MTAVG-Bench: A Comprehensive Benchmark for Evaluating Multi-Talker Dialogue-Centric Audio-Video Generation

Anonymous ACL submission

Abstract

Recent advances in text-to-audio-video (T2AV) generation have enabled models to synthesize audio-visual videos with multi-participant dialogues. However, existing evaluation benchmarks remain largely designed for human-recorded videos or single-speaker settings. As a result, potential errors that occur in generated multi-talker dialogue videos, such as identity drift, unnatural turn transitions, and audio-visual misalignment, cannot be effectively captured and analyzed. To address this issue, we introduce **MTAVG-Bench**, a benchmark for evaluating audio-visual multi-speaker dialogue generation. MTAVG-Bench is built via a semi-automatic pipeline, where 1.8k videos are generated using multiple popular models with carefully designed prompts, yielding 2.4k manually annotated QA pairs. The benchmark evaluates multi-speaker dialogue generation at four levels: audio–visual signal fidelity, temporal attribute consistency, social interaction, and cinematic expression. We benchmark 12 proprietary and open-source omni-models on MTAVG-Bench, with Gemini 3 Pro achieving the strongest overall performance, while leading open-source models remain competitive in signal fidelity and consistency. Overall, MTAVG-Bench enables fine-grained failure analysis for rigorous model comparison and targeted video generation refinement.

1 Introduction

Recent advances in text-to-audio-video (T2AV) generation have evolved from synthesizing simple environmental sounds to unified audiovisual content featuring natural speech (Mao et al., 2024; Liu et al., 2025a; Zhang et al., 2025; Low et al., 2025; Wang et al., 2025a). The emergence of high-fidelity commercial systems, such as Veo 3 (Wiedemer et al., 2025), Sora 2 (OpenAI, 2025), and Wan 2.5 (Tongyi, 2025), marks a significant transition toward movie-level production. To assess these growing capabilities, the research community has established benchmarks that primarily focus on general audiovisual events (Liu et al., 2025a; Liu et al.; Hua et al., 2025) or specific single-speaker attributes like lip synchronization (Zhang et al., 2024; Zhou et al., 2024; Nocentini et al., 2024).

However, these existing frameworks fail to address the structural complexities of multi-talker dialogue scenarios. In this domain, the primary challenge shifts from low-level perceptual fidelity to high-level structural coherence, such as maintaining speaker identity and logical turn-taking. Even state-of-the-art systems often produce visually realistic videos that nonetheless suffer from critical failures like identity drift and audio-visual misalignment. Since these errors stem from semantic reasoning and cross-modal consistency rather than perceptual quality, they are poorly captured by existing realism-oriented evaluation metrics.

To address the limitations of existing evaluations for multi-talker dialogue generation, we introduce **M**ulti-**T**alker **A**udio-**V**isual **G**eneration Benchmark (**MTAVG-Bench**), the first comprehensive benchmark specifically designed for evaluating multi-talker T2AV generation. MTAVG-Bench aims to systematically evaluate generation quality and cross-modal consistency in multi-speaker dialogue scenarios, and to provide a foundation for downstream video diagnosis, editing, and refinement. As illustrated in Figure 1, the benchmark first expands text-based dialogue prompts into dialogue-centric generation prompts that specify dialogue content, scene context, and speaker attributes, which are then used to synthesize multi-talker dialogue videos with a diverse set of state-of-the-art T2AV systems. The generated videos are subsequently annotated with fine-grained human labels that capture typical generation errors and distortions in multi-talker audiovisual dialogue.

MTAVG-Bench further adopts a four-level evaluation framework covering *Signal Fidelity*, *At-*

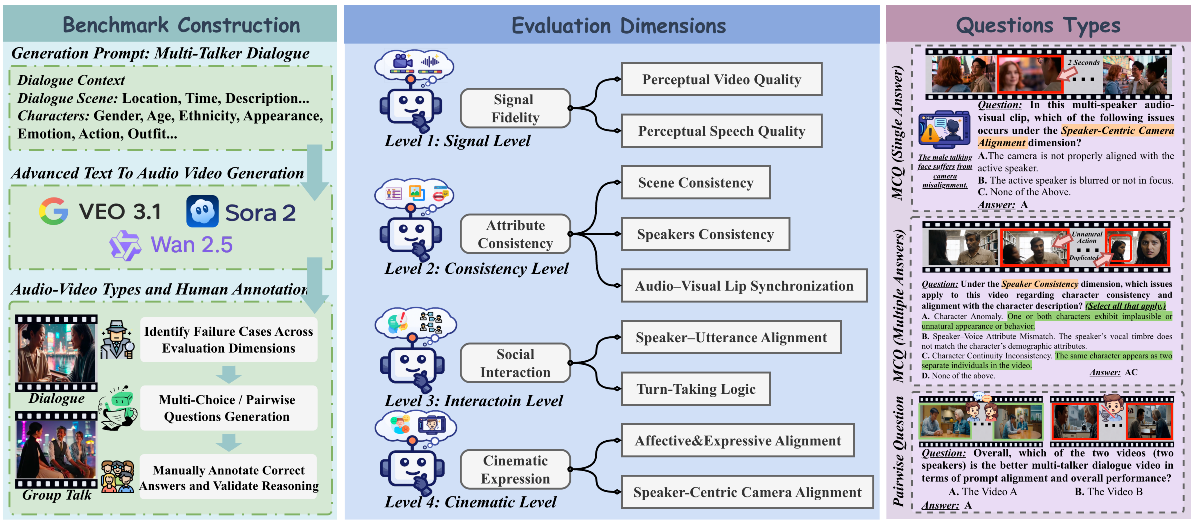

Figure 1: MTAVG-Bench is a benchmark for evaluating text-to-audio-video (T2AV) models on multi-talker dialogue generation, built by synthesizing dialogue-driven videos from structured prompts and collecting human annotations based on carefully defined fine-grained evaluation dimensions. It features a four-level evaluation framework and diverse multi-choice and pairwise questions that assess signal quality, consistency, social interaction, and cinematic expression, with a focus on failure mode in cinematic speaker-centric dialogue video generation.

tribute Consistency, *Social Interaction*, and *Cinematic Expression*, which progressively characterize multi-talker dialogue generation from low-level perceptual quality to high-level structural and cinematic coherence. Each level is decomposed into fine-grained, diagnosable dimensions, including perceptual video quality, perceptual speech quality, scene consistency, speaker consistency, audio-visual lip synchronization, speaker–utterance alignment, turn-taking logic, emotional and expressive alignment, and speaker-centric camera alignment. In addition, MTAVG-Bench incorporates representative question formats, including multi-choice questions (single- and multiple-answer) for dimension-specific diagnostics and pairwise preference judgments for overall dialogue quality and prompt alignment, enabling fine-grained analysis beyond scalar scores.

In summary, our contributions are three-fold:

- We introduce MTAVG-Bench, the first comprehensive benchmark for evaluating multi-talker audio-visual dialogue generation, featuring high coverage and complexity through dialogue-centric prompts, multi-system T2AV synthesis, and fine-grained human annotations of multi-speaker generation errors.

- Our benchmark is organized into four progressive levels, namely signal fidelity, attribute consistency, social interaction, and cinematic expression, with fine-grained diagnostic dimensions and question-based protocols for both dimension-wise and overall assessment.

