evaluate cross-image reasoning, visual comparison, and multi-image information integration. (4) *Hallucination.* This domain evaluates whether model responses remain faithful to the visual input. We use HallusionBench [48] and MMHal-Bench [49], which measure visual consistency and hallucination in multimodal generation. (5) *Video understanding.* This domain evaluates spatio-temporal reasoning and motion understanding in videos. We use Video-MME [50], LVBench [51], MLVU [52], LongVideoBench [53], and MotionBench [54], covering both varying video lengths.

Speech Understanding and Generation. Speech evaluation covers automatic speech recognition, speech translation, audio understanding, speech question answering, and speech generation. For speech understanding, we evaluate on standard ASR benchmarks, including AISHELL-1 [55], AISHELL-2 [56], WenetSpeech [57], LibriSpeech [58], GigaSpeech [59], and VoxPopuli [60]; speech translation on CoVoST 2 [61]; multi-task audio understanding on MMAU and MELD [62]; and

Table 2: Vision-language results (instruct mode).

Benchmark Size	Gemini 2.5 Flash -	InternVL3.5 8B	Qwen3-VL 8B	Qwen3-Omni 30B-A3B	MiniCPM-o 4.5 9B
STEM & General					
OpenCompass	78.5	75.8	76.5	75.7	77.6
MMBench EN v1.1	86.6	79.5	84.5	84.9	87.6
MMBench CN v1.1	86.0	80.0	84.7	84.1	87.2
MathVista	75.3	78.4	77.2	75.9	80.1
MMVet	81.4	83.1	73.7	74.8	74.4
MMMUS	76.3	73.4	69.6	69.1	67.6
MMStar	75.8	69.3	70.9	68.5	73.1
AI2D	87.7	84.0	85.7	85.2	87.6
MMT-Bench (val)	70.0	66.7	60.9	70.4	69.7
MM-IFEval	75.8	56.3	59.4	65.7	66.3
Document & OCR					
OCRBench	864	840	896	880	876
TextVQA (val)	74.3	78.2	82.9	84.1	83.8
DocVQA (val)	93.0	92.3	96.1	95.4	94.7
OmniDocBench (EN)↓	0.214	0.322	0.255	0.216	0.109
OmniDocBench (CN)↓	0.290	0.416	0.319	0.363	0.162
Hallucination					
HallusionBench	59.1	54.5	61.1	59.7	63.2
MMHal-Score	4.6	3.8	4.7	4.6	4.7
MMHal-Hallrate↓	23.9	34.7	29.9	31.6	24.3
Multi-Image					
Mantis-Eval	72.8	70.5	74.2	78.3	79.7
MUIRBench	74.5	55.8	64.4	61.9	72.0
MMSI-Bench	12.1	–	11.3	14.2	16.6
Video					
Video-MME (w/o subs)	75.6	66.0	71.4	70.5	70.4
LVBench	62.2	–	58.0	50.2	50.9
MLVU (M-Avg)	77.8	70.2	78.1	75.2	76.5
LongVideoBench (val)	–	62.1	66.4	66.9	66.0
MotionBench	–	62.3	59.5	61.7	61.4

spoken question answering on VoiceBench [63], Speech TriviaQA [64], Speech Web Questions [65], and Speech CMMU [66]. For speech generation, we evaluate speech quality, intelligibility, speaker similarity, long-form generation, and emotion/style control using SeedTTS Test [67], LongTTS [68], Expresso [69], and ESD [70].

Text Capability. We compare MiniCPM-o 4.5 with its language backbone, Qwen3-Instruct-8B [10], to assess whether omni-modal training preserves core text abilities. Our benchmark suite spans instruction following, world knowledge, multilingual understanding, reasoning, and code generation. Specifically, we use IFEval [71] for instruction following; MMLU [72] and CMMLU [73] for knowledge and multilingual understanding; BBH [74], MATH-500 [75], and GSM8K [76] for reasoning and mathematics; and HumanEval [77] and MBPP [78] for code generation.

