AV-Dialog: Spoken Dialogue Models with Audio-Visual Input

Anonymous ACL submission

Abstract

Dialogue models falter in noisy, multi-speaker environments, often producing irrelevant responses and awkward turn-taking. We present AV-Dialog, the first multimodal dialog framework that uses both audio and visual cues to track the target speaker, predict turn-taking, and generate coherent responses. By combining acoustic tokenization with multi-task, multi-stage training on monadic, synthetic, and real audio-visual dialogue datasets, AV-Dialog achieves robust streaming transcription, semantically grounded turn-boundary detection and accurate responses, resulting in a natural conversational flow. Experiments show that AV-Dialog outperforms audio-only models under interference, reducing transcription errors, improving turn-taking prediction, and enhancing human-rated dialogue quality. These results highlight the power of seeing as well as hearing for speaker-aware interaction, paving the way for spoken dialogue agents that perform robustly in real-world, noisy environments.

1 Introduction

Dialogue models are moving closer to natural, human-like interaction (Défossez et al., 2024; Veluri et al., 2024), but real-world deployment remains challenging. Real environments are complex with background noise, overlapping talk, and interfering speakers. This setting is known as the "cocktail party problem": the difficulty of attending to a target speaker amid simultaneous talkers and noise. Current models rely solely on speech inputs, making them brittle in these settings; often losing track of the target speaker, producing irrelevant responses, and breaking natural turn-taking.

We argue that overcoming this limitation requires looking as well as listening. Humans address the "cocktail party problem" by combining auditory and visual cues, using lip movements and gaze to focus on the speaker and learn turn-taking cues (Mcdermott, 2009; Best et al., 2023).

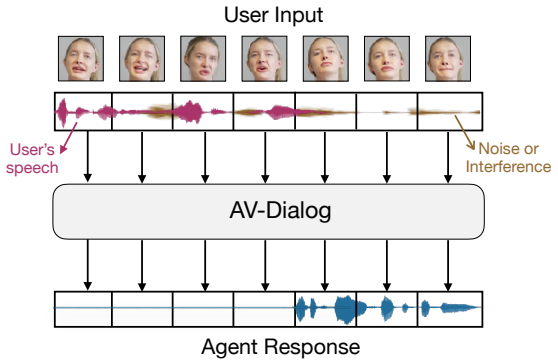

Figure 1: AV-Dialog understands audio-visual input from the target user (purple waveform), accurately detects the appropriate time to take a turn in the conversation, and outputs responses (blue waveform), even in the presence of interfering speakers (brown waveform).

Inspired by this, we present AV-Dialog, a novel audio-visual framework for dialogue modeling. Designing such a framework requires meeting three key challenges: First, the model must continuously process audio and video in a streaming manner, isolating the target speaker even when background noise or louder interfering speakers are present. Second, it must detect turn-taking cues and respond appropriately, maintaining conversational flow despite overlapping or interfering speech. Third, the system must produce coherent responses to the intended speaker without being misled by distractors or environmental noise. large-scale text-only

Our paper presents the first spoken dialog models with audio-visual input that address the above challenges. We make the following contributions:

• **Multimodal dialogue modeling.** We start with a pre-trained large language model (LLAMA3-8B) (Dubey et al., 2024) and train it to process audio and video input in a streaming manner with 40ms chunks. The model learns to extract text tokens for the target speaker under interference and predict turn-change tokens for natural conversational timing. We explore two architectures: *dual* and *unified*. In the dual architecture, the multimodal model outputs transcriptions and turn-taking to-

kens that trigger a second LLM-based text backbone for high-quality response generation, either via in-context learning or instruction-tuning on dialogue data for greater naturalness. The unified architecture instead uses a single model to perform AV understanding, turn prediction, and response generation jointly. Our results show that explicit turn-change supervision is not only essential for dual-model setups but also improves the generation quality of the unified model.

• **Acoustic tokenization for noisy, multi-speaker settings.** Unlike prior dialogue models (Défossez et al., 2024; Veluri et al., 2024) that rely on semantic tokenizers (e.g., HuBERT) trained on single-speaker speech, we use general-purpose acoustic tokens, Descript Audio Codec (DAC) (Kumar et al., 2023) for multimodal dialogue modeling. Because acoustic tokens preserve both semantic and raw acoustic information, they enable inherent speaker differentiation based on voice characteristics. Thus, we can better address the "cocktail party problem," maintaining robustness to noise and interfering speakers across a range of Signal-to-Noise Ratios (SNRs), indicating the noise level of the input speech. In ablation studies, replacing semantic with acoustic tokens both reduces word error rate for streaming AVSR from 67% to 31.7% under strong multi-speaker interference as well as enables more timely responses.

• **Multi-task, multi-stage training recipe.** Open audio-visual dialogue datasets are much smaller than text-based chat corpora, making robust training challenging. We address this with a multi-task, two-stage training strategy: the first stage trains the base LLaMA model with text prediction, ASR, AVSR and audio captioning tasks to strengthen audio-visual understanding and align with original text embeddings. The second stage fine-tunes the model on real audio-only and audio-visual conversational datasets to learn natural turn-taking and conversation context. We further improve robustness with synthetic mixture augmentation, simulating noisy, multi-speaker environments. This task-oriented approach enables AV-Dialog to acquire complementary skills from each dataset, enhancing transcription accuracy, turn-taking prediction, and dialogue quality under challenging conditions.

We compare AV-Dialog with Moshi-7B (Défossez et al., 2024), a state-of-the-art spoken dialogue model. Results show that adding the visual modality boosts turn-taking prediction accuracy from 54% to 79% in the presence of interfering speakers. Human evaluation (N=18) further demonstrates a +1.75-point MOS improvement in dialogue naturalness and a +1.99-point MOS gain in response relevance and helpfulness.

2 Related work

Audio-visual speech recognition. A related task is Audio-Visual Speech Recognition (AVSR) (Rouditchenko et al., 2024; Hong et al., 2023), where models like AV-HuBERT (Shi et al., 2022) learn speech representations from synchronized audio and video. Recent work combines pre-trained audio (Radford et al., 2023) and video (Shi et al., 2022) with language models to improve word error rates (Cappellazzo et al., 2025a). While AVSR systems excel at speech recognition, they are not designed for generative dialogue, turn-taking, or full-duplex interaction. In contrast, AV-Dialog extends audio-visual fusion beyond recognition to enable grounded conversational agents. Moreover, most AVSR models operate offline with full-recording access (Rouditchenko et al., 2024; Cappellazzo et al., 2025a), whereas our system performs streaming inference and incorporates AVSR as a multi-task objective. We therefore compare AV-Dialog with state-of-the-art streaming AVSR models such as Auto-AVSR (Ma et al., 2023) in §4.1.

Dialog models. Recent work on dialog models generate spoken responses from a given prompt. Notably, SpeechGPT (Zhang et al., 2023) is fine-tuned on speech-only data and multimodal instructions for spoken question answering. Multimodal models like SpiritLM (Nguyen et al., 2024) accept speech or text as prompts and generate responses in either modality, while prior non-open source models (Park et al., 2024) handle audio-visual inputs but require explicit prompting. Unlike AV-Dialog, these systems do not model turn-taking and thus do not know when to respond, a key component of human-like dialog interaction. (Liao et al., 2025) leverage audio and visual cues for turn prediction, but their approach is non-streaming, does not support full-duplex interaction, and requires clean text transcripts as input.

Recent full-duplex dialogue models like dGSLM (Lakhotia et al., 2021), Moshi (Défossez et al., 2024), and SyncLLM (Veluri et al., 2024) generate responses concurrently with user input by predicting intent or turn-endings without explicit prompts. However, relying solely on text and

semantic speech tokens limits their ability to track the target speaker in noisy, multi-speaker settings. In contrast, AV-Dialog integrates visual cues and general-purpose acoustic tokens for robust speaker tracking and dialogue generation under challenging signal-to-noise (SNR) conditions.

3 AV-Dialog Models

In human conversation, we process rich acoustic and visual cues to understand and respond via speech. Prevalent dialogue models, however, emphasize the modality the agent must generate (speech & language) while only considering the semantic representations of speech and/or ignore visual cues; making them brittle to noise and interference. In contrast, combining audio and visual context enables accurate turn-boundary detection and timely responses. To this end, we develop a dialogue framework built on audio-visual understanding that infers the user's complete intent, both what is said and when they intend to yield the floor.

As shown in Fig. 2B, AV-Dialog's dual-model architecture comprises two components: an audio-visual dialogue understanding module and a text backbone. The latter is implemented with instruction-tuned LLAMA3-8B (Dubey et al., 2024), though any text LLM or API can be used.