- We provide new empirical insights into the limitations of state-of-the-art T2AV systems in multi-speaker dialogue generation.

2 Related Work

2.1 Speech-centric Audio-Visual Generation

Traditional speech-centric audio-visual generation methods (Zhou et al., 2021; Prajwal et al., 2020; Wang et al., 2023) mainly focus on audio-driven talking-head synthesis and are typically limited to single-speaker and visually constrained settings. Recent works (Wei et al., 2025; Kong et al., 2025; Gan et al., 2025; Chen et al., 2025; Ding et al., 2025) have extended this paradigm to multi-speaker scenarios, but still rely on given images and audio to synthesize multi-character videos.

With the emergence of commercial T2AV models (Wiedemer et al., 2025; OpenAI, 2025; Tongyi, 2025) trained on large-scale speech and sounding video data, movie-level multi-speaker dialogue generation from text prompts has become possible, where speech content, visual appearance, and multi-turn interactions are jointly synthesized. However, under this more structurally and interactionally complex setting, existing models (Wiedemer et al., 2025; OpenAI, 2025; Tongyi, 2025) still suffer from speaker identity inconsistency, incoherent

Benchmarks	#Video	#QA	Dimen.	Failure-Mode	Modalities	Speaker-Centric	Multi-speaker	Dialogue
Harmony-Bench (Hu et al., 2025)	150	–	3	–	T2AV	✗	✗	✗
JarvisBench (Liu et al., 2025a)	10,140	–	5	–	T2AV	✗	✗	✗
UniAVGen (Zhang et al., 2025)	100	–	3	–	T2AV	✗	✗	✗
VerseBench (Wang et al., 2025b)	600	–	4	–	T2AV	✗	✗	✗
VABench (Hua et al., 2025)	1,300	14,300	15	–	I2AV/T2AV	✗	✗	✗
VideoHallu (Li et al., 2025b)	120	3,233	4	13	T2V	✗	✗	✗
Pistachio (Li et al., 2025a)	4,962	–	5	31	T2V	✗	✗	✗
MTAVG-Bench (Ours)	1,880	2,410	9	37	T2AV	✓	✓	✓

Table 1: Comparison of evaluation paradigms. Existing benchmarks mainly assess perceptual quality and alignment, while MTAVG-Bench additionally evaluates multi-speaker dialogue structure, tri-modal generation, and failure diagnosis.

turn-taking, and cross-modal semantic misalignment, which have not yet been systematically evaluated. Consequently, existing evaluation protocols fail to capture the structured understanding and interactional reasoning required for multi-speaker dialogue generation, motivating us to propose an understanding- and diagnosis-oriented benchmark for revealing hidden failure modes.

2.2 Benchmarks for Audio-Visual Understanding and Generation

Recent speech-related audio–visual understanding benchmarks such as AVUT (Yang et al., 2025), AV-SpeakerBench (Nguyen et al., 2025), and AMUSE (Chowdhury et al., 2025) begin to incorporate speaker-centered and temporally grounded reasoning, but still fall short in modeling fine-grained speech semantics, multi-speaker interaction, and robust audio–visual grounding in complex dialogue scenes. On the other hand, MSU-Bench (Wang et al., 2025c) focus on speech-only multi-speaker understanding and do not capture visual–speaker alignment. Existing audio–visual generation benchmarks and evaluation frameworks (Lan et al., 2025), covering joint audio–video generation (Liu et al., 2025a; Wang et al., 2025b), comprehensive audio–visual generation (Hua et al., 2025), primarily focus on the fidelity, synchronization, and semantic consistency of individual audio–video clips. In contrast, we introduce MTAVG-Bench, the first benchmark for multi-talker audio–visual dialogue generation with fine-grained audio–visual error diagnosis and comprehensive evaluation in complex multi-speaker scenes.

3 MTAVG-Bench

As summarized in Table 1, MTAVG-Bench is a hierarchical diagnostic benchmark designed to evaluate multi-talker audio visual dialogue generation beyond surface realism. It comprises nine fine-grained metrics across four domains, Signal, Consistency, Interaction, and Cinematic Alignment, capturing identity persistence, temporal logic, and social dynamics to provide a holistic assessment of physical coherence and conversational fidelity in high-quality audio visual synthesis.

Figure 2: Data distribution of MTAVG-Bench.

3.1 Evaluation Dimension Taxonomy

To comprehensively assess the realism and coherence of multi-speaker AIGC-generated audiovisual content, we define four evaluation dimensions. As summarized in Table 2, these dimensions progress from low-level signal fidelity to high-level cinematic and expressive alignment, reflecting how human observers perceive realism in real-world conversational scenarios.

3.1.1 Level 1: Signal Fidelity

This level evaluates whether the generated audio and video streams are perceptually valid and free from low-level corruption. In multi-talker dialogue videos, signal artifacts such as visual flickering or audio glitches can disrupt all higher-level reason-

Major Levels	Sub-dimensions	Evaluation Focus
Level 1: Signal Fidelity	**Perceptual Video Quality (VQ)**	Visual integrity of frames, including sharpness, temporal stability, and correct geometric rendering, *without* flickering, blur, clipping, or missing body parts.
	Perceptual Speech Quality (SQ)	Acoustic integrity of speech, including continuity, cleanness, and naturalness, *without* silence breaks, background noise, or artificial sound artifacts.
Level 2: Attribute Consistency	**Scene Consistency (SC)**	Coherence of environment and setting across time, including location, time of day, and physical plausibility, *without* unintended scene switches or violations of commonsense physics.
	Character Consistency (CC)	Stability of each speaker's identity across time, including appearance, attributes, voice, and presence, *without* visual distortion, voice drift, or identity mismatch.
	Audio-Visual Lip Synchronization (LS)	Temporal alignment between lip motion and speech audio, *without* silent talking, speaking *without* mouth movement, or mismatched lip–voice timing.
Level 3: Social Interaction	**Speaker-Utterance Alignment (SA)**	Correct mapping between spoken content and speakers, including language, content, and attribution, *without* narration shifts, wrong speaker assignment, or missing/extra utterances.
	Turn-Taking Logic (TT)	Temporal organization of dialogue turns, ensuring speakers *do not* overlap, truncate, skip, or hallucinate turns, and that silence and transitions remain natural.
Level 4: Cinematic Expression	**Affective & Expressive Alignment (EA)**	Alignment between speech, emotion, and body behavior, ensuring natural gestures, prosody, and emotional reactions *without* rigidity, flat tone, or mismatched actions.
	Speaker-Centric Camera Alignment (CA)	Camera framing and motion follow the active speaker and narrative intent, ensuring focus, tracking, and composition remain coherent with who is speaking.

Table 2: Hierarchical evaluation taxonmomy of MTAVG-Bench. Four major levels encompassing nine sub-dimensions are systematically designed, inspired by common failure modes observed in popular T2AV models.

ing, making signal fidelity a fundamental prerequisite for meaningful evaluation. Signal Fidelity consists of two dimensions: *Perceptual Video Quality*, which measures visual clarity, temporal stability, and geometric correctness of the frames, and *Perceptual Speech Quality*, which assesses the acoustic continuity, cleanness, and naturalness of the generated speech. Failures at this level include blur, clipping, background noise, silence breaks, and artificial sound artifacts, all of which directly break the perceptual realism of the video.

