

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

Yang-Hao Zhou^{1,†,‡}, Haitian Li^{2,†}, Rexar Lin^{1,†},
Heyan Huang^{1,*}, Jinxing Zhou³, Changsen Yuan⁴, Tian Lan¹, Ziqin Zhou⁵,
Yudong Li⁶, Jiajun Xu⁷, Jingyun Liao⁷, Yiming Cheng⁶,
Xuefeng Chen⁷, Xian-Ling Mao¹, Yousheng Feng⁷

¹Beijing Institute of Technology, ²Shanghai University, ³OpenNLP Lab, ⁴Beijing University of Technology,

⁵The University of Adelaide, ⁶Tsinghua University, ⁷Inkeverse Group Limited.

*Corresponding author. †Equal Contribution. ‡Project Leader.

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, structural failures in generated multi-talker dialogue videos, such as identity drift, unnatural turn transitions, and audio-visual misalignment, cannot be effectively diagnosed. To address this issue, we introduce **MTAVG-Bench**¹, a failure-driven diagnostic benchmark for multi-talker dialogue-centric audio-video generation. MTAVG-Bench is built via a semi-automatic pipeline, where 1.8k videos are generated using mainstream T2AV models with carefully designed prompts, yielding 2.4k manually annotated QA pairs for fine-grained failure diagnosis. The benchmark evaluates multi-speaker dialogue generation at four levels: audio-visual signal fidelity, temporal attribute consistency, social interaction, and cinematic expression. Built on a hierarchical failure taxonomy and a targeted QA protocol, MTAVG-Bench is primarily designed to evaluate whether proprietary and open-source omni-models can reliably identify failure modes in multi-speaker T2AV outputs. 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 audio-visual

¹Our benchmark is available at <https://mortyzhou77.github.io/MTAVG-Bench-project-page>

content featuring natural speech (Mao et al., 2024; Liu et al., 2025a; Zhang et al., 2025; Low et al., 2025; HaCohen et al., 2026; Wang et al., 2025a; Team et al., 2026). 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 audio-visual 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) and interactive talking face (Zhou et al., 2022, 2025; Zhu et al., 2025).

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 **Multi-Talker Audio-Visual Generation Benchmark (MTAVG-Bench)**, a diagnostic benchmark for multi-talker dialogue-centric audio-video generation. MTAVG-Bench aims to systematically characterize structural failures and cross-modal inconsistencies in multi-speaker dialogue scenarios, while providing 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

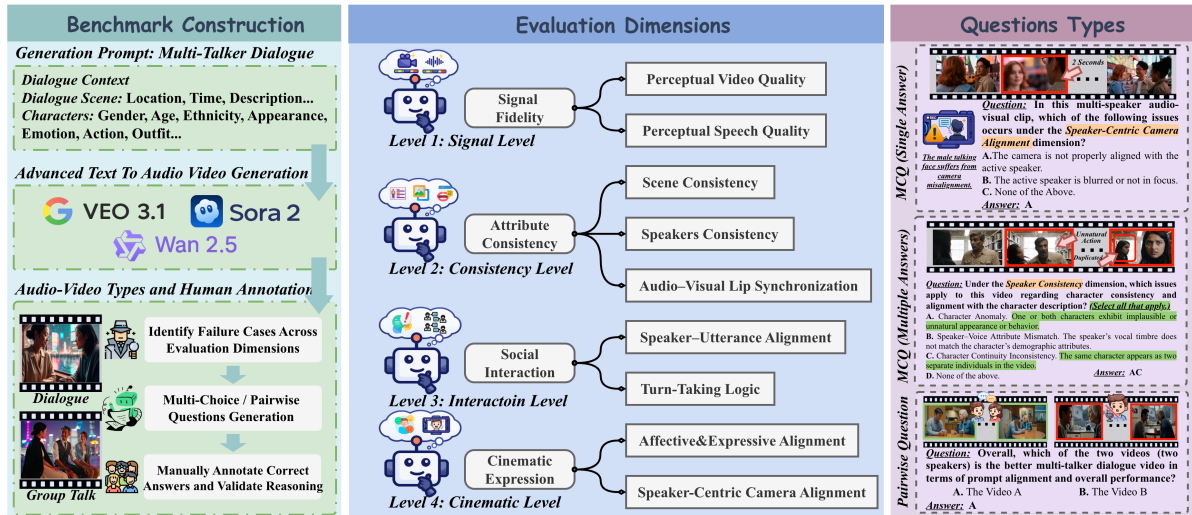


Figure 1: MTAVG-Bench is a benchmark for evaluating text-to-audio-video (T2AV) models on multi-talker dialogue generation, built by synthesizing dialogue-driven videos from structured prompts and collecting human annotations based on carefully defined fine-grained evaluation dimensions. It features a 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.