3.1 Audio-Visual Dialogue Understanding

We propose an AV dialogue understanding model, fine-tuned on a base text-LLM, with two essential capabilities for voice interfaces: recognizing user speech and detecting intent to yield the conversation floor. It must also operate as a streaming model, so we can update the response LLM's KV-cache as the user speaks.

Our model processes multi-stream inputs: at each timestep n, a continuous visual stream V_n from a visual encoder and 16 audio streams $\mathbf{A}_n = [A_{n,1}, \ldots, A_{n,16}]$ from an audio tokenizer. It outputs two streams: (1) a text stream U_n representing the user's input, and (2) a turn event stream T_n from the AV understanding module, predicting when the agent should take the conversation floor.

Both streams are synchronized and operate on 40 ms chunks. Both embeddings are projected to the transformer's model dimension via separate linear layers and summed with the previous timestep's text embedding to produce the final embedding $e_n = \mathcal{L}_A(\sum_{i=1}^{16} \mathcal{E}(A_{n,i})) + \mathcal{L}_V(V_n) + \mathcal{E}(U_{n-1}) + \mathcal{E}(T_{n-1})$. Here, $\mathcal{E}(\cdot)$ denotes the embedding layer, $\mathcal{L}_V(\cdot)$ the visual projection layer, and $\mathcal{L}_A(\cdot)$ the audio projection layer.

At the AV-Dialog model output z_n, two linear heads estimate the distributions of U_n and T_n, conditioned on all preceding sub-sequences.

$$\boldsymbol{\sigma}(L_U(z_n)) \approx \mathbb{P}[U_n|\mathbf{A}_{\leq n}, V_{\leq n}, U_{<n}, T_{<n}],$$
$$\boldsymbol{\sigma}(L_T(z_n)) \approx \mathbb{P}[T_n|\mathbf{A}_{\leq n}, V_{\leq n}, U_{<n}, T_{<n}]$$

$\boldsymbol{\sigma}(\cdot)$ is the softmax operation, $L_U(\cdot)$ the linear header for the text stream, and $L_T(\cdot)$ the linear header for the turn event stream.

3.1.1 Audio-Visual Encoding

Most prior turn-taking and spoken dialogue models rely on speaker-invariant semantic speech representations (Nguyen et al., 2022; Veluri et al., 2024; Défossez et al., 2024). While effective in clean settings, they struggle in real-world, "cocktail-party" environments. Models like HuBERT (Hsu et al., 2021), though robust to uncorrelated background noise, can amplify spurious speech interference due to their speaker invariance.

We instead leverage general audio representations, enabling inherent speaker differentiation based on voice characteristics. Our AV-Dialog model uses the high-fidelity Descript Audio Codec (DAC) (Kumar et al., 2023) tokenizer, where each 40 ms chunk is encoded into 16 DAC codebooks.

We incorporate visual cues of the speaker because they (1) enable robust target speech identification in noisy environments, (2) enhance speech perception and understanding (Shi et al., 2022; Cappellazzo et al., 2025b), and (3) provide crucial signals for estimating turn boundaries. Specifically, we use the dlib library (dlib) to detect face regions in first-person video and extract continuous lip-centric visual representations via a pre-trained AV-HuBERT model (Shi et al., 2022).

3.1.2 Output Streams

The AV dialogue understanding module outputs two token streams: i) time-aligned transcription of user's speech U_n, and ii) turn-taking event labels T_n, using the L_U and L_T heads, respectively.

If a user's word begins at t_{start}, the model predicts its tokens from timestep $\lceil t_{start}/25 \rceil + d$, where d is a small delay providing a reasonable context for recognizing the word. When no word is uttered, it outputs a silence token `<EMP>`.

For turn boundaries, we adopt Pairwise-TurnGPT's (Leishman et al., 2024) turn-taking event taxonomy: (1) `Normal turn`, agent speaks

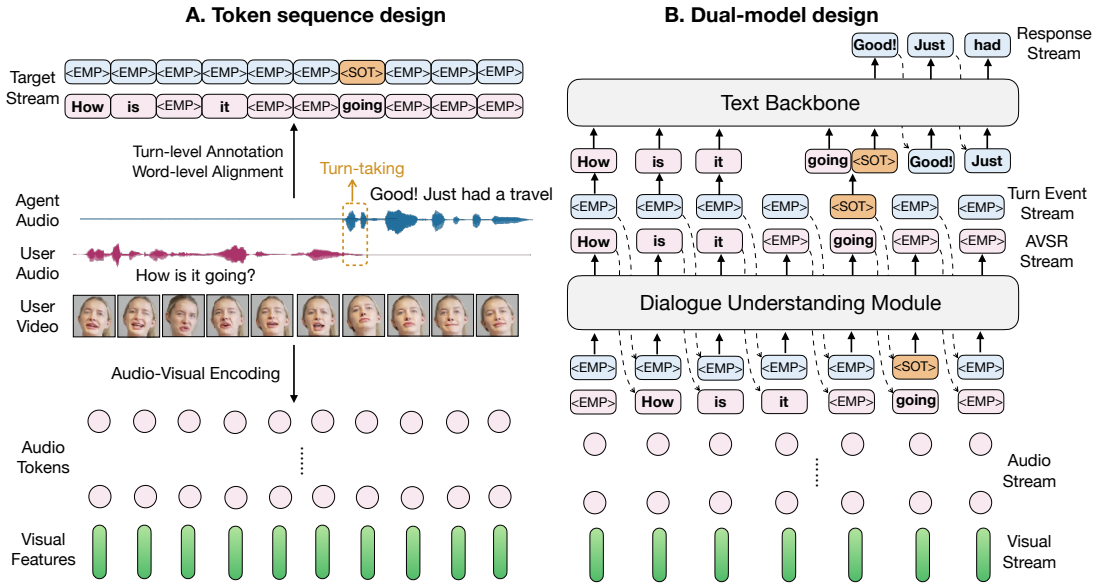

Figure 2: Token sequence and dual-model design. **A.** We use a DAC tokenizer to encode audio into 16 audio token streams and use AV-HuBERT to convert video to continuous visual features. We use turn-level annotation and word-level alignment to generate the target output text stream. **B.** shows our dual-model pipeline for AV-dialog. The AV dialogue understanding module recognizes user speech and detects potential turn-taking events, while the text backbone generates high-quality responses once turn-taking is triggered.

after the user finishes; (2) `Overlapping turn`, agent begins before user finishes, i.e, partial overlap; (3) `Backchannel`, short interjections (e.g., "hmm", "yeah"). The model predicts special token `<SOT>` for both `Normal` and `Overlapping` turns, and `<SOB>` for `Backchannels`. For timesteps without turn taking events, we output `<EMP>`. An example turn-taking event stream is in Fig 2A.

3.2 Response Generation

Our audio-visual dialogue model, in its dual-model mode, streams user speech recognition and turn-taking predictions directly into a text-based LLM backbone (Fig. 2B). This design combines the responsiveness of instruction-tuned LLMs with the flexibility to swap in different text LLMs or APIs.

The text backbone operates in two states: *LISTENING* and *SPEAKING*. In *LISTENING*, non-silence tokens from the AV understanding module are streamed as the user's input. When a turn-taking token appears, the model switches to *SPEAKING*, generating responses autoregressively. If new user speech tokens arrive mid-response, the model yields the floor and re-enters *LISTENING*.

To make response more natural and human-like, we explore two methods:

• **In-Context Learning (ICL):** We apply in-context learning (Brown et al., 2020) and add few-shot dialogue examples from SEAMLESS INTERACTION (InterAct) (Agrawal et al., 2025) training sets to the text backbone's system prompt (see §C.1.1).

• **Instruction Tuning (IT):** We finetune a chat-oriented LLM on real human dialogues using instruction tuning (Ouyang et al., 2022) to improve naturalness and responsiveness. The finetune hyperparameters can be found in (see §C.1.2).

The generated text is then converted to speech via the streaming TTS module Mimi (Défossez et al., 2024).

3.3 Training Strategy

A key challenge in training our AV-dialog understanding module is the scarcity of large-scale aligned audio-visual conversational data. To address this, we leverage diverse data sources: text datasets, monadic audio/audio-visual data, and real dyadic audio-only and audio-visual conversations. We build on a pre-trained text LLM, LLAMA-3-8B, and employ a two-stage, multi-task training approach to progressively develop the audio-visual understanding needed for a robust dialogue model.