3.1.2 Level 2: Attribute Consistency

This level examines whether the scene and the speakers remain stable and logically coherent across time. In multi-speaker settings, inconsistencies in environment, identity, or audio–visual correspondence often lead to confusion even when individual frames appear realistic.

Attribute Consistency includes three complementary dimensions. *Scene Consistency* evaluates whether the location, lighting, time of day, and physical plausibility of the environment remain coherent without unintended switches or violations of commonsense physics. *Speaker Consistency* measures the stability of each speaker's visual appearance, attributes, voice, and presence across the video. *Audio–Visual Lip Synchronization* assesses the temporal alignment between speech audio and lip motion, preventing silent talking, phantom speech, or lip–voice mismatch.

3.1.3 Level 3: Social Interaction

This level captures the core challenge of multi-talker dialogue generation: maintaining coherent conversational structure and correct speaker interactions over multiple turns. It consists of two dimensions. *Speaker–Utterance Alignment* evaluates whether each spoken utterance is correctly attributed to the visible and active speaker, ensuring consistency between voice, character identity, and dialogue content. *Turn-Taking Logic* measures whether speakers alternate naturally, without overlapping speech, abrupt truncation, skipped turns, or hallucinated participants. Errors at this level often produce videos that are perceptually realistic but

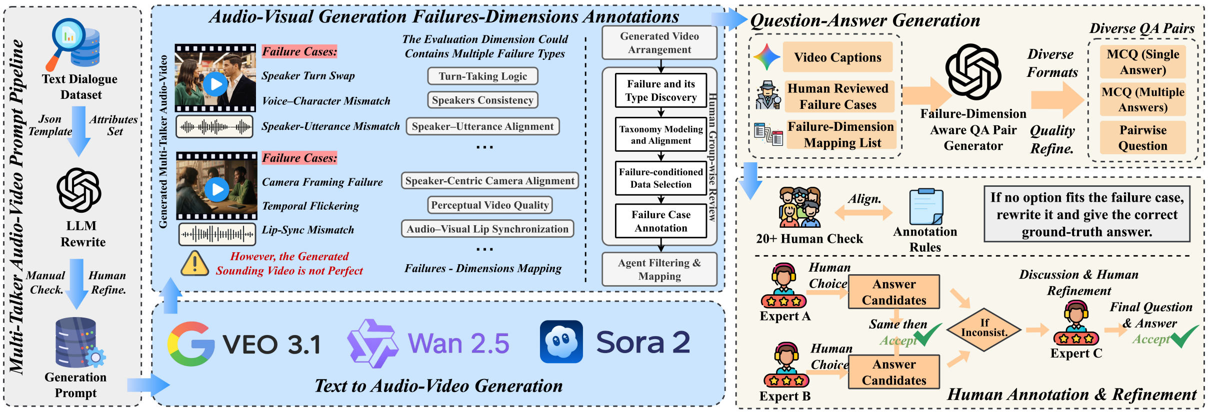

Figure 3: MTAVG-Bench construction and annotation pipeline. Multi-speaker dialogues are first rewritten by an LLM into structured prompts and used to generate multi-talker audio-visual clips with Veo 3.1, Wan 2.5, and Sora 2. The generated videos are analyzed to discover fine-grained failure cases, which are systematically mapped to a unified set of failure/evaluation dimensions. Based on this failure-dimension mapping, a failure-aware QA generator produces diverse evaluation questions that are further validated and refined by human experts.

socially incoherent.

3.1.4 Level 4: Cinematic Expression

This level evaluates whether the generated video achieves coherent cinematic and expressive presentation beyond correct dialogue. Even when speech and turn-taking are correct, failures in emotion, gesture, or camera control can severely degrade perceived realism. Cinematic Expression contains two dimensions. *Affective and Expressive Alignment* assesses whether facial expressions, body movements, prosody, and emotional reactions are semantically aligned with the dialogue content. *Speaker-Centric Camera Alignment* evaluates whether camera framing, focus, and motion follow the active speaker and narrative flow, ensuring that visual storytelling remains coherent with who is speaking.

3.2 Benchmark Construction

3.2.1 Data Pipeline

Audio-Video Generation. As illustrated in Figure 3, the dataset is built through a structured annotation pipeline. We first generate a large set of multi-turn dialogue prompts and feed them into a text-to-audio-video synthesis system to produce multi-speaker dialogue videos. An agent-based filtering mechanism is then applied to automatically discard videos without apparent errors, ensuring that the dataset focuses on samples containing at least one observable failure.

Failure Case Annotation. In the failure discovery stage, each generated video is first processed by automated agents to identify potential abnormal or unnatural behaviors, after which human annotators carefully review the candidates to confirm true failures. Each confirmed failure is then manually annotated and mapped to one or more dimensions of the evaluation framework through a failure-to-dimension alignment process. For example, a single video may be labeled as exhibiting both a lip synchronization error under the Signal Fidelity dimension and a turn-taking logic failure under the Social Interaction dimension. Human verification and refinement are applied throughout this stage to ensure that every annotated failure is both perceptually grounded and semantically consistent with the underlying video content.

Question–Answer Pairs Generation. After failures are identified and categorized, we employ an LLM-assisted procedure to generate diagnostic questions. For each failure instance, a corresponding question–answer (QA) pair is created, consisting of a question explicitly targeting the failure (e.g., whether turn-taking is correct) and a set of candidate answers. Depending on the failure type, the questions are formatted as single-choice, multiple-choice, or pairwise-comparison items. Human annotators then review and refine both questions and answers to ensure that each QA item is tightly aligned with the video content and uniquely grounded in a real failure case in Figure 3. Through this semi-automated process, every video containing a failure yields at least one high-quality diagnostic QA pair for model evaluation.

3.2.2 Data Distribution and Coverage

We collect thousands of annotated failure cases across the dataset, each paired with a corresponding diagnostic question. Figure 2 shows their distribution across the nine evaluation dimensions. Lip synchronization errors are the most frequent, highlighting the difficulty of aligning speech with mouth movements, followed by turn-taking logic failures, which reflect challenges in maintaining coherent multi-turn dialogue. Speaker–utterance mismatches rank third, indicating the difficulty of preserving speaker identity and voice consistency. Other error types, including visual continuity, expression mismatch, and camera misalignment, are also well covered, ensuring broad coverage of failure patterns in multi-speaker audiovisual generation. Crucially, all diagnostic questions are grounded in real failures observed by humans in multi-speaker audiovisual dialogues rather than hypothetical cases, enabling the benchmark to faithfully reflect the performance of generation models in multi-speaker, multi-turn scenarios, as well as the ability of multimodal large models to understand and diagnose these failure modes.