Omni-modal and Streaming Interaction. We evaluate omni-modal understanding on benchmarks where video and audio input streams are naturally time-aligned, including Daily-Omni [79], WorldSense [80], Video-Holmes [81], JointAVBench [82], AVUT-Human [83], FutureOmni [84], and Video-MME-Short with audio [50]. For full-duplex streaming, the model must continuously perceive incoming streams while producing timely responses. Due to the limited availability of benchmarks for real-time omni-modal full-duplex interaction, we report results on LiveSports-3K-CC [27], an audio-free full-duplex benchmark. Qualitative demonstrations involving simultaneous vision, speech, and text streams are provided on our demo website.

6.2 Vision-Language Results

As shown in Table 2 and Table 3, MiniCPM-o 4.5 demonstrates strong performance across a wide range of vision-language tasks under both instruct and thinking modes.

Table 3: Vision-language results (thinking mode).

Benchmark Size	Gemini 2.5 Flash	GPT-5	Qwen3-VL 8B	Qwen3-Omni 30B-A3B	MiniCPM-o 4.5 9B
STEM & General					
OpenCompass	79.9	79.7	77.3	78.5	78.2
MMBench EN v1.1	87.1	85.5	85.3	88.2	89.0
MMBench CN v1.1	87.3	85.6	85.5	87.7	87.6
MathVista	79.4	81.9	81.4	80.0	81.0
MMVet	81.2	77.6	69.8	74.8	73.6
MMMUS	77.7	81.8	74.1	75.6	70.2
MMStar	76.5	75.7	75.3	74.9	73.6
HallusionBench	63.5	65.2	65.4	62.8	62.6
AI2D	88.7	89.5	84.9	86.1	88.5
MMT-Bench (val)	70.7	72.7	68.1	70.9	69.7
MM-IFEval	75.7	83.1	73.5	69.9	68.2
Document & OCR					
OCRBench	853	807	819	859	879
TextVQA (val)	73.8	77.8	77.8	80.8	79.8
DocVQA (val)	92.8	91.3	95.3	94.2	92.3

Comprehensive Capability. MiniCPM-o 4.5 achieves an average score of 77.6 on OpenCompass [32], a comprehensive collection of 8 popular vision-language benchmarks, in instruct mode and 78.2 in thinking mode. With only 9B parameters, it consistently outperforms models of similar scale, such as InternVL3.5-8B [85] and Qwen3-VL-8B [4], as well as larger models like Qwen3-Omni-30B [7], while close to leading proprietary models including Gemini 2.5 Flash [86] and GPT-5 [87].

OCR and Document Analysis. MiniCPM-o 4.5 exhibits the best performance in document parsing. It achieves strong results on OmniDocBench [44] for both English and Chinese, significantly outperforming other general models with larger parameter size, such as Qwen3-Omni-30B-A3B. On OCRBench [41], TextVQA [42], and DocVQA [43], MiniCPM-o 4.5 is on par with top-tier models.

Multi-Image Understanding. Benefiting from enhanced data coverage and quality of multi-image datasets, MiniCPM-o 4.5 outperforms all baselines on Mantis-Eval [45] and MMSI-Bench [47] as shown in Table 2. It also yields a competitive score on MUIRBench [46]. These results indicate strong performance on cross-image understanding, which is essential for real-world applications.

6.3 Speech Results

Audio Understanding. As shown in Table 4, MiniCPM-o 4.5 demonstrates broad audio understanding capability. On ASR, it remains close to the leading systems across both Chinese and English benchmarks, with the best results on GigaSpeech and VoxPopuli. More importantly, its advantages extend to semantic speech tasks. MiniCPM-o 4.5 leads on CoVoST 2 en→zh, MELD, VoiceBench AlpacaEval, and Speech TriviaQA, indicating that the model can leverage speech-conditioned representations for translation, audio reasoning, instruction following, and knowledge-intensive speech QA. At the same time, the remaining gaps on Speech Web Questions and Speech CMMU show that retrieval-like factual QA and Chinese speech knowledge QA are still challenging.