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 audio-visual dialogue.

MTAVG-Bench further adopts a four-level evaluation framework covering *Signal Fidelity*, *Attribute 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, a failure-driven diagnostic benchmark for multi-talker

dialogue-centric audio-video generation, featuring high coverage and complexity through dialogue-centric prompts, multi-system T2AV synthesis, and fine-grained human annotations of multi-speaker failure cases.

- MTAVG-Bench 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 current omni-modal evaluators and state-of-the-art T2AV systems in multi-speaker dialogue settings.

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; Zhou 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.

Benchmarks	#Video	#QA	Dimen.	Failure-Mode	Modalities	Speaker-Centric	Multi-speaker	Dialogue
Harmony-Bench (Hu et al., 2025)	150	–	3	–	T2AV	✗	✗	✗
JavisBench (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.

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 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

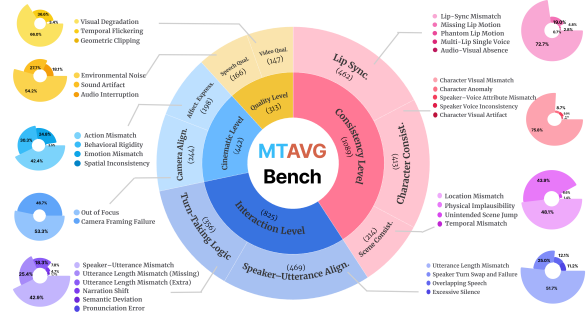


Figure 2: Data distribution of MTAVG-Bench.

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.

3.1 Evaluation Dimension Taxonomy

To comprehensively assess the realism and coherence of multi-speaker AIGC-generated audio-visual 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.

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, <i>without</i> flickering, blur, clipping, or missing body parts.
	Perceptual Speech Quality (SQ)	Acoustic integrity of speech, including continuity, cleanness, and naturalness, <i>without</i> 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, <i>without</i> 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, <i>without</i> visual distortion, voice drift, or identity mismatch.
	Audio-Visual Lip Synchronization (LS)	Temporal alignment between lip motion and speech audio, <i>without</i> silent talking, speaking <i>without</i> 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, <i>without</i> narration shifts, wrong speaker assignment, or missing/extra utterances.
	Turn-Taking Logic (TT)	Temporal organization of dialogue turns, ensuring speakers <i>do not</i> 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 <i>without</i> 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 taxonomy of MTAVG-Bench. Four major levels encompassing nine sub-dimensions are systematically designed, inspired by common failure modes observed in popular T2AV models.

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 reasoning, 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

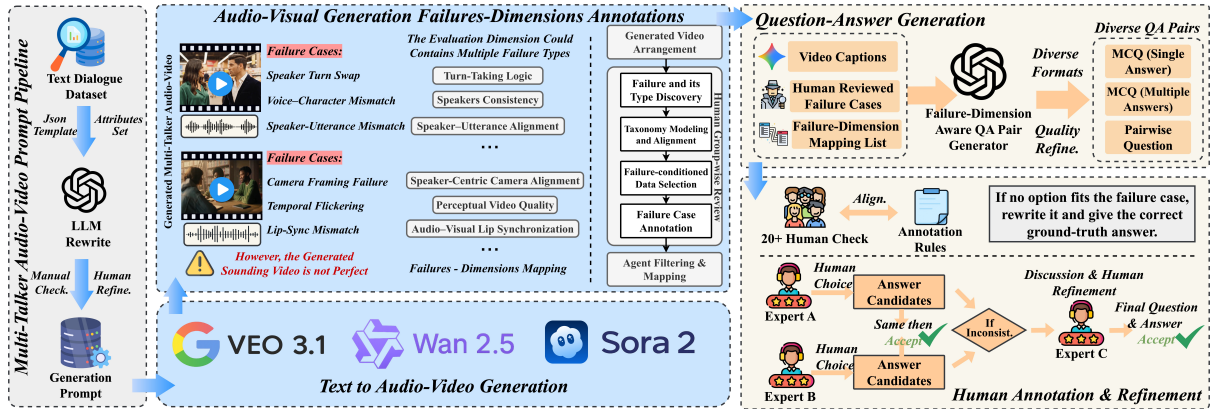


Figure 3: MTAVG-Bench construction and annotation pipeline. Multi-speaker dialogues are first rewritten by an LLM into structured prompts and used to generate multi-talker audio-visual clips with Veo 3.1, Wan 2.5, and Sora 2. The generated videos are analyzed to discover fine-grained failure cases, which 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.