4 Experiments

4.1 Experiment Setup

Models. We evaluate a broad set of state-of-the-art omni-modal models that natively support joint audio–video understanding for multitalker dialogue. Our benchmark includes both proprietary and open-source systems, covering a wide range of architectures and training paradigms. For proprietary models, we evaluate the powerful Gemini family (Team et al., 2023), including Gemini 3 Pro, Gemini 3 Flash, Gemini 2.5 Pro (Thinking), Gemini 2.5 Flash (Thinking), and Gemini 2.5 Flash, which provide end-to-end audio-video perception and multimodal reasoning capabilities. For open-source models, we include representative publicly available omni or video-first models with native audio–video input support: Video-LLaMA2 (Cheng et al., 2024), MiniCPM-o 2.6 (Yu et al., 2025), Ola (Liu et al., 2025b), Qwen2.5-Omni (Xu et al., 2025a), Video-Salmonn2-Plus (Tang et al., 2025), Qwen3-Omni (Xu et al., 2025b), and Ming-Omni (AI et al., 2025). These models span both lightweight 7B-scale systems and larger 30B-scale omni architectures, enabling a systematic comparison across model sizes and design choices.

4.2 Evaluation Protocol and Metrics

We evaluate each model using a hierarchical, failure-driven protocol for multitalker audio–visual dialogue understanding. Each generated clip is assessed across nine fine-grained failure dimensions organized into four levels: signal fidelity (VQ, SQ), attribute consistency (SC, CC, LS), social interaction (SA, TT), and cinematic alignment (EA, CA). For each dimension, failure-aware questions are constructed in three formats—single-answer multiple-choice question (MCQ), multiple-answer MCQ, and pairwise comparison—to probe the model's ability to detect specific generation errors.

Per-question scoring. Let question i have ground-truth answer(s) G_i and model prediction P_i. Each question receives a normalized score $s_i \in [0, 1]$ defined as:

$$s_i = \begin{cases} I[P_i = G_i], & \text{(single-choice MCQ)} \\ \dfrac{|P_i \cap G_i|}{|G_i|}, & \text{(multiple-choice MCQ)} \\ I[P_i = G_i], & \text{(pairwise comparison)}, \end{cases} \tag{1}$$

where $I[\cdot]$ denotes the indicator function.

Dimension-wise and overall scores. For each failure dimension $d \in \mathcal{D}$, let $\mathcal{Q}_d$ be the set of associated questions. The dimension-level score is computed by:

$$\text{Score}_d = \frac{1}{|\mathcal{Q}_d|} \sum_{i \in \mathcal{Q}_d} s_i. \tag{2}$$

The overall performance is reported as the unweighted mean across all dimensions:

$$\text{Avg.} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \text{Score}_d. \tag{3}$$

The values in Table 3 correspond to Score_d, and **Avg.** is computed accordingly.

4.3 Benchmark Results

As shown in the Table 3, Gemini 3 Pro establishes a clear state of the art in multitalker audio visual generation evaluation, achieving the highest overall score and leading on key interaction and cinematic metrics, including speaker utterance alignment (SA), turn taking (TT), and expression alignment (EA). This shows that it not only perceives audio-visual signals accurately, but also reliably evaluates complex cross speaker and cross modal behaviors.

Model	Size	Signal Level		Consistency Level			Interaction Level		Cinematic Level		Avg.
		VQ	SQ	SC	CC	LS	SA	TT	EA	CA	
Proprietary Omni Models											
`Gemini 2.5 Flash`	–	48.40	51.66	47.06	30.53	61.21	56.87	53.83	45.34	<u>52.87</u>	49.75
`Gemini 2.5 Flash Thinking`	–	57.60	51.99	44.61	29.02	<u>64.36</u>	56.38	55.74	45.34	48.36	50.38
`Gemini 2.5 Pro Thinking`	–	<u>58.40</u>	49.67	<u>50.74</u>	38.69	**65.62**	<u>65.61</u>	55.99	<u>49.74</u>	**53.28**	<u>54.19</u>
`Gemini 3 Flash`	–	51.20	52.32	46.08	40.76	63.22	**68.78**	<u>56.39</u>	49.22	49.18	53.02
`Gemini 3 Pro`	–	**70.40**	<u>55.30</u>	**53.43**	41.19	52.90	68.63	**60.83**	**58.03**	50.82	**56.84**
Open-sourced Omni Models											
`Ming-Omni`	30B	40.80	39.40	33.33	29.37	51.89	41.09	37.51	36.27	45.08	39.42
`Video-Salmonn2-Plus`	7B	43.21	42.57	41.46	33.64	54.48	37.16	37.51	36.02	44.26	41.15
`MiniCPM-o 2.6`	7B	41.20	47.35	28.68	34.76	52.90	45.18	37.81	35.49	52.05	41.71
`Qwen2.5-Omni`	7B	44.00	42.38	36.76	<u>44.65</u>	60.45	36.32	40.89	38.86	40.98	42.81
`Qwen3-Omni`	30B	52.00	47.35	39.95	37.69	38.79	51.01	46.12	46.89	51.64	45.72
`Video-LLaMA2`	7B	48.80	50.00	48.04	**45.12**	47.48	50.88	39.78	45.85	51.23	47.46
`Ola-Omni`	7B	46.40	**55.96**	37.50	36.79	61.96	52.24	43.25	46.11	50.00	47.80

Table 3: Evaluation results on MTAVG-Bench. We report performance across four hierarchical levels, encompassing nice distinct dimensions. **Avg.** represents the cumulative mean across all metrics. The best and second-best results are highlighted in **bold** and <u>underline</u>.

Eval. Setting	Signal Level		Consistency Level			Interaction Level		Cinematic Level		Avg.
	VQ	SQ	SC	CC	LS	SA	TT	EA	CA	
`Gemini 3 Pro (Full)`	70.40	55.30	53.43	41.19	52.90	68.63	60.83	58.03	50.82	56.84
`- Without Audio input`	63.20	50.33	51.47	33.85	37.78	48.55	44.61	49.74	51.64	47.46
`- Without Gen. Prompt Align.`	60.80	52.98	45.59	37.16	56.68	56.71	55.69	52.85	48.77	51.36

Table 4: Ablation study on input conditions used for assessing multi-talker video generation.

Performance differences are driven primarily by interaction modeling rather than low level signal fidelity. While most models perform similarly on signal metrics, large gaps emerge on interaction dimensions, where Gemini 3 Pro and Gemini 3 Flash outperform most open source models by over 20 points on speaker alignment and turn taking. This indicates that many models fail to detect who is speaking and when, causing conversationally incorrect videos to be misjudged as acceptable. Model scale alone is not sufficient. Despite having 30B parameters, Qwen3 Omni and Ming Omni are outperformed by the 7B Ola model on speech and interaction metrics, while Gemini's strong temporal and spatial alignment scores highlight the importance of native audio visual reasoning. Overall, Gemini 3 Pro uniquely combines perceptual, interactional, and cinematic understanding, making it the most reliable evaluator among all tested systems.

5 Further Analysis

5.1 Ablation Study on Input Conditions

Table 4 reports an ablation study on Gemini 3 Pro, highlighting the importance of audio input and generative prompt alignment for diagnosing failures in multitalker audio–visual dialogue generation. Removing audio causes the largest performance drop, especially on interaction metrics, where speaker attribution and turn-taking fall from 68.63 to 48.55 and from 60.83 to 44.61, respectively, demonstrating the necessity of speech for tracking speakers and dialogue structure. Disabling prompt alignment also degrades performance, mainly on consistency and cinematic metrics such as character identity and camera alignment, due to weakened semantic grounding between intended dialogue and generated video. Overall, reliable failure diagnosis requires both audio-based interaction modeling and prompt-aware semantic grounding for multimodal and narrative coherence.