Speech Generation. As shown in Table 5, MiniCPM-o 4.5 demonstrates clear advantages in speech clarity and expressive control. It achieves the lowest CER/WER on SeedTTS Test-ZH and SeedTTS Test-EN, showing reliable bilingual speech generation. On LongTTS, it obtains a much lower English WER than the baselines, indicating better stability for long-form English generation, while remaining close to CosyVoice2 on Chinese CER. It also performs best on Espresso and ESD, suggesting stronger emotion and style control for expressive speech synthesis.

6.4 Text Results

As shown in Table 6, MiniCPM-o 4.5 outperforms its backbone LLM in most text-only tasks, specifically across complex reasoning, mathematics, coding, and instruction following. This suggests

Table 4: Results on audio understanding benchmarks. For ASR benchmarks, lower is better; *: VoiceBench AlpacaEval scores are rated on a scale from 1 to 5.

Benchmark Size	Kimi-Audio 9B	Qwen3-Omni 30B-A3B	MiniCPM-o 4.5 9B
Automatic Speech Recognition			
AISHELL-1↓	0.6	0.6	0.9
AISHELL-2↓	2.6	2.3	2.5
WenetSpeech test-net↓	6.3	4.7	5.9
WenetSpeech test-meeting↓	5.4	5.9	5.7
LibriSpeech test-clean↓	1.3	1.2	1.4
LibriSpeech test-other↓	2.4	2.5	2.8
GigaSpeech test↓	9.4	8.7	8.5
VoxPopuli V1-En↓	8.0	6.4	6.2
Speech Translation			
CoVoST 2 en→zh	36.6	46.6	49.9
CoVoST 2 zh→en	18.3	29.4	26.4
Multi-task Audio Understanding			
MMAU	68.4	77.5	76.9
Meld	59.1	56.8	60.2
Speech Question Answering			
VoiceBench AlpacaEval*	4.46	4.74	4.81
Speech TriviaQA	41.9	62.9	75.5
Speech Web Questions	46.4	74.9	70.2
Speech CMMU	67.0	47.8	59.2

Table 5: Speech generation results. Lower is better for CER and WER; N/A: not supported; *: Neutral reference audio is used for evaluation.

Model	SeedTTS Test-ZH		SeedTTS Test-EN		LongTTS		Emotion/Style Control	
	CER↓	SIM-o	WER↓	SIM-o	EN WER↓	ZH CER↓	Expresso*	ESD*
CosyVoice2	1.45	74.8	2.57	65.2	14.80	5.27	17.9	53.4
Qwen3-Omni	1.41	N/A	3.39	N/A	17.33	18.99	N/A	N/A
MiniCPM-o 4.5	0.86	74.5	2.38	64.9	3.37	6.58	29.8	82.1

that a strategic balance of textual and multimodal data allows the model to retain its text capabilities while acquiring strong multimodal capabilities.

6.5 Omni-modal and Streaming Results

Omni-modal Understanding. MiniCPM-o 4.5 demonstrates strong omni-modal understanding capabilities as shown in table 7. It achieves the best results on five of the seven benchmarks, namely Daily-Omni, WorldSense, Video-Holmes, JointAVBench, and AVUT-Human. Despite its small parameter-size, it remains competitive on FutureOmni and Video-MME-Short (w/ audio).

Full-Duplex Results. Table 8 evaluates whether models can respond appropriately while continuously receiving visual streams. MiniCPM-o 4.5 achieves a win rate of 54.4 on LiveSports-3K-CC, outperforming LiveCC and StreamingVLM by 12.9 and 8.8 points, respectively. This improvement suggests that Omni-Flow is effective for continuous visual interaction: by organizing perception and

Table 6: Results on text benchmarks.