3.3.1 Stage 1: Audio-Visual Understanding

The first stage focuses on aligning text, audio, and visual modalities through four multi-task objectives (see §B.1 for training hyper-parameters):

• *Text continuation:* Utilize large-scale text-only datasets for text continuation pre-training objective, to preserve robust language understanding and avoid catastrophic forgetting of textual data.

• *Speech comprehension:* Train on monaural speech datasets, LibriLight (Kahn et al., 2020), MLS (Pratap et al., 2020), and VP400k (Wang et al., 2021), on the ASR task to provide the model with acoustic comprehension of human speech.

• *Audio captioning:* Use the large audio dataset, Audioset (Gemmeke et al., 2017), for audio captioning task to achieve general audio comprehension.

• *Audio-visual alignment:* Train AVSR task using VoxCeleb2 (Nagrani et al., 2017), which is a large audio-visual monadic dataset. This enables the model to learn visual features linked to speech, fostering multimodal learning across text, audio, and visual modalities. To further improve AV understanding in noisy conditions, we apply the synthetic mixing augmentation from §3.3.3 on this dataset.

3.3.2 Stage 2: Learning about Conversations

We train the model on audio-only and audio-visual conversational data to learn natural dialogue dynamics. We use Fisher (Cieri et al., 2004) for audio-only and InterAct (Agrawal et al., 2025) for audio-visual conversations, optimizing two tasks: (1) streaming AVSR and (2) turn-taking event prediction. Synthetic mixing augmentation (§3.3.3) is also applied for robustness in noisy settings. Training hyperparameters are detailed in §B.2.

To prepare target sequences, we align words and turns by converting conversations into synchronized token streams. We deploy Whisper-Large (Radford et al., 2022) to acquire word-level timestamps. The first token of each word is placed at the $\lceil t_{start}/25 \rceil + d$ token in the AVSR stream, where t_{start} is the start timestamp from Whisper. The special turn event token is placed at $\lfloor t_{turn}/25 \rfloor$, where t_{turn} is the timestamp of the annotated turn event. Note that, the $d = 1s$ is also added to the AVSR stream but not to the Turn event stream to avoid introducing additional delays to the response.

3.3.3 Synthetic Mixing Augmentation

We apply synthetic mixing to simulate noisy, multi-speaker environments. For each training sample, with 20% probability, we use clean audio as input, with 40% probability, we mix the clean audio with background noise randomly sampled from from MUSAN (Snyder et al., 2015), and with 40% probability, we mix with 1–4 interference speakers from the same dataset. The input SNR is uniformly sampled between –8 dB and 8 dB. This augmentation enables the model to understand audio-visual cues in complex, real-world conditions.

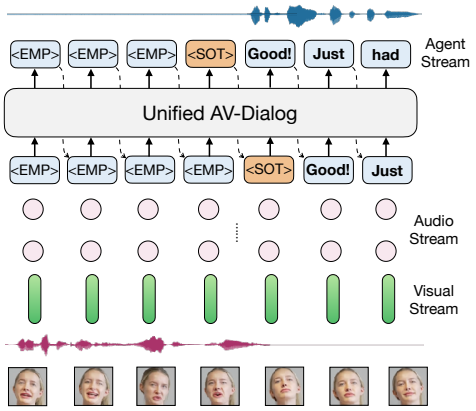

Figure 3: Unified AV-Dialog model. It takes the audio-visual input and predicts the turn-taking events. When the special turn-taking token is generated, the AV-Dialog model generates the response on the same output stream.

3.4 Unified AV-Dialog Model

We also explore a unified model variant where the AV-Dialog module directly generates full-duplex responses, eliminating the need for a text backbone (see §A for algorithmic latency analysis). This requires two key modifications:

• *Model and token design:* We remove the AVSR stream and instead add time-aligned agent response tokens interleaved with the turn-taking event stream (Fig. 3). Empirically, we found removal of AVSR stream improves unified model performance, enabling it to predict turn-taking events and then generate responses directly.

• *Training setup:* In Stage 2, we train on the Fisher and InterAct datasets to generate aligned text responses alongside turn-taking predictions (Fig. 3). Training hyperparameters are provided in §B.3.

4 Experiments

For each AV dialog sample, one side is randomly chosen as the user, whose audio and visual tokens are streamed into the model while output tokens are generated simultaneously. We focus on three aspects: (1) how well the model understands audio-visual input, (2) how accurately it predicts turn-taking, and (3) the quality of its responses.

To test robustness in noisy conditions, we evaluate under three conditions:

• *Clean:* Clean raw audio as input.

• *BG:* Clean raw audio mixed with background noise (music, chatter) from MUSAN (Snyder et al., 2015), input SNR range is -8dB to 12dB.

• *Interf:* Clean raw audio mixed with 1-4 interfering speakers from the same dataset as the target

WER(%) on Voxceleb2 ↓	Clean	BG	Interf
Auto-AVSR	26.8	48.2	71.8
Ours	**17.4**	**35.6**	**38.8**
WER(%) on LRS2 ↓	Clean	BG	Interf
Auto-AVSR	15.8	34.0	60.0
Ours	**9.53**	**24.0**	**28.4**

Table 1: Benchmarking streaming AVSR on the test set of Voxceleb2 and LRS2. We compare WER (%) between AUTO-AVSR and our AV-Dialog model.

WER(%)↓	Clean	BG	Interf
Auto-AVSR	32.2	60.4	93.0
Ours (A)	28.6	68.0	92.2
Ours (V)	67.8	67.8	67.8
Ours (A+V)	16.3	37.4	30.8

Table 2: Streaming AVSR on the InterAct test set. Ours (A): trained and test on audio-only input. Ours (V): trained and tested on visual-only input. Ours (A+V): trained and tested on audio-visual input.

speaker, input SNR range is -8dB to 12dB.

4.1 Audio-Visual Understanding Evaluation

We evaluate streaming AVSR using word error rate (WER). We compare our model with the state-of-the-art streaming AVSR model, Auto-AVSR (Ma et al., 2023). More recent AVSR works (Roudditchenko et al., 2024; Cappellazzo et al., 2025a) focus on offline settings, where models process the full recording before inference, which is an easier task. So, we benchmark against Auto-AVSR on Voxceleb2 and LRS2 dataset.

For a fair comparison, we first benchmark on the Voxceleb2 test set, as both our model and Auto-AVSR are trained on its training set. Since Voxceleb2 lacks text labels, we use Whisper-Large to transcribe clean speech as ground truth for WER. Table 1 compares audio-only, video-only, and audio-visual models. AV-Dialog achieves consistently lower WER than Auto-AVSR, with audio-visual input yielding the best performance. In Table 1, we also compare our model with Auto-AVSR on the LRS2 test set with manually labeled transcripts. Auto-AVSR is trained on LRS2, while our model is not, demonstrating out-of-domain generalization of our model's audio-visual understanding.

We also evaluate our models for the streaming AVSR task on the test-set of the InterAct dataset. We use the transcription from InterAct as our ground-truth text to compute WER. As shown in Table. 2, our AV-dialogue model achieves much lower WER than audio-only or visual-only input,

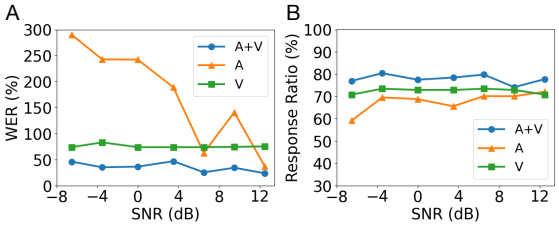

Figure 4: Model performance across different SNRs. Plot A shows the WER of streaming AVSR task on different SNRs of noisy audio input. Plot B shows the response ratio of the turn-taking prediction on different SNRs of noisy audio input.

demonstrating that combining modalities greatly improves AV understanding, especially under noise and interference. To further assess robustness, we evaluate across SNR ranges (Fig. 4A), averaging WER over multiple noise samples for BG and Interf scenarios. The results show our AV model remains robust across varying SNR levels.

4.2 Turn-Taking Prediction Evaluation

We evaluate our model using turn-taking events from (Nguyen et al., 2023) and measure floor-transfer offset (FTO), which is the duration between turn transitions, where negative FTO indicates overlap and positive FTO indicates a gap.

We extract the agent's turn start timestamp using the SOT token and compute FTO as the gap between the user's turn end (obtained from ground-truth turn annotations) and the agent's turn start.

Metrics. We report three metrics:

- *Response Ratio:* percentage of FTOs within –2s to 3s, the typical range in InterAct's human conversations (around 90% of FTOs in the InterAct test set are in this range).
- *FTO Error:* mean absolute error (MAE) between generated and ground-truth FTOs.
- *Median FTO:* the median value of FTOs.