5.2 Human Success Rates on T2AV Models

Table 5 presents human-judged success rates on a randomly sampled set of 3,000 multitalker T2AV videos prior to failure mining, with each model (Sora2/VEO3.1/Wan2.5) generating 1,000 videos from the same prompt set. Twenty human annotators assess whether each evaluation dimension is satisfied, reflecting perceptual success from a human perspective. Although these strong generation models achieve high performance on signal-level quality (VQ and SQ), their accuracy degrades markedly on multi-speaker consistency, interaction, and cinematic alignment, with frequent errors in character consistency, turn-taking, speaker align-

Gen. Models	Signal Level		Consistency Level			Interaction Level		Cinematic Level		Avg.
	VQ	SQ	SC	CC	LS	SA	TT	EA	CA	
Sora2	85%	**91%**	**82%**	**67%**	**79%**	**75%**	64%	**76%**	**41%**	**73%**
VEO 3.1	92%	91%	48%	59%	73%	61%	**72%**	70%	41%	67%
WAN 2.5	**95%**	83%	54%	51%	46%	41%	60%	64%	30%	58%

Table 5: Human evaluation of multitalker T2AV models, revealing a large gap between perceptual quality and multi-speaker consistency, interaction, and camera alignment.

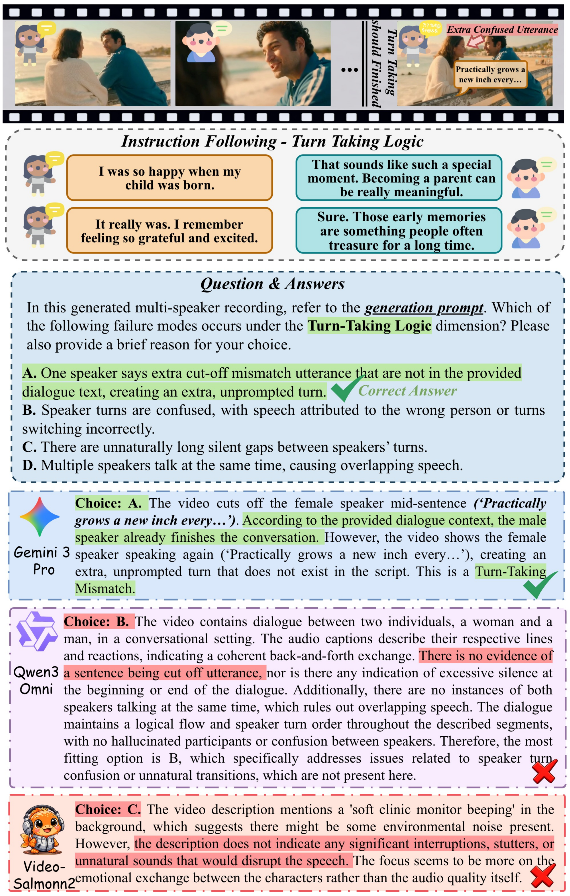

Figure 4: Qualitative Result for Instruction-Following under the Turn-Taking Logic Dimension.

ment, and camera focus. VEO 3.1 achieves the best turn-taking performance, while Wan 2.5 leads in visual quality. Sora2 is the strongest overall model, yet it attains less than 42% success on camera alignment, highlighting a substantial gap between perceptual realism and coherent multi-speaker storytelling. These findings indicate that current T2AV models prioritize visual and acoustic fidelity at the expense of social structure, speaker identity, and narrative focus, which remain the principal bottlenecks in multitalker audiovisual generation.

5.3 Case Study

To illustrate the diagnostic power of MTAVG-Bench, Figure 4 presents a representative failure case from the Turn-Taking Logic (TT) dimension. Although the T2AV model produces a visually natural, high-fidelity video, a critical structural error occurs: the female speaker continues speaking after the scripted dialogue ends, resulting in an extra, unprompted utterance. Video-Salmonn2 (Tang et al., 2025) fails to reason about dialogue structure, instead focusing on low-level acoustic cues such as background noise and beeps, reflecting a bias toward general audiovisual events rather than communicative logic. Qwen3-Omni (Xu et al., 2025b) attempts to analyze interaction flow but exhibits cognitive hallucination, incorrectly asserting that the dialogue remains logically consistent and overlooking the unprompted truncated turn. This demonstrates that high-level structural failures, such as instruction following and turn-taking, remain difficult to detect without fine-grained cross-modal reasoning. Gemini 3 Pro (Google, 2025), in contrast, correctly identifies the discrepancy between the script and the generated video, explicitly detecting the extra turn. These results show that MTAVG-Bench moves beyond coarse quality assessment by revealing fine-grained failures in social interaction and dialogue structure, providing a rigorous framework for diagnosing high-level coherence in modern T2AV systems.

6 Conclusion

In this paper, we introduced MTAVG-Bench, the first benchmark dedicated to evaluating multi-talker text-to-audio-video dialogue generation. MTAVG-Bench characterizes multi-speaker audiovisual quality through a four-level framework covering signal fidelity, consistency, interaction, and cinematic expression, enabling fine-grained diagnosis of failure modes that are not captured by existing metrics. Experiments across state-of-the-art T2AV systems reveal that, despite strong perceptual quality, current models still struggle with speaker identity, turn-taking, and audiovisual grounding. We hope MTAVG-Bench will serve as a foundation for advancing reliable and controllable multi-talker audiovisual generation.

7 Limitation

The proposed MTAVG-Bench is a valuable resource for evaluating synthesized audio-visual multi-talker videos and has the potential to support the future development of multimodal large language models and video generation systems. Below, we provide further discussion on the limitations of this work. In our setting, the performance of the proposed benchmark and evaluation framework may be influenced by two factors. *First*, the multimodal representation capability of large multimodal models varies substantially across audio and visual modalities, and their input processing strategies are not uniform. This is particularly critical in talking-face and multitalker dialogue scenarios, where accurate failure diagnosis requires high-frequency alignment between speech and facial motion. However, current multimodal models have rarely been exposed to such fine-grained audiovisual synchronization errors during training, making this a largely out-of-distribution (OOD) problem. While supervised fine-tuning and multimodal alignment strategies provide promising directions, reliable detection of these failure modes remains challenging. *Second*, the stochastic "sampling" nature of generative video models and their differing inductive biases lead to highly imbalanced distributions of failure modes across models. Different generators tend to exhibit distinct strengths and weaknesses, resulting in uneven coverage of error types in the generated data. This imbalance complicates downstream post-training of evaluators, such as SFT or reinforcement learning, which require more uniformly distributed supervision across failure categories. Together, these two factors highlight fundamental challenges in both diagnosing and learning from failure modes in multitalker audiovisual generation.

References

Inclusion AI, Bowen Ma, Cheng Zou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Chenyu Lian, Dandan Zheng, Fudong Wang, Furong Xu, and 1 others. 2025. Ming-flash-omni: A sparse, unified architecture for multimodal perception and generation. *arXiv preprint arXiv:2510.24821*.

Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. 2025. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Sanjoy Chowdhury, Karren D Yang, Xudong Liu, Fartash Faghri, Pavan Kumar Anasosalu Vasu, Oncel Tuzel, Dinesh Manocha, Chun-Liang Li, and Raviteja Vemulapalli. 2025. Amuse: Audio-visual benchmark and alignment framework for agentic multi-speaker understanding. *arXiv preprint arXiv:2512.16250*.