Model	IFEval-PLS	BBH	CMMLU	MMLU	HumanEval	MBPP	Math500	GSM8K	Avg
Qwen3-8B-Instruct	83.0	69.4	78.7	81.7	86.6	75.9	84.0	93.4	81.6
MiniCPM-o 4.5	84.7	81.1	79.6	77.0	86.6	76.7	77.0	94.5	82.1

Table 7: Omni-modal benchmark results in simplex settings.

Benchmark Size	Gemini 2.5 Flash	Qwen3-Omni	MiniCPM-o 4.5
	-	30B-A3B	9B
Daily-Omni	79.3	70.7	80.2
WorldSense	52.6	54.0	55.7
Video-Holmes	51.3	50.4	64.3
JointAVBench	55.6	53.1	60.0
AVUT-Human	65.4	74.2	78.6
FutureOmni	55.6	62.1	56.1
Video-MME-Short (w/ audio)	85.5	81.3	84.7

Table 9: Ablation results of different length rewards.

Length Reward	Benchmarks Avg.		Length Reduction Avg.	
	Thinking	Instruct	Thinking	Instruct
No Length Reward	73.5	70.9	—	—
Kimi K1.5-Style [30]	73.0	70.1	50.7%	20.2%
Ours	74.3	70.9	35.3%	20.5%

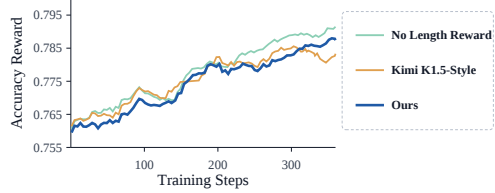


Figure 6: Training set accuracy using different length penalty methods.

response along a shared timeline, the model can better ground its responses in the evolving scene instead of relying on delayed or fragmented visual context.

Table 8: Vision-only full-duplex benchmark results.

Benchmark Size	LiveCC	StreamingVLM	MiniCPM-o 4.5
	8B	8B	9B
LiveSports-3K-CC	41.5	45.6	54.4

6.6 Analysis

Ablation of Length Reward. We ablate the length penalty design to examine the trade-off between response efficiency and task performance. We compare the length penalty used in Kimi K1.5 [30] with our proposed smooth length penalty. As shown in Table 9, the K1.5 penalty reduces response length by 49.2%, but lowers the average benchmark score from 75.0 to 74.5. In contrast, our length penalty achieves a milder length reduction of 35.3% while improving the average score to 75.8, suggesting a better efficiency-performance trade-off. The training curves in Figure 6 help explain the degradation under the K1.5 penalty. Its training accuracy drops in the later stage, indicating that an aggressive length penalty can compete with the accuracy reward and suppress further optimization. Our method avoids this instability through variance-aware smoothing, which reduces the penalty strength when length differences are small. The performance gain of our method appears to come from a different effect. We observe that the smoother penalty reduces unnecessary reasoning on perception-heavy tasks, where overly long analyses can introduce redundant or even misleading intermediate steps. This may partly explain the improvement on HallusionBench, where our method increases the score from 60.6 to 63.2.

Comparison of Speech Generation Modes.

Table 10 compares three speech generation modes: non-interleaved generation, our fixed-text interleaving, and our dynamic-text interleaving strategy TAIL. Fixed-text interleaving achieves the best CER/WER, suggesting that chunked streaming generation can improve pronunciation accuracy over synthesizing speech after the full text is generated. TAIL is designed

Table 10: MiniCPM-o 4.5 speech generation quality of different modes. We report results on Seed TTS test set.

Interleaving Mode	ZH CER↓	ZH SIM-o↑	EN WER↓	EN SIM-o↑
No interleave	1.44	74.1	2.70	64.9
Fixed text	0.86	74.5	2.38	64.9
Dynamic text (TAIL)	1.04	74.1	3.93	65.1

Table 11: Efficiency comparison of different inference frameworks. We report the real-time factor on different hardware configurations. Lower results indicates higher inference efficiency.