Baselines. We compare with Moshi (Défossez et al., 2024), a state-of-art spoken dialogue model.

- **Moshi:** We deploy the Moshi checkpoints (Défossez et al., 2024) to generate responses.
- **SE+Moshi:** We first apply the speech enhancement (SE) model Demucs (Defossez et al., 2020), then feed output to Mosh to generate responses.
- **Separation+Moshi:** We first apply X-TF-GridNet (Hao et al., 2024), a state-of-the-art target speaker separation model, to the noisy mixture. We then feed output to Moshi to generate responses.

Results. As shown in Table 3, the audio-visual

Model	Clean			BG			Interf		
Metrics	Response Ratio(↑)	FTO Err(↓)	Median FTO	Response Ratio(↑)	FTO Err(↓)	Median FTO	Response Ratio (↑)	FTO Err(↓)	Median FTO
Moshi	54.0%	3.48	-0.72	53.8%	3.20	-0.12	52.5%	3.27	-0.52
SE + Moshi	55.9%	3.66	-0.6	56.0%	3.24	-0.7	54.4%	3.36	-0.84
Separation + Moshi	47.8%	3.60	-0.92	51.7%	3.47	-0.42	52.7%	3.27	-0.92
Ours (A)	73.2%	1.87	1.04	70.2%	2.66	1.02	65.8%	2.81	1.2
Ours (V)	72.5%	2.37	0.98	72.5%	2.37	0.98	72.5%	2.37	0.98
Ours (A+V)	**74.5%**	1.86	1.16	**78.3%**	1.96	1.12	**78.8%**	1.81	1.18
Ours (Unified)	68.1%	**1.68**	1.92	75.6%	**1.49**	1.76	75.9%	**1.67**	1.76
GT	-	0	1.5	-	0	1.5	-	0	1.5

Table 3: Turn-taking evaluation under different noise and interference conditions.

Noise condition	Clean		BG		Interf	
Model	PPL(↓)	Pickup Ratio(↑)	PPL(↓)	Pickup Ratio(↑)	PPL(↓)	Pickup Ratio(↑)
Moshi	44.1	23.8%	52.4	19.4%	46.4	18.4%
SE + Moshi	50.8	26.4%	50.4	24.5%	56.1	19.1%
Ours (ICL)	25.8	**66.6%**	24.6	**68.1%**	**23.1**	**67.8%**
Ours (IT)	**23.2**	32.5%	**24.0**	30.3%	23.8	36.7%
Ours (Unified)	31.0	29.6%	29.8	35.5%	32.5	31.3%
GT	51.7	-	51.7	-	51.7	-

Table 4: Semantic evaluation of dialogue responses under different noise conditions.

dialogue model achieves the highest response ratio across all noisy scenarios: 74.5% (*Clean*), 78.3% (*BG*), and 78.8% (*Interf*), substantially outperforming the baseline Moshi model, which reaches only 50%. Adding visual input to the audio-only model improves turn-taking accuracy by 1.3% (Clean), 8.1% (BG), and 13% (Interf). The unified audio-visual model shows a slight drop in response ratio but achieves the lowest FTO errors. Note that the GT median FTO is around 1.5s for the InterAct conversation dataset which consists of casual conversations between strangers. The detailed FTO distribution visualization can be found in §E.

To assess robustness under noise, in Fig. 4B, we also evaluate turn-taking prediction across different SNR ranges for both *BG* and *Interf* scenarios.

4.3 Semantic Evaluation

We evaluate the semantic quality of dialogue responses, comparing the Moshi baselines with three audio-visual dialogue variants:

- **Ours (ICL):** Dual-model pipeline using in-context learning with example conversations from InterAct in the prompt of the LLAMA3-8B text backbone model (see §C.1).
- **Ours (IT):** Dual-model pipeline fine-tuned via instruction tuning on Fisher, and InterAct datasets for the text backbone model (see §C.2).
- **Ours (Unified):** Unified model trained to generate text responses from audio-visual input.

We run our models end-to-end and compute the perplexity (PPL) of agent-generated turns. To further assess text quality, we use the Prometheus (Kim et al., 2023) LLM as an evaluator framework, performing relative/pairwise comparisons rather than absolute scoring, which better aligns with human judgment (Kiritchenko and Mohammad, 2017; Liusie et al., 2023). For each evaluation, the LLM compares the ground-truth InterAct response with the model-generated text. We compute the *Pickup Ratio* as the fraction of responses in which the LLM prefers the model-generated text over the ground truth (see details in §D).

As shown in Table 4, the baseline Moshi and SE+Moshi models achieve the lowest Pickup Ratio according to the LLM evaluator. The audio-visual dialogue model using in-context learning (ICL) achieves the highest Pickup Ratio among all methods. In contrast, the same dual-model pipeline fine-tuned via instruction tuning (IT) and the unified audio-visual model show a reduced Pickup Ratio of 30–40%. This drop is likely because InterAct dialogues often contain casual, unpredictable conversation; fine-tuning the generation task on such data can degrade response quality. For perplexity (PPL), both the ICL and IT audio-visual models achieve the lowest values. Finally, the response quality for the unified model is worse than the cascaded model settings for our AV-dialogue models. This is line with recent observations in the related domain of speech-to-speech dialog models (Hu

Model	N-MOS(↑)	H-MOS(↑)
SE + Moshi	2.39	2.10
Ours (Dual+ICL)	**4.14**	**4.09**
Ours (Unified)	3.54	3.02
GT	3.92	3.62

Table 5: Human evaluation result. N-MOS: Mean Opinon Score on Naturalness of response. H-MOS: Mean Opinion Score on Helpfulness of response.

et al., 2025b), where cascade model responses outperform unified models.

4.4 Human Evaluation

We conducted a human evaluation with 18 participants to assess the end-to-end performance of our audio-visual dialogue model. Model text outputs were converted to speech using the Moshi streaming TTS (Défossez et al., 2024), ensuring a fair comparison since both systems used the same TTS. Participants were given dialogue transcripts and audio, including both user turns and model responses.

We randomly selected 15 samples from the Inter-Act test set across Clean, BG, and Interf conditions (details and SNRs in §G). For each sample, participants evaluated four conditions: (1) SE+Moshi, (2) dual-model + ICL, (3) unified model, and (4) ground truth. Each participant rated 8 dialogue sets, with randomized method order to avoid bias. Ratings followed the Mean Opinion Score (MOS) protocol (ITU-T P.808 (ITU-T, 2018)) on a 5-point Likert scale, evaluating Naturalness (N-MOS) and Helpfulness (H-MOS) (see §F).

Table 5 shows that both our dual and unified models outperform the SE+Moshi baseline in naturalness and helpfulness. The dual model with in-context learning achieves the best results. The unified model drops by 0.5 in naturalness and 1.02 in helpfulness compared to the dual model. This is likely due to (1) the limited size of real conversational data and (2) the fact that the conversations in the dataset are mostly casual chit-chat, often containing low-quality responses, random topic shifts, and limited logical reasoning.

These results highlight that the dual model benefits from using real-world data primarily for turn-taking modeling while leveraging the pretrained text backbone for stronger generation quality. Notably, the MOS trends align with the LLM evaluator pick ratios in our semantic evaluation.

4.5 Ablation studies

We first compare acoustic tokens with semantic tokens. Using the same training setup, we trained the AV model with DinoSR (Liu et al., 2023) semantic tokens instead of DAC tokens, as DinoSR is a newer, improved semantic representation compared to HuBERT. Table 6 shows that acoustic tokens outperform semantic tokens in both streaming AVSR and turn-taking prediction tasks.

AVSR WER(%) ↓	Clean	BG	Interf
DinoSR (A)	24.9	89.0	239.2
DinoSR (A+V)	26.9	83.0	67.0
Acoustic (A)	28.6	60.0	63.4
Acoustic (A+V)	16.3	37.4	30.8
Response Ratio(%) ↑	Clean	BG	Interf
DinoSR (A)	67.3	63.3	63.2
DinoSR (A+V)	69.5	49.1	47.8
Acoustic (A)	73.2	70.2	65.8
Acoustic (A+V)	74.5	76.9	78.8

Table 6: Acoustic & semantic token comparison.

AVSR WER(%) ↓	Clean	BG	Interf
Ours(A+V)	16.3	37.4	30.8
Ours(No Stage 1)	58.9	95.1	86.7
Ours(No Audio Diag)	22.6	37.8	31.8
Ours (No augmentation)	17.5	121.3	160.2

Table 7: Ablation Study on the training recipe.