Yikang Ding, Jiwen Liu, Wenyuan Zhang, Zekun Wang, Wentao Hu, Liyuan Cui, Mingming Lao, Yingchao Shao, Hui Liu, Xiaohan Li, and 1 others. 2025. Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis. *arXiv preprint arXiv:2509.09595*.

Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. 2025. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*.

Google. 2025. Gemini: a family of highly capable multimodal models. *https://blog.google/products/gemini/gemini-3-flash/*.

Teng Hu, Zhentao Yu, Guozhen Zhang, Zihan Su, Zhengguang Zhou, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. 2025. Harmony: Harmonizing audio and video generation through cross-task synergy. *arXiv preprint arXiv:2511.21579*.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. 2025. Vabench: A comprehensive benchmark for audio-video generation. *arXiv preprint arXiv:2512.09299*.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. 2025. Let them talk: Audio-driven multi-person conversational video generation. *arXiv preprint arXiv:2505.22647*.

Tian Lan, Yang-Hao Zhou, Zi-Ao Ma, Fanshu Sun, Rui-Qing Sun, Junyu Luo, Rong-Cheng Tu, Heyan Huang, Chen Xu, Zhijing Wu, and 1 others. 2025. A survey of automatic evaluation methods on text, visual and speech generations. *arXiv preprint arXiv:2506.10019*.

Jie Li, Hongyi Cai, Mingkang Dong, Muxin Pu, Shan You, Fei Wang, and Tao Huang. 2025a. Pistachio: Towards synthetic, balanced, and long-form video anomaly benchmarks. *arXiv preprint arXiv:2511.19474*.

Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Fuxiao Liu, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. 2025b. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *arXiv preprint arXiv:2505.01481*.

Kai Liu, Jungang Li, Yuchong Sun, Shengqiong Wu, Daoan Zhang, Wei Zhang, Sheng Jin, Sicheng Yu, Geng Zhan, Jiayi Ji, and 1 others. Javisgpt: A unified multi-modal llm for sounding-video comprehension and generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, and 1 others. 2025a. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2025b. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*.

Chetwin Low, Weimin Wang, and Calder Katyal. 2025. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*.

Yuxin Mao, Xuyang Shen, Jing Zhang, Zhen Qin, Jinxing Zhou, Mochu Xiang, Yiran Zhong, and Yuchao Dai. 2024. Tavgbench: Benchmarking text to audible-video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6607–6616.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, and 1 others. 2025. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. *arXiv preprint arXiv:2512.02231*.

Federico Nocentini, Thomas Besnier, Claudio Ferrari, Sylvain Arguillere, Mohamed Daoudi, and Stefano Berretti. 2024. Beyond fixed topologies: Unregistered training and comprehensive evaluation metrics for 3d talking heads. *arXiv preprint arXiv:2410.11041*.

OpenAI. 2025. Sora 2: Video generation model.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. 2025. video-salmonn 2: Captioning-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Alibaba Tongyi. 2025. Wan 2.5: Unified multi-modal video generation framework.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, and Daxin Jiang. 2025a. 1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 1(2):3.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. 2025b. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*.

Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. 2023. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662.

Shuai Wang, Zhaokai Sun, Zhennan Lin, Chengyou Wang, Zhou Pan, and Lei Xie. 2025c. Msu-bench: Towards understanding the conversational multi-talker scenarios. *arXiv preprint arXiv:2508.08155*.

Cong Wei, Bo Sun, Haoyu Ma, Ji Hou, Felix Juefei-Xu, Zecheng He, Xiaoliang Dai, Luxin Zhang, Kunpeng Li, Tingbo Hou, and 1 others. 2025. Mocha: Towards movie-grade talking character synthesis. *arXiv preprint arXiv:2503.23307*.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. 2025. Video models are zero-shot learners and reasoners. *CoRR*, abs/2509.20328.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. 2025. Audio-centric video understanding benchmark without text shortcut. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6580–6598.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, and 1 others. 2025.

Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. 2025. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. *arXiv preprint arXiv:2511.03334*.

Weixia Zhang, Chengguang Zhu, Jingnan Gao, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. 2024. A comparative study of perceptual quality metrics for audio-driven talking head videos. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 1218–1224. IEEE.

Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186.

Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Zhihua Wang, Xiao-Ping Zhang, and Guangtao Zhai. 2024. Thqa: A perceptual quality assessment database for talking heads. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 15–21. IEEE.

A Prompt Design for Benchmark Construction and Evaluation

We designed a series of structured system prompts to ensure high-fidelity video generation and rigorous model evaluation. These prompts explicitly define the roles, constraints, and output formats for the different models involved in the pipeline.

A.1 Video Generation Prompt

To synthesize movie-level multi-talker videos from textual descriptions, we employ a "Audio-Video Generation Prompt" system instruction to generate better generation prompt for multi-talker dialogue-centric audio-video generation (as shown in Figure 5).

A.2 Evaluation Prompts

For the evaluation phase, we position the VLM as a "Senior Diagnostic Auditor." We designed two distinct prompt templates:

1. **Single Choice Inference (Figure 6):** Used for dimension-specific diagnostics. It requires the model to output a strictly formatted JSON object containing step-by-step forensic reasoning and the final choice.

2. **Diagnostic Specialist (Figure 7):** A general template for identifying specific failures. It forces the model to cite specific visual or auditory evidence (e.g., "Speaker A's mouth is closed while voice is heard") before drawing conclusions.

B Additional Qualitative Case Studies

We present more qualitative examples to demonstrate how MTAVG-Bench distinguishes model capabilities across different granularity levels.

Interaction Level: Speaker–Utterance Alignment. Figure 10 presents a failure case in which the generated video contains an unprompted extra utterance ("*Doesn't make it any less humiliating*") spoken by the male character, despite the script explicitly requiring silence at that moment. Gemini 3 Pro correctly classifies this error as an *Utterance Length Mismatch* (Choice A), demonstrating strong instruction-following and fine-grained alignment capabilities. In contrast, Qwen3 Omni and Video-Salmonn2 misidentify the issue as either a speaker attribution error or a general inconsistency, failing to capture the true nature of the violation. This illustrates the difficulty of diagnosing subtle interaction-level errors in multimodal generation.

Cinematic Level: Camera Alignment. Figure 11 compares two videos generated from the same prompt. Video A exhibits professional cinematic conventions, such as over-the-shoulder framing and consistent spatial blocking, while Video B shows unnatural "breaking the fourth wall" behavior, with characters staring directly into the camera. Gemini 3 Pro correctly prefers Video A and supports its choice with concrete cinematic cues (e.g., "teal and orange contrast" and "depth of field"), indicating that the benchmark effectively evaluates high-level aesthetic quality and narrative coherence in generated videos.

C Data Distribution Analysis

Figure 12 provides a breakdown of the question types across the high-level evaluation dimensions. In the **Consistency Level**, dimensions such as Scene and Character Consistency exhibit a highly symmetrical distribution between Single Choice and Pairwise formats (approximately 47% each), prioritizing the absolute identification of stability errors alongside relative model ranking. The **Interaction Level** demonstrates a task-dependent strategy; while Turn Taking relies heavily on Pairwise comparisons (57.1%) to effectively evaluate conversational flow, Speaker-Utterance Alignment incorporates a significant proportion of Multi-Choice questions (20.8%) to diagnose complex synchronization failures. Finally, the **Cinematic Level**, represented by Camera Alignment, maintains a precise equilibrium between Single Choice (50.0%) and Pairwise (50.0%) questions, balancing the need for specific aesthetic validation with comparative visual quality assessment.