Framework	RTX 4090	RTX 5080	RTX 5070	DGX Spark
PyTorch	OOM	OOM	OOM	2.42
llama.cpp-omni (Ours)	0.21	0.32	0.40	0.17

for the more challenging full-duplex setting, where text and speech must stay temporally aligned. Although it slightly sacrifices recognition accuracy, especially on English WER, it maintains reasonable overall speech quality, hitting a practical trade-off between streaming interaction and speech generation quality.

7 Efficient Real-Time Inference

To support real-time full-duplex omni-modal streaming interaction in practical deployment scenarios, we develop an efficient inference framework based on llama.cpp [88], termed *llama.cpp-omni*. The framework is tailored to the streaming interaction paradigm of MiniCPM-o 4.5, enabling smooth execution across multiple hardware platforms. Beyond improving runtime efficiency, we also validate its generalization across different operating systems, including macOS, Windows, and Linux. We further provide a lightweight demo system, allowing users to quickly deploy MiniCPM-o 4.5 on their own hardware and smoothly experience its diverse capabilities, including real-time speech, vision-language, and full-duplex omni-modal interaction. Table 11 compares the real-time factor (RTF) of different inference frameworks across several hardware configurations. Compared with the PyTorch implementation, *llama.cpp-omni* substantially reduces RTF on all evaluated devices and requires less memory cost.

8 Conclusion

Contributions. We present MiniCPM-o 4.5, a 9B open-source MLLM for real-time full-duplex omni-modal interaction. By continuously perceiving visual and auditory streams while generating speech responses, MiniCPM-o 4.5 moves beyond conventional turn-based multimodal interaction and enables a more human-like interaction paradigm. It achieves this capability with practical edge efficiency, requiring less than 12GB RAM during deployment, while also approaching Gemini 2.5 Flash in vision-language capabilities and delivering frontier image and video understanding performance among open-source MLLMs at this scale. We further introduce the unified omni-modal streaming framework Omni-Flow, as the key technique behind MiniCPM-o 4.5, that aligns multimodal inputs and outputs along a shared temporal axis, providing a general formulation for full-duplex and proactive multimodal interaction.

Limitations. MiniCPM-o 4.5 is still an early exploration of real-time full-duplex omni-modal interaction and remains limited in several aspects. First, its foundation capability and robustness in long, dynamic real-world streaming interactions still require further improvement and validation. Second, speech generation in omni-modal streaming mode can occasionally be unstable, including mispronunciation or unintended mixing between English and Chinese. Third, although our web demo enables convenient access, users may experience increased latency or missing output fragments under unstable network conditions; local deployment with *llama.cpp-omni* can better support smooth real-time interaction. Finally, the model’s proactive behavior is still relatively simple, leaving richer context-aware planning and self-initiated assistance for future work.

References

- [1] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *ArXiv preprint*, abs/2408.01800, 2024.
- [2] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan,

- Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025.
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025.
 - [4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025.
 - [5] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer, 2024.
 - [6] 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 (ICCV)*, pages 11975–11986, October 2023.
 - [7] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025.
 - [8] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 23–29 Jul 2023.
 - [9] Zhuoyuan Yao, Di Wu 0061, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *Interspeech*, volume 2021, pages 4054–4058, 2021.
 - [10] Qwen Team. Qwen3 Technical Report, 2025.
 - [11] Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming, 2024.
 - [12] Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, Yuxin Zhang, Zhao You, Brian Li, Changyi Wan, Hanpeng Hu, Jiangjie Zhen, Siyu Chen, Song Yuan, Xuelin Zhang, Yimin Jiang, Yu Zhou, Yuxiang Yang, Binxing Jiao, Daxin Jiang, Heung-Yeung Shum, Jiansheng Chen, Jing Li, Xiangyu Zhang, and Yibo Zhu. Step-audio 2 technical report, 2025.
 - [13] Chi-Yuan Hsiao, Ke-Han Lu, Kai-Wei Chang, Chih-Kai Yang, Wei-Chih Chen, and Hung-yi Lee. Analyzing mitigation strategies for catastrophic forgetting in end-to-end training of spoken language models. *arXiv preprint arXiv:2505.17496*, 2025.