Next, we compare different training strategies (Table 7). *No Stage 1* indicates skipping Stage 1 while training on the LLaMA3-8B model, while *No Audio Dialogue* excludes the audio-only dialogue dataset during Stage 2 fine-tuning. Including the audio-only dialogue dataset improves performance. *No augmentation* trains the second stage without Synthetic Mixing Augmentation. Results show a significant drop without Stage 1 or without Synthetic Mixing Augmentation.

We also evaluate the impact of explicit turn-taking supervision strategy on the unified model in Appendix. §J. Finally, we conduct ablation studies on the effect of different audio channels of tokenizer and visual input distortion in Appendix. §J.

5 Conclusion

We introduced AV-Dialog, the first streaming audio-visual dialogue system that integrates audio, vision, turn-taking, and response generation. Using acoustic tokenization, multi-stage training, and explicit turn-event supervision, it achieves robust performance in noisy, multi-speaker environments, outperforming audio-only baselines in transcription, turn-taking, and response quality. Human evaluations confirm that AV-Dialog enables more natural, helpful, and speaker-aware conversations, underscoring the value of combining listening and looking for real-world multimodal dialogue.

6 Limitations and Risks

Limitations. While AV-Dialog advances full-duplex dialogue in noisy environments, its performance can be further improved. It currently does not explicitly model non-verbal auditory cues (e.g., laughter, sighs) or visual cues (e.g., facial expressions, gestures) beyond lip movements. Enhancing the understanding and generation of these multimodal signals could make interactions more human-like. Finally, factors like poor lighting, occlusions (e.g., hands covering the mouth) or extreme head poses, can impair lip movement extraction (Shi et al., 2022), affecting speaker tracking and speech understanding. Developing lip encoders that are robust to such conditions is a promising and complementary direction for future work. This performance gap between cascaded and unified approaches is an important observation, and we would like to note that this has been highlighted as a limitation of unified approaches by prior work as well (Hu et al., 2025a; Veluri et al., 2024; Nguyen et al., 2023). Multiple prior works report a similar performance gap in response quality compared to cascaded baselines, with the primary driving factor being the lack of large-scale channel separated spoken dialog data.

Ethical considerations. Like any advanced AI enabling human-like interaction, AV-Dialog presents key ethical challenges. It may produce misleading dialogue, particularly under noisy or ambiguous conditions, requiring rigorous evaluation and ongoing monitoring. While audio-visual capture (e.g., lip movements, voices) is common in voice conferencing platforms like Zoom, it still demands strict attention to privacy. To prevent misuse such as exploitation in online scams, methods like speech watermarking could help safeguard against abuse.

References

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeffrey M. Girard, Martin Gleize, and 65 others. 2025. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *CoRR*, abs/2506.22554.

Virginia Best, Alex D. Boyd, and Kamal Sen. 2023. An effect of gaze direction in cocktail party listening. *Trends in Hearing*, 27:23312165231152356. PMID: 36691678.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. 2025a. Large language models are strong audio-visual speech recognition learners. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. 2025b. Large language models are strong audio-visual speech recognition learners. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *International Conference on Language Resources and Evaluation*.

Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real time speech enhancement in the waveform domain. *Preprint*, arXiv:2006.12847.

dlib. Dlib c++ library. `https://dlib.net/`.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *Preprint*, arXiv:2410.00037.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Fengyuan Hao, Xiaodong Li, and Chengshi Zheng. 2024. X-tf-gridnet: A time–frequency domain target speaker extraction network with adaptive speaker embedding fusion. *Information Fusion*, 112:102550.

Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. 2023. Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18783–18794, Los Alamitos, CA, USA. IEEE Computer Society.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Żelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, and Boris Ginsburg. 2025a. Efficient and direct duplex modeling for speech-to-speech language model. *arXiv preprint arXiv:2505.15670*.

Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Żelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, and Boris Ginsburg. 2025b. Salm-duplex: Efficient and direct duplex modeling for speech-to-speech language model. *Preprint*, arXiv:2505.15670.

ITU-T. 2018. Recommendation P.808: Subjective evaluation of speech quality with a crowdsourcing approach. ITU-T Recommendation P.808, International Telecommunication Union.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. 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)*, pages 7669–7673. https://github.com/facebookresearch/libri-light.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Svetlana Kiritchenko and Saif M Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. *arXiv preprint arXiv:1712.01765*.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, and Emmanuel Dupoux. 2021. Generative spoken language modeling from raw audio. *Preprint*, arXiv:2102.01192.

Sean Leishman, Peter Bell, and Sarenne Wallbridge. 2024. Pairwiseturngpt: a multi-stream turn prediction model for spoken dialogue. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue*.

Zikai Liao, Yi Ouyang, Yi-Lun Lee, Chen-Ping Yu, Yi-Hsuan Tsai, and Zhaozheng Yin. 2025. Beyond words: Multimodal llm knows when to speak. *arXiv preprint arXiv:2505.14654*.

Alexander H Liu, Heng-Jui Chang, Michael Auli, Wei-Ning Hsu, and Jim Glass. 2023. Dinosr: Self-distillation and online clustering for self-supervised speech representation learning. *Advances in Neural Information Processing Systems*, 36:58346–58362.

Adian Liusie, Potsawee Manakul, and Mark JF Gales. 2023. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. *arXiv preprint arXiv:2307.07889*.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-avsr: Audio-visual speech recognition with automatic labels. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

T Aleksandra Ma, Sile Yin, Li-Chia Yang, and Shuo Zhang. 2025. Real-time audio-visual speech enhancement using pre-trained visual representations. *arXiv preprint arXiv:2507.21448*.

Josh Mcdermott. 2009. The cocktail party problem. *Current biology : CB*, 19:R1024–7.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *CoRR*, abs/1706.08612.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2022. Generative spoken dialogue language modeling. *Preprint*, arXiv:2203.16502.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, and 1 others. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. 2024. Spirit-lm: Interleaved spoken and written language model. *Preprint*, arXiv:2402.05755.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Se Park, Chae Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeonghun Yeo, and Yong Ro. 2024. Let's go real talk: Spoken dialogue model for face-to-face conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16348, Bangkok, Thailand. Association for Computational Linguistics.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. 2024. Whisper-flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation. In *Interspeech 2024*, pages 2420–2424.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21390–21402, Miami, Florida, USA. Association for Computational Linguistics.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. 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. Association for Computational Linguistics.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15757–15773. Association for Computational Linguistics.

A Algorithm Latency Analysis

Moshi is built on a 7B backbone model, while our unified system uses a single 8B LLaMA model. Our dual-model architecture extends this by running two such models in parallel, which naturally increases peak memory consumption.

Because system latency depends on hardware and software optimizations, which is not the focus of our paper, we focus instead on algorithmic latency which is a platform-invariant metric. In AV-Dialog, both the audio tokenizer (DAC) and the visual encoder operate causally at 25 Hz. However, the AV-HuBERT visual encoder introduces a 2-frame lookahead (Ma et al., 2025), resulting in an overall algorithmic latency of approximately 120 ms. In our dual-model setup, output tokens from the understanding module are streamed directly to the text backbone with KV-cache in parallel, enabling immediate response generation during turn-taking without additional delay.

By contrast, Moshi (Défossez et al., 2024) processes 80 ms audio chunks, achieving an algorithmic latency of about 80 ms. Note that AV-Dialog employs Moshi's TTS model. Although our system's latency is somewhat higher, it is mainly limited by the visual encoder's lookahead: an aspect that could be further reduced by pretraining a visual encoder with a smaller or zero lookahead window. Despite this limitation compared to Moshi-like natively full-duplex models, we believe our approach of predicting agent's start-of-the-turn bridges the naturalness gap while also leveraging superior helpfulness and knowledge of standalone text backbones.

B Training Hyper-parameters

B.1 Stage-1 Training

In Stage 1, we trained the original LLAMA3-8B with sequence length 4096. We use a learning rate of $3e^{-5}$ on the transformer block and a learning rate of $1.5e^{-4}$ on embedding layers and audio/visual adapters. The model is trained with 500 step warmup and trained for 50k iterations on 128 A100 GPUs with a per-gpu batch size of 1.

The proportion of each task and dataset in the stage1 training is as follows:

• *Text continuation* (48.0%): Arxiv (16.0%), B3g (20.0%) and Wikipedia_en (12.0%).

• *Speech comprehension* (32.0%): LibriLigh (13.76%), MLS (13.12%) and VP400k (5.12%).

• *Audio captioning* (4.0%): AudioSet (4.0%).

• *Audio-visual alignment (AVSR)* (16.0%): Voxceleb2 (16.0%).