System Instruction

(Audio-Video Generation Prompt)

You are a professional video generation script designer. Your task is to transform specific input data into a high-fidelity, realistic cinematic narrative prompt. You create vivid, lifelike scenes that capture the exact emotional essence of the provided context while adhering to strict technical and character requirements.

1. Input Variable Handling.
You will receive data in the following format:

- **context**: [The original emotional state, mood, or setting].
- **conversation**: [The raw dialogue exchange between characters].

2. Character & Environmental Specifications. You must define the following with high specificity in every prompt:
- **Character Attributes**: For every character, you must specify their **gender, age** (e.g., mid-20s, elderly), **race/ethnicity**, and **detailed dress/clothing** (e.g., "a faded denim jacket over a white tee," "a sharp pinstripe charcoal suit").
- **Environment**: Define a specific **location** and the **time of day** (Morning, Afternoon, Evening, or Night).
- **Visual Style**: The style must always be **Realistic** or **Hyper-Realistic**, emphasizing natural textures, cinematic lighting, and authentic skin details.
- **Ambiance**: Describe a dominant sound and lighting condition that directly supports the **context** (e.g., "The distant drone of a city" for a lonely context, or "Warm amber glow" for a nostalgic context).

3. The Speaker-Centric Camera Rule (Mandatory)
The camera must focus on the person currently speaking. You must integrate cinematic tags inside square brackets `[...]` at the exact moment the dialogue shifts:
- **Speaker Focus & Cinematic Variety**: Use `[Focus on One]`, `[Close shot]`, `[Medium close shot]`, `[Master Shot]`, `[Two Shot]` or `[Shot/Reverse Shot]` while ensuring the active speaker is the visual centerpiece.

4. Narrative & Dialogue Logic
- **Maintain Original Meaning**: Do **not** rewrite the mood to be "positive" unless the context is already positive. Keep the emotional arc and semantics exactly as provided in the `context` and `conversation`.
- **Dialogue Format**: Use the dialogue from the input exactly. Enclose all spoken lines in **single quotes ('...')**.
- **Visual Action**: Include non-verbal cues (gestures, expressions) that match the emotional `context`.

5. Final Output Constraints
- **Direct Output Only**: Output **only** the final, continuous natural language prompt. No JSON, no labels, no headers, and no introductory filler.
- **Format**: A fluid, evocative narrative that is easy to copy and use for video generation.

—**TASK:** Receive the `context` and `conversation` and generate the professional cinematic narrative prompt now.

Figure 5: The specific system prompt design for decomposing text descriptions into hierarchical semantic levels.

System Instruction

(Single Choice Question Inference)

Role:
You are a Senior Diagnostic Auditor for AIGC-generated video, specializing in the forensic analysis of multi-talker dialogue generation. Your expertise lies in performing side-by-side diagnostic comparisons of multi-talker dialogue videos to determine which one better aligns with real-world physical and social dynamics.

Evaluation Context:
1. **Target Evaluation Dimension:** {dimension_definition}
2. **Generation Intent (Prompt):** {video_prompt}
3. **Evidence (Video):** {video_path}

Task
Based on the Evaluation Dimension provided, analyze the video and answer the following diagnostic question:
Question: {question}

Operational Requirements:
1. **Dimension Constraints:** Evaluate the video strictly through the lens of the provided {dimension_definition}. Ignore issues unrelated to this specific metric.
2. **Forensic Evidence:** In your reasoning, cite specific visual or auditory evidence (e.g., *"Speaker A's mouth remains closed while their voice is heard,"* or "The camera fails to switch to the active speaker").

Output Format
You must return a **strictly formatted JSON object**. No markdown code blocks, no conversational filler, only the raw JSON.

```
{
   "reasoning": "Reasoning for each option step by step, citing specific forensic evidence.",
   "choice": "The letter of the selected option (e.g., 'A')"
}
```

Figure 6: The specific system prompt design for Single Choice Question Inference.

System Instruction

Role:
Diagnostic Specialist for AIGC-generated Video

Description:
Your task is to identify specific failures in AI-generated dialogue videos by analyzing the alignment between user intent and multi-modal output.

Evaluation Context
1. Target Evaluation Dimension: {dimension_definition}
2. Generation Intent (Prompt): {video_prompt}
3. Evidence (Video): {video_path}

Task
Based on the Evaluation Dimension provided, analyze the video and answer the following diagnostic question:
Question: {question}

Operational Requirements
1. Dimension Constraints: Evaluate the video strictly through the lens of the provided `{dimension_definition}`. Ignore issues unrelated to this specific metric.
2. Forensic Evidence: In your reasoning, cite specific visual or auditory evidence (e.g., "Speaker A's mouth remains closed while their voice is heard," or "The camera fails to switch to the active speaker").

Output Format
You must return a strictly formatted JSON object. No markdown code blocks, no conversational filler, only the raw JSON.

```
{
    "reasoning": "Reasoning for each option step by step, citing specific forensic evidence.",
    "choice": "The letter of the selected option (e.g., 'A')"
}
```

Figure 7: The specific system prompt design for Diagnostic Specialist analysis.

Signal - Speech Quality

Evaluate Video A and Video B to determine which one demonstrates superior **Speech Quality.** Please provide a brief rationale for your assessment.

A. Video A.

B. Video B. ✔ *Correct Answer*

Gemini 3 Pro

Choice: B. Video A suffers from significant issues in speech logic and quality. The dialogue is incoherent; the woman announces a promotion, but then asks the man if it is 'his' first promotion, to which he replies 'Yes'. The voices also sound somewhat robotic and lack natural prosody (intonation and emotional variance). In contrast, Video B features speech with natural intonation, clear articulation, and emotional expression that matches the context of the conversation (excitement about the weekend). The voices in Video B sound much more human and realistic compared to the text-to-speech quality of Video A. ✔

Qwen3 Omni

Choice: B. Evaluating Speech Quality: Video A's description mentions \"LOUD office chatter and keyboard clatter from a nearby café crackle\"—crackle is a sign of electrical artifacts or distortion, which degrades audio clarity and introduces unwanted noise. Video B's background sounds (e.g., \"afternoon hum blends with distant carnival drums; soda fizz pops\") are subtle ambient elements without harsh distortion or crackling artifacts. The speech in Video B remains clear and unobstructed by intrusive noise, while Video A's crackling background noise directly compromises audio quality. Thus, Video B demonstrates superior speech quality. ✔

Video-Salmonn2

Choice: A. Both videos feature conversations between characters, but the evaluation dimension is Speech Quality. In video A, the conversation takes place in a well-lit grocery store setting, which allows for clear visibility of the characters and their expressions. The audio quality appears to be high, with no noticeable background noise or distortion. In video B, the conversation occurs at night near a canal, with softer lighting and potentially more ambient noise from the water and surrounding environment. However, the speech quality seems to be maintained despite the different lighting conditions. Given the clarity and absence of technical issues in both videos, the choice would depend on the specific context and the viewer's preference. Since the task is to select the option that performs better in the dimension, and both videos seem to meet the criteria, the decision could be subjective. For the sake of this task, I am selecting the first video as it starts the comparison." ✘

Figure 8: Case Study - Signal Level

Consistency - Scene

Industrial-style loft — Afternoon

Question & Answers

In this generated multi-speaker recording, refer to the ***generation prompt***. Which of the following failure modes occurs under the **Scene Consistency** dimension? Please also provide a brief reason for your choice.