- [14] 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, 2025.
- [15] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens, 2024.
- [16] 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.
- [17] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE signal processing letters*, 6(1):1–3, 1999.
- [18] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.
- [19] Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. <https://github.com/snakers4/silero-vad>, 2024.
- [20] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.
- [21] Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition, 2023.
- [22] Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukas Burget. Leveraging self-supervised learning for speaker diarization, 2024.
- [23] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain, 2021.
- [24] Qiyang Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. CapsFusion: Rethinking Image-Text Data at Scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14022–14032. IEEE, 2024.
- [25] MiniCPM Team, Chaojun Xiao, Yuxuan Li, Xu Han, Yuzhuo Bai, Jie Cai, Haotian Chen, Wentong Chen, Xin Cong, Ganqu Cui, et al. Minicpm4: Ultra-efficient llms on end devices. *arXiv preprint arXiv:2506.07900*, 2025.
- [26] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025.
- [27] Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Livecc: Learning video llm with streaming speech transcription at scale, 2025.
- [28] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv preprint*, abs/2402.03300, 2024.
- [29] Shudong Liu, Hongwei Liu, Junnan Liu, Linchen Xiao, Songyang Gao, Chengqi Lyu, Yuzhe Gu, Wenwei Zhang, Derek F Wong, Songyang Zhang, and Kai Chen. Compassverifier: A unified and robust verifier for llms evaluation and outcome reward. *arXiv preprint arXiv:2508.03686*, 2025.
- [30] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with llms. *ArXiv preprint*, abs/2501.12599, 2025.

- [31] Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIIF-V: Open-Source AI Feedback Leads to Super GPT-4V Trustworthiness, 2024.
- [32] OpenCompass Contributors. OpenCompass: A Universal Evaluation Platform for Foundation Models. <https://github.com/open-compass/opencompass>, 2023.
- [33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.
- [34] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- [35] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.
- [36] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE, 2024.
- [37] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proc. of ICLR*. OpenReview.net, 2024.
- [38] Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram is Worth a Dozen Images. In *European Conference on Computer Vision (ECCV)*, 2016.
- [39] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- [40] Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. *ArXiv preprint*, abs/2504.07957, 2025.
- [41] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 2024.
- [42] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. TextVQA: Towards VQA requiring reasoning about text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [43] Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

- [44] Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations, 2024.
- [45] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning. *ArXiv preprint*, abs/2405.01483, 2024.
- [46] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024.
- [47] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025.
- [48] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusion-bench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14375–14385. IEEE, 2024.
- [49] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *ArXiv preprint*, abs/2309.14525, 2023.
- [50] Chaoyou Fu, Yuhao Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. 2025.
- [51] Weihao Wang, Zehao He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. LVBench: An Extreme Long Video Understanding Benchmark. *ArXiv preprint*, abs/2406.08035, 2024.
- [52] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701, 2025.
- [53] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.
- [54] Wenyi Hong*, Yean Cheng*, Zhuoyi Yang*, Weihao Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. MotionBench: Benchmarking and Improving Fine-grained Video Motion Understanding for Vision Language Models, 2024.
- [55] Hui Bu, Jiatong Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5. IEEE, 2017.
- [56] Jiatong Du, Xingyu Na, Xuechen Liu, and Hui Bu. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*, 2018.
- [57] Binbin Zhang, Hang Lv, Haowen Guo, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP*, pages 6182–6186. IEEE, 2022.
- [58] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP*, pages 5206–5210. IEEE, 2015.