The input and output token sequence design for Stage 1 training is shown in Fig. 5. The special tokens <ASR>, <Trans>, and <AC> serve as prefixes for different tasks. We also introduce a special <NULL> token: when a modality is missing in the input stream for a given task, it is filled with <NULL>, whose embedding vector is all zeros after the embedding layer. If <NULL> appears in the target stream, its loss is not computed. We apply cross-entropy loss on the output text stream.

B.2 Stage-2 Training for Dual Model

In Stage 2, we fine-tuned the Stage 1 model with a sequence length of 4096. We used a learning rate of $2e^{-5}$ for the transformer blocks, embedding layers, and audio/visual adapters. The model was trained with a 500-step warm-up over 10k iterations on 32 A100 GPUs, with a per-GPU batch size of 1. The proportion of each task and dataset in Stage 2 fine-tuning is as follows:

• *Audio-only conversation* (55.0%): Fisher (10.0%) and InterAct (45.0%).

• *Audio-Visual conversation* (45.0%): InterAct (45.0%).

The input and output token sequence design for Stage 2 training is shown in Fig. 6. We also apply the <NULL> token in the same way as Stage 1. We compute the cross-entropy loss on both the AVSR stream U and Turn event stream T and compute their average. In the AVSR stream, the loss weight for text tokens is set to 1.0, while the silence token <EMP> is set to 0.1. In the Turn-Event stream, the loss weight for Turn-taking token <SOT> is set to 2.5, the loss weight for the backchannel token <BOT> is set to 1.0, and the loss weight for the the silence token <EMP> is set to 0.1.

B.3 Stage-2 Training for Unified Model

In Stage 2 of our unified model, we fine-tuned the pretrained Stage 1 model with a sequence length of 4096. We used a learning rate of $2e^{-5}$ for the transformer blocks, embedding layers, and audio/visual adapters. The model was trained with a 500-step warm-up over 10k iterations on 32 A100 GPUs, with a per-GPU batch size of 1.

The proportion of each task and dataset in Stage 2 fine-tuning is as follows:

• *Audio-only conversation* (55.0%): Fisher (10.0%), InterAct (45.0%).

• *Audio-Visual conversation* (45.0%): InterAct (45.0%).

The input and output token sequence design for the unified model training at Stage 2 is shown in Fig. 7. We apply the <NULL> token in the same way as Stage 1. We compute the cross-entropy loss on the output stream T. The loss weight for text tokens is set to 1.0, the loss weight for <EMP> is set to 0.1, the loss weight for <SOT> is set to 2.5, and the loss weight for <BOT> is set to 1.0.

C Text Backbone Hyper-parameters

C.1 In-Context Learning

The prompting and few-shot samples are provided as:

"Carefully read the user prompt. You follow these instructions:

You are a helpful assistant that engages in natural, casual conversation. Respond like a human would - be conversational, use natural language, and don't be overly formal.

Here are some examples of the conversational style you should adopt:

Example1:

user: Hey, Siobhan, what's up? You seem troubled.

assistant: Yeah, I am. I'm just having a hard time. I needed someone to talk to.

user: Of course, man, I'm always here for you. What's going on?

assistant: It's just everything. Work is stressing me out. My relationship is falling apart, and I feel like I'm losing touch with my friends. I don't know what to do.

user: Well, let's start with work then. What's going on there?

assistant: It's just that... Everything is so demanding and I can't keep up. I'm constantly behind and it feels like I'm never gonna catch up.

Example2:

user: Hey, thanks for taking my motorcycle off my hands. I really appreciate it.

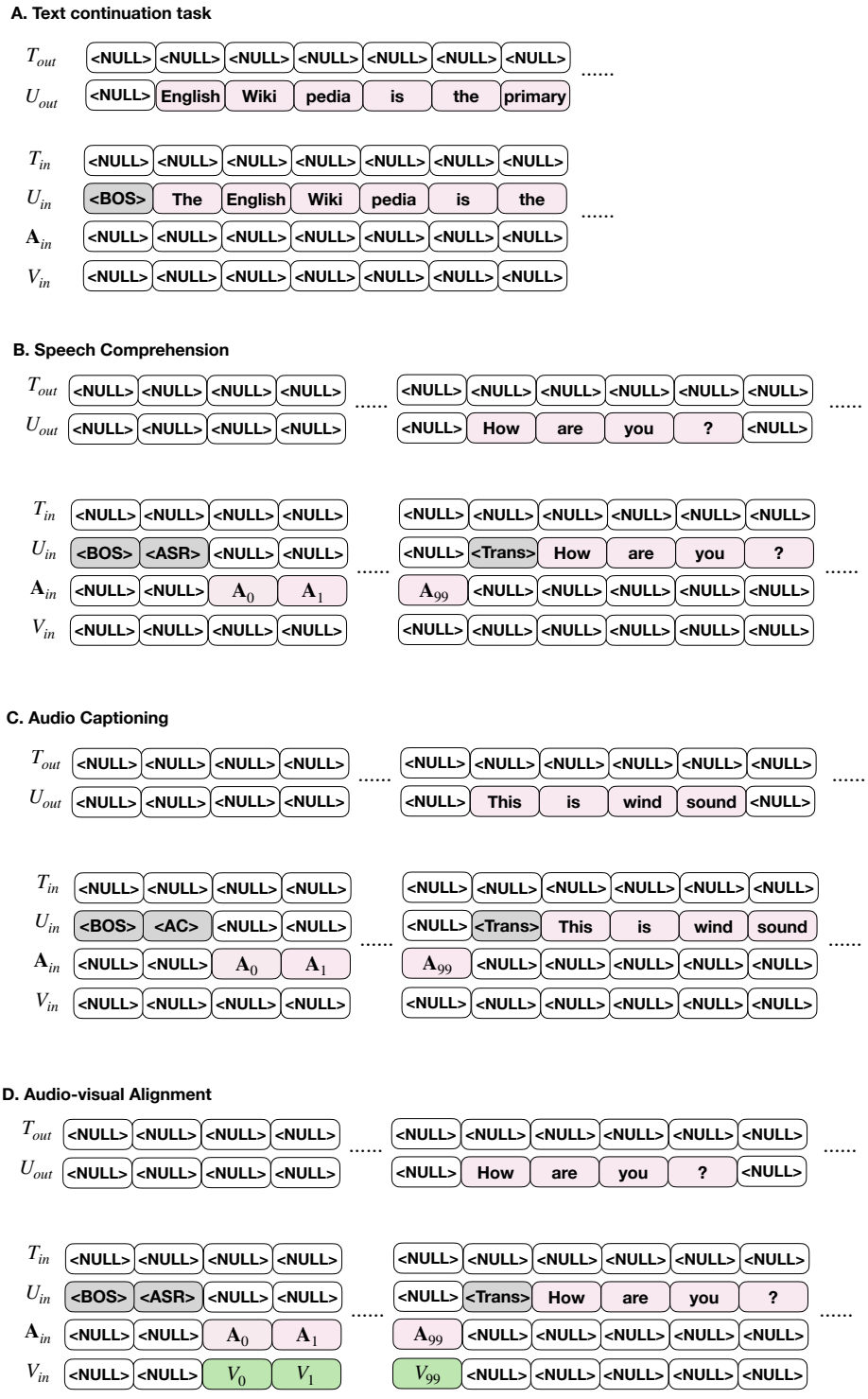

Figure 5: Training input and output tokens design for different tasks at Stage 1.

assistant: No problem, man. It's a nice bike. I'm sure I'll enjoy riding it.

user: I'm just glad to be rid of it, to be honest. It was costing me a lot in repairs and maintenance.

assistant: Yeah, motorcycles can be pretty expensive to keep up. They're also a lot of fun to ride. So it's worth it in my opinion.

user: I'm not so sure. I think I'd rather just... take the bus from now on. It's cheaper and less hassle.

assistant: That's a valid point. But sometimes it's nice to have the freedom that a motorcycle provides. You can go where you want, when you want.

user: I guess that's true, but it's just not worth the expense for me anymore.