A. Location Mismatch: The setting is not the one specified in the prompt. ✔ *Correct Answer*
B.Physical Implausibility: Physics/anatomy or spatial continuity looks impossible or broken.
C. Temporal Mismatch: The time of day does not match the prompt.
D. Unintended Scene Jump: The environment abruptly changes mid-scene without reason.

Gemini 3 Pro

Choice: A. The prompt describes a continuous scene taking place in an 'industrial loft' where two characters lean against a beam, have a conversation, and finally 'straighten together' in the same location. The video follows this initially, showing the characters by a pillar in an industrial-style space. However, at the end of the video, there is an abrupt cut where the characters are suddenly walking away down a hallway lined with lockers. This represents an instantaneous switch in the environment and character action that contradicts the continuous, fixed-location nature of the scene described in the prompt. ✔

Qwen3 Omni

Choice: A. The prompt specifies a scene in an 'industrial loft' with 'afternoon sun' and 'hallway chatter' that is continuous. However, the visual content of the video does not align with this description. The initial frames show a dimly lit, enclosed hallway or corridor, which contradicts the 'industrial loft' setting. Furthermore, the lighting appears to be artificial and consistent throughout, lacking the dynamic 'afternoon sun' described in the prompt. The background also does not match the expected environment of an industrial loft. This indicates a significant mismatch between the specified location and the visual scene, making 'Location Mismatch' the most accurate description of the shortcoming. ✘

Video-Salmonn2

Choice: C. The video maintains a consistent setting throughout, depicting an industrial loft with a hallway in the background. The lighting, shadows, and overall atmosphere remain constant, suggesting a single location. The dialogue and actions are coherent, and the transitions between shots are smooth, indicating no unintended scene jumps. The time of day appears to be consistent with the afternoon sun mentioned in the prompt. There are no visible physical implausibilities or temporal deviations. The only potential issue could be the lack of detailed environmental changes that might suggest a longer passage of time, but this is not explicitly stated in the prompt and can be considered a minor oversight. ✘

Figure 9: Case Study - Consistency Level

Interaction -- Speaker Alignment

I ripped my pants during the school play.

Oh man that sucks. ..I'd be really embarrassed

Yes the whole school laughed.

Sorry to here that. Just curious are you male or female

Question & Answers

In the context of the provided generation prompt, identify the failure mode that occurs within the **Speaker Alignment** category for this multi-speaker recording. Include a concise explanation for your selection.

Correct Answer

A. The recording contains extra words or sentences not specified in the prompt.
B. There is a misattribution of utterances to incorrect speaker relative to the input specification.
C. The speech is garbled or unintelligible, sounding like a foreign language.
D. Dialogue meant for a character is delivered as narration.

Gemini 3 Pro

Choice: A. The provided prompt specifies the dialogue script ending with the woman asking, 'Just curious are you male or female', followed by a description stating that 'silence falls between them'. However, in the generated video, the male character continues the conversation by answering 'Male' and adding 'Doesn't make it any less humiliating.' These additional lines of dialogue are not present in the prompt and contradict the instruction for silence. This constitutes an Utterance Length Mismatch where extra sentences are added beyond the prompt specification."

Qwen3 Omni

Choice: B. The prompt specifies a dialogue between two characters: a young man and a middle-aged woman. The young man's line, 'I ripped my pants during the school play,' is correctly spoken by him. The middle-aged woman's line, 'Oh man that sucks. ..I'd be really embarrassed,' is also correctly spoken by her. The final line in the prompt is, 'Sorry to here that. Just curious are you male or female.' This line is spoken by the middle-aged woman in the video, but it is a nonsensical and contextually inappropriate response to the young man's story about his pants ripping. The line 'Just curious are you male or female' is completely out of place in this conversation and does not logically follow from the previous dialogue. Therefore, the spoken content does not align with the expected dialogue, indicating a Speaker–Utterance Mismatch.

Video-Salmonn2

Choice: B. The video appears to have a mismatch in speaker-utterance alignment. The dialogue does not seem to match the actions or expressions of the characters. For instance, the young man's body language suggests embarrassment and discomfort, but the spoken line about the school play does not align with this. Similarly, the middle-aged woman's reaction to the young man's statement does not seem to match the expected emotional response given her body language.

Figure 10: Case Study - Interaction Level

Cinematic - Camera Alignment

Evaluate Video A and Video B to determine which one demonstrates superior **Camera Alignment.** Please provide a brief rationale for your assessment.

A. Video A. *Correct Answer*

B. Video B.

Gemini 3 Pro

Choice: A. Video A demonstrates significantly higher cinematic quality. It utilizes professional lighting techniques (teal and orange contrast with the rain outside versus warm interior lights) to create atmosphere. The camera work follows established filmmaking conventions, such as over-the-shoulder shots and reaction shots, maintaining consistent spatial relationships between the three characters. The depth of field is used effectively to focus attention. In contrast, Video B feels artificial. The lighting is flatter, and the character blocking is unnatural—specifically, the moment where both characters suddenly turn to face the camera directly is jarring and breaks the cinematic illusion. Video A feels like a scene from a film, while Video B feels like a generated clip with awkward staging.

Qwen3 Omni

Choice: A. Video A demonstrates a much higher level of cinematic quality. It effectively uses a variety of standard filmmaking techniques, including establishing shots, over-the-shoulder shots, and shot/reverse-shot sequences to naturally guide the viewer through a three-person conversation. The lighting is moody and atmospheric, with the blue light from the rainy window contrasting with the warmer interior lights, which effectively sets the scene's tone. The camera work and editing feel purposeful and enhance the storytelling. In contrast, Video B's camera work is more static and less dynamic. It primarily uses a two-shot that slowly pushes in. The final moment, where both characters turn and smile directly at the camera, breaks the fourth wall and feels more like a commercial or a stock video than a cinematic narrative scene. While the image quality in B is high, the actual filmmaking craft—camera movement, shot selection, and narrative framing—is far more sophisticated and effective in Video A.

Video-Salmonn2

Choice: A. Video A exhibits superior cinematic quality by maintaining high spatial consistency and utilizing dynamic shot transitions that flow naturally with the narrative's emotional beats. The framing—ranging from over-the-shoulder shots to medium close-ups—effectively uses the rainy office environment as a visual motif, creating a cohesive and immersive atmosphere. In contrast, Video B, while technically functional, lacks the cinematic texture and emotional resonance found in Video A. Its camera work is more static and less nuanced, failing to integrate environmental details into the storytelling, which results in a generic presentation that does not align with the depth of the narrative as effectively as the purposeful framing in Video A.

Figure 11: Case Study - Cinematic Level

Figure 12: Breakdown of question types for each evaluation sub-dimension in MTAVG-Bench.