- [59] Guoguo Chen, Wei Chai, Jiatong Wang, et al. Gigaspeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Interspeech*, pages 3670–3674, 2021.
- [60] 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 *ACL-IJCNLP*, pages 993–1003, 2021.
- [61] Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. CoVoST 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*, 2020.
- [62] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL*, pages 527–536, 2019.
- [63] Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024.
- [64] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, pages 1601–1611, 2017.
- [65] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, pages 1533–1544, 2013.
- [66] Haoran Li et al. CMMLU: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*, 2023.
- [67] 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.
- [68] Chengyao Wang, Zhisheng Zhong, Bohao Peng, Senqiao Yang, Yuqi Liu, Haokun Gui, Bin Xia, Jingyao Li, Bei Yu, and Jiaya Jia. MGM-Omni: Scaling omni LLMs to personalized long-horizon speech. *arXiv preprint arXiv:2509.25131*, 2025.
- [69] Tu Anh Nguyen, Wei-Ning Hsu, Antony D’Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. Espresso: A benchmark and analysis of discrete expressive speech resynthesis. In *Interspeech*, pages 4823–4827, 2023.
- [70] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional speech dataset (ESD): A multi-style emotional speech dataset for speech synthesis and voice conversion. In *Interspeech*, pages 3361–3365, 2021.
- [71] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.
- [72] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ICLR*, 2021.
- [73] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, 2024.

- [74] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, 2023.
- [75] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- [76] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [77] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [78] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- [79] Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.
- [80] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.
- [81] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.
- [82] Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. *arXiv preprint arXiv:2512.12772*, 2025.
- [83] Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. *arXiv preprint arXiv:2503.19951*, 2025.
- [84] Qian Chen, Jinlan Fu, Changsong Li, See-Kiong Ng, and Xipeng Qiu. Futureomni: Evaluating future forecasting from omni-modal context for multimodal llms. *arXiv preprint arXiv:2601.13836*, 2026.
- [85] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internv13. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.
- [86] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, and Inderjit Dhillon et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.
- [87] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.
- [88] ggml-org. llama.cpp: Llm inference in c/c++. <https://github.com/ggml-org/llama.cpp>, 2023. Accessed: 2026-04-28.

9 Appendix / supplemental material

Optionally include supplemental material (complete proofs, additional experiments and plots) in appendix. All such materials **SHOULD be included in the main submission**.

A Model Configuration

Table 12 lists the architectural hyperparameters of each component. The full model contains 9.34B learnable parameters and uses bfloat16 precision.

Table 12: Architectural hyperparameters of MiniCPM-o 4.5.

Component	Hyperparameter	Value
Visual Encoder (SigLIP ViT, 417.8M)		
	Hidden dimension	1,152
	Layers	27
	Attention heads	16
	FFN dimension	4,304
	Activation	GELU _{tanh}
	Patch size	14×14
Visual Resampler (88.9M)		
	Query tokens	64
	Embedding dimension	4,096
	Attention heads	32
Audio Encoder (Whisper Medium encoder, 307.2M)		
	Hidden dimension	1,024
	Layers	24
	Attention heads	16
	FFN dimension	4,096
	Activation	GELU
	Mel-frequency bins	80
Audio Projector (21.0M)		
	Architecture	Two-layer MLP with ReLU
	Dimensions	$1024 \rightarrow 4096 \rightarrow 4096$
LLM Backbone (Qwen3-8B, 8,189.2M)		
	Hidden dimension	4,096
	Layers	36
	Attention heads	32
	KV heads (GQA)	8
	Head dimension	128
	FFN dimension	12,288
	Activation	SiLU
	Normalization	RMSNorm ($\epsilon=10^{-6}$)
	Vocabulary size	151,748
	Max context length	40,960
	RoPE θ	10^6
	Weight tying	None
Backbone-to-Decoder Projector (10.5M)		
	Architecture	Two-layer MLP with ReLU
	Dimensions	$4096 \rightarrow 768 \rightarrow 768$
Speech Token Decoder		
	Text embedding layer	116.8M
	Text vocabulary size	152,064
	Transformer	188.8M
	Hidden dimension	768
	Layers	20
	Attention heads	12
	KV heads	12
	FFN dimension	3,072
	Activation	SiLU
	Max context length	4,096
	Speech codebook size	6,562
	Speech number of codebooks	1
	Speech token frame rate	25/s