Example3: user: Would you rather have the ability to speak to the past or send messages to the

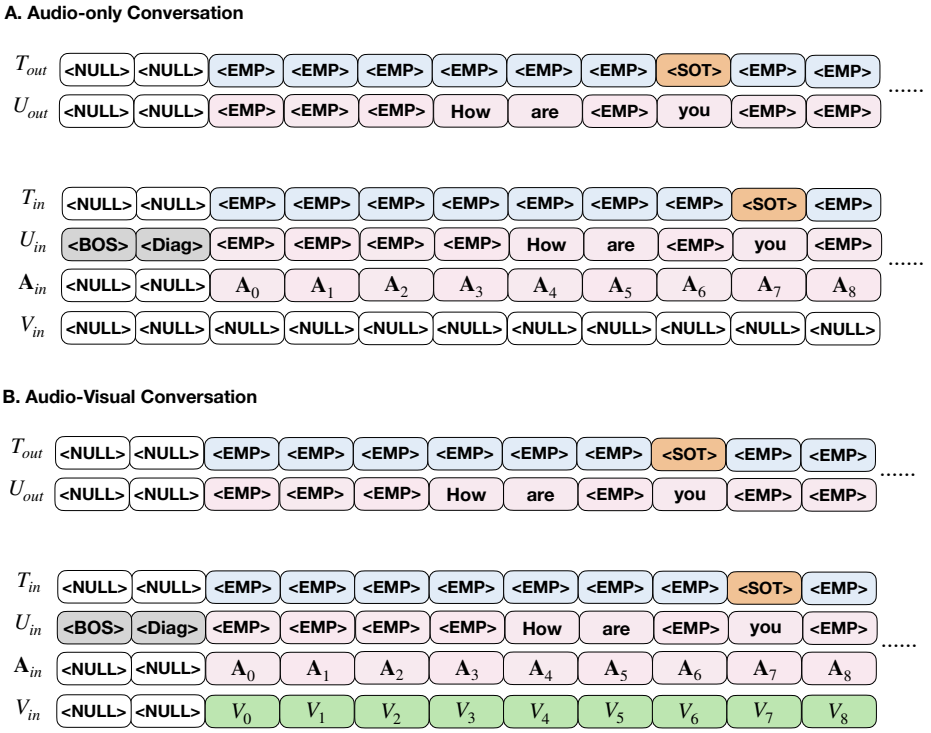

Figure 6: Training input and output tokens design for different tasks at Stage 2.

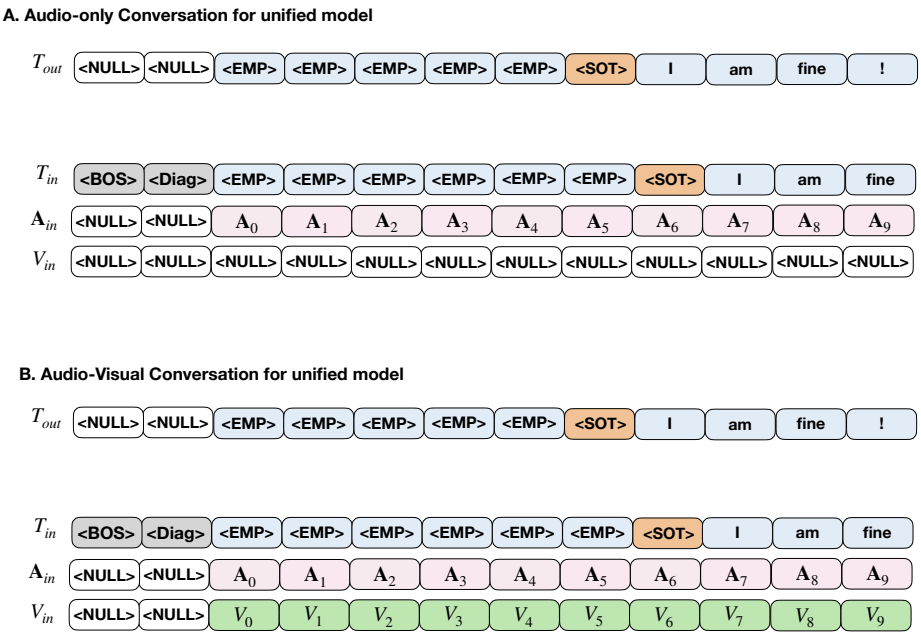

Figure 7: Training input and output tokens design for multi-tasks at Stage 2 for the unified model.

future?

assistant: I would rather send messages to the future.

user: Really? But we can kind of already do that, don't you think, with certain apps and maybe hiding a letter in your house or something?

assistant: Well, to me, the progress, I like the progress, what's going on, and I can't just wait to see what else up there.

user: What kind of messages would you want to send?

assistant: And messages are positive, always positive.

Now respond to the user in this same natural conversational style. Remember your instructions."

C.2 Instruction Tuning

We finetune LLAMA3-8B-INSTRUCT with a sequence length of 8192 on human conversational datasets, including 22.2% InterAct dataset, 3.7% Fisher dataset. We pre-processing the dialogue dataset by applying the original instruction-tuning template from Llama3-8B-Instruct. We use a low learning rate of $1e^{-5}$ and finetuned it for only 3000 steps on 32 A100 GPUs with a per-gpu batch size of 1.

D Evaluation Prompting for Prometheus

We deploy PROMETHEUS-7B-V2.0 as our LLM evaluator on the generated response. We randomly shuffle the order of ground-truth and generated response. The rubric description is as follow:

"Does Agent respond in a way that is generally related to the user's input and current conversation? Minor topic drift, informal language, brevity, or slight ambiguity should not be penalized. Ignore formatting, punctuation, and minor inconsistencies."

E FTO Distribution

Fig. 8 visualizes the different FTO distributions for different model configurations:

- A. Moshi: state of the art speech-only dialogue model
- B. SE+Moshi: cascade of a speech enhancement model (Demcus) and Moshi
- C. Semantic(A): our dual-model approach with semantic tokens (DinoSR) and audio-only input
- D. Semantic(A+V): our dual-model approach with semantic tokens (DinoSR) and audio-visual input
- E. Acoustic(A): our dual-model approach with acoustic tokens (DAC) and audio-only input
- F. Acoustic(V): our dual-model approach with acoustic tokens (DAC) and visual-only input
- G. Acoustic(A+V): our dual-model approach with acoustic tokens (DAC) and audio-visual input
- G. Unified(A+V): our unified-model approach with acoustic tokens (DAC) and audio-visual input

F N-MOS and H-MOS

N-MOS scores for Naturalness are defined as follows:

- 1. Bad - Response is not normal English or does not make sense.
- 2. Poor - Response is normal English but not coherent to the user's input.
- 3. Fair - Response is somewhat plausible and coherent
- 4. Good - Response is plausible and coherent
- 5. Excellent - Response is highly plausible and coherent

M-MOS scores for Meaningfulness are defined as follows:

- 1. Bad - essentially nothing in common with human-like conversation
- 2. Poor - very little natural and human-like conversation
- 3. Fair - substantial differences from human-like and natural conversation
- 4. Good - minor differences from human-like and natural conversation
- 5. Excellent - basically indistinguishable from human-like and natural conversation

G Samples distribution of real-human evaluation

The properties of the samples used in human evaluation are shown in Table. 8.

H User study participants

The human evaluation study was performed under our institution's IRB. All participants provided consent and were recruited from our institutions and nearby areas. Participants ranged from 18 to 40 years old, with 35% identifying as female and the

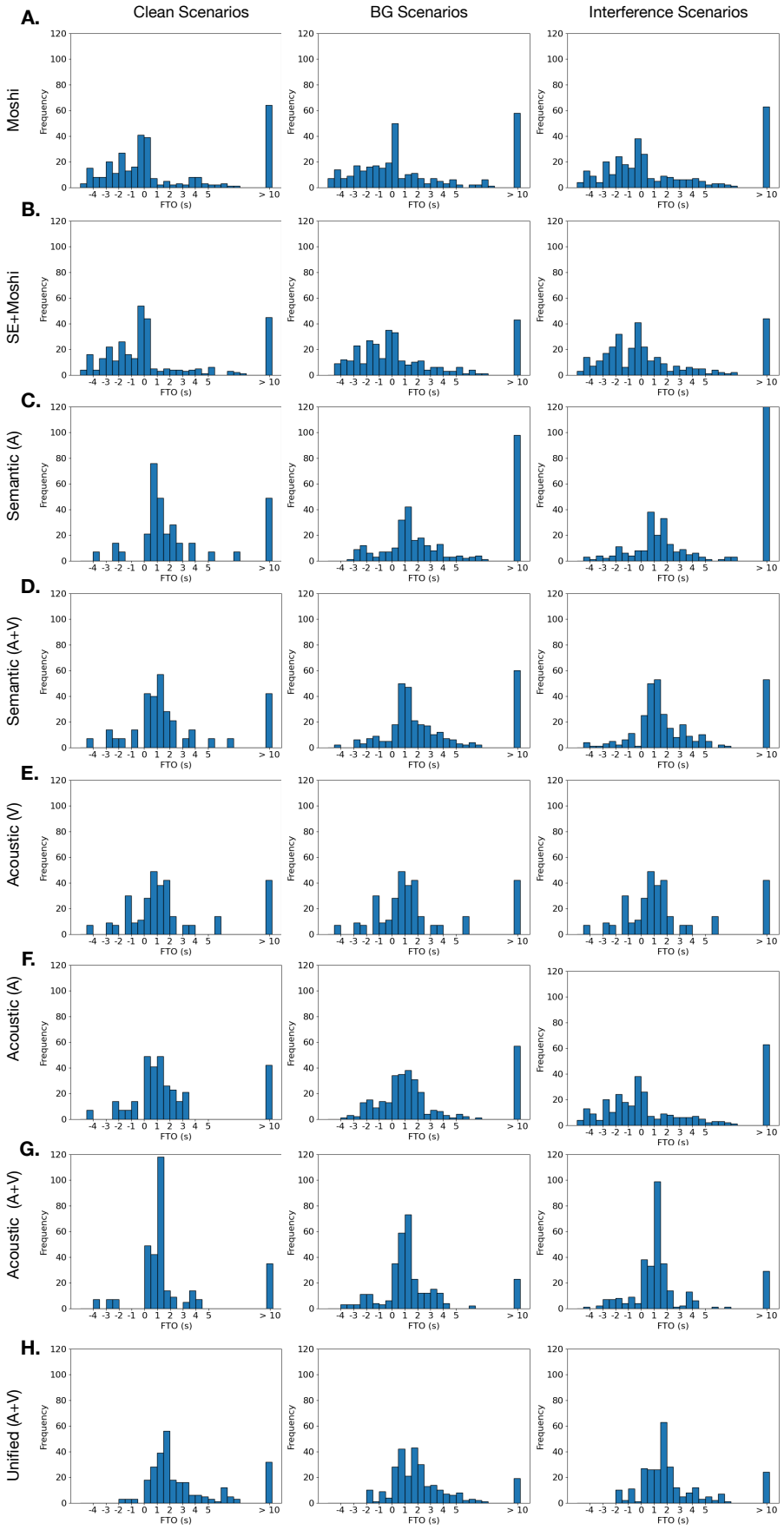

Figure 8: Distribution of FTO of different model configurations. On the x-axis of each figure, the label "> 10" also accounts for samples where the model does not respond at all.

Sample Index	Noise Scenario	SNR(dB)
1	BG	9.0
2	BG	0.0
3	clean	∞
4	Interf	3.0
5	clean	∞
6	Interf	-2.99
7	Interf	2.99
8	clean	∞
9	clean	∞
10	Interf	-3.0
11	BG	3.0
12	BG	6.0
13	Interf	-7.0
14	BG	3.0
15	BG	0.0

Table 8: Distribution of the samples used in human evaluation.

AVSR WER(%) ↓	Clean	BG	Interf
16 channels	16.3	37.4	30.8
8 channels	19.2	36.6	40.1
4 channels	23.8	65.0	48.8
Response Ratio(%) ↑	Clean	BG	Interf
16 channels	74.5	78.3	78.8
8 channels	79.1	78.4	77.1
4 channels	73.8	78.6	72.5

Table 9: Ablation Study on the audio channel number.

rest as male. They are recruited from both technical background and non-technical background.

I Example of AV-Dialog

We provide some samples of our AV-Dialog input and output in Figure. 9

J Ablation Study

J.1 Effect of audio channels

To evaluate the impact of audio compression quality, we experimented with varying the number of audio token channels. Since our Residual Vector Quantization (RVQ) audio tokenizer encodes information hierarchically (coarse-to-fine), we compared the full 16 channels against reduced inputs of 8 and 4 channels. As shown in Table. 9, reducing the input to 8 channels results in only a slight performance drop. Another interesting observation is that the AV understanding task is more vulnerable to the audio channel drop, because AV understanding task relies heavily on the fine-grained acoustic details captured in the deeper RVQ levels, which are lost when channels are aggressively pruned.

AVSR WER(%) ↓	Clean	BG	Interf
0% drop rate	16.3	37.4	30.8
5% drop rate	17.3	34.0	32.6
10% drop rate	18.4	35.8	33.5
25% drop rate	17.9	35.6	35.6

Table 10: Ablation Study on the visual frame drop.

Response ratio(%) ↑	Clean	BG	Interf
w explicit turn-taking	68.1	75.6	75.9
w/o explicit turn-change	48.0	35.1	38.0
LLM-evaluator pick-up ratio(%) ↑			
w explicit turn-change	29.6	35.5	31.3
w/o explicit turn-change	29.0	22.1	18.2

Table 11: Ablation study on turn-taking supervision in our Unified Model.

J.2 Effect of Visual Frame Drop

We conducted additional experiments on video modality distortion (e.g., occlusions, back-facing, or extreme angles). In our pipeline, such distortions cause the face detection module to fail, resulting in dropped visual frames. We simulated this by randomly dropping video frames at different rates. The results in Table. 10 show that our model can be robust to the video frame dropped due to the severe distortion.

J.3 Effect of Explicit Turn-taking Supervision

We also evaluate the impact of explicit turn-taking supervision on the unified model. In Stage 2, we trained a version without the <SOT> token, forcing it to generate response tokens directly. Table 11 shows that this supervision is crucial for both turn-taking prediction and response quality.

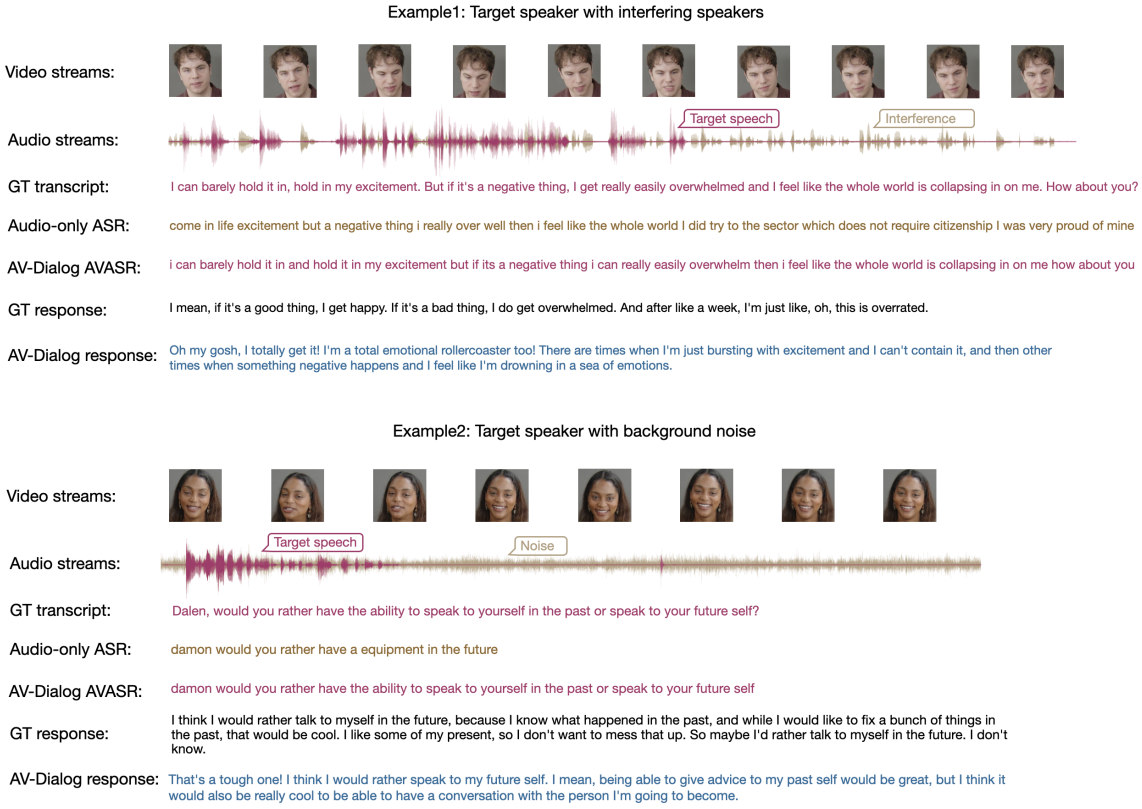

Figure 9: Data samples for our AV-dialog. Example 1 has interfering speakers while Example 2 has significant background noise. Video and audio streams are input to AV-Dialog. GT transcript is the ground-truth transcription of the target speaker. Audio-only ASR is the audio-only model output. AV-Dialog AVASR is the output of AVSR stream from AV-Dialog. GT response is the ground-truth response and AV-Dialog response is the output of our dialogue model.