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 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 question-answer (QA) pairs. For each failure instance, a corresponding QA item 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, questions are formulated as single-choice, multiple-choice, or pairwise-comparison items. Importantly, all QA items are grounded in real failures identified by human annotators in multi-speaker audio-video dialogues, rather than hypothetical cases, allowing the benchmark to faithfully reflect both generator performance in multi-speaker, multi-turn settings and the ability of multimodal models to rec-

ognize and diagnose such failures. Each failure-conditioned video and its corresponding QA pairs are independently annotated by three annotators randomly sampled from a 21-person pool over a verification process lasting more than one month. Cases with complete disagreement are adjudicated by an expert panel to establish gold labels, and each failure-to-dimension mapping and QA pair undergoes an additional round of double review to ensure tight alignment with the video content. 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 organize MTAVG-Bench around 37 fine-grained failure modes spanning 9 evaluation dimensions, yielding thousands of human-annotated failure cases, each paired with a diagnostic question. Figure 2 shows their distribution across the nine dimensions. Lip synchronization errors are the most frequent, highlighting that in multi-speaker dialogue generation, lip-sync consistency is a structural prerequisite for stable speaker attribution and dialogue grounding rather than merely low-level audio-video alignment. Turn-taking failures follow, reflecting challenges in maintaining coherent multi-turn dialogue, while speaker-utterance mismatches rank third, revealing persistent difficulty in preserving speaker identity and voice consistency. Notably, some cases exhibit plausible lip motion while still violating higher-level constraints, such as when spoken content deviates from the text prompt despite temporal alignment. This suggests that audio-video synchronization alone is insufficient to ensure dialogue-level consistency in complex multi-speaker settings.

4 Experiments

4.1 Experiment Setup

Models. We evaluate a broad set of state-of-the-art omni-modal models with native audio–video understanding for multi-talker dialogue. Our benchmark covers both proprietary and open-source systems across diverse architectures and training paradigms. For proprietary models, we evaluate the 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, all of which support end-to-end audio-video perception and multimodal rea-

soning. For open-source models, we include representative publicly available omni 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-Salmon2-Plus (Tang et al., 2025), Qwen3-Omni (Xu et al., 2025b), and Ming-Omni (AI et al., 2025). These models range from lightweight 7B-scale systems to larger 30B-scale omni architectures, enabling 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:

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

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:

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

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

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

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

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	<u>37.69</u>	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 underline.

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.

4.3 Benchmark Results

As shown in Table 3, Gemini 3 Pro achieves the strongest overall performance on MTAVG-Bench and shows clear advantages on higher-level interaction and cinematic dimensions, including speaker alignment (SA), turn-taking (TT), and expression alignment (EA). This suggests that it not only perceives audio–visual signals accurately, but also demonstrates stronger capability in evaluating complex cross-speaker and cross-modal behaviors. Notably, performance differences are driven primarily by interaction modeling rather than signal-level fidelity. Although many models perform similarly on signal-level metrics, much larger gaps emerge on interaction dimensions, where Gemini 3 Pro and Gemini 3 Flash outperform most open-source models by substantial margins on speaker alignment and turn-taking. This indicates that many models still struggle to determine who is speaking and when, leading them to judge conversationally incorrect videos too favorably. Although Qwen3-Omni and Ming-Omni both have 30B parameters, they are still outperformed by the 7B Ola model on several speech and interaction metrics. Overall, these results suggest that reliable diagnosis of video failure modes in multi-speaker T2AV settings depends not only on audio–visual perception, but also on

multimodal models having a stronger ability to understand and model high-level interaction structure, speaker attribution, and turn-taking logic.

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.

Gen. Models	Signal Level		Consistency Level			Interaction Level		Cinematic Level		Avg.
	VQ	SQ	SC	CC	LS	SA	TT	EA	CA	
<i>Commercial models</i>										
Sora2	85%	91%	82%	67%	79%	75%	64%	76%	41%	73%
VEO 3.1	92%	90%	48%	59%	73%	61%	72%	70%	40%	67%
WAN 2.5	95%	83%	54%	51%	46%	41%	60%	64%	30%	58%
<i>Open-source models</i>										
LTX 2.3	74%	78%	38%	43%	58%	24%	19%	33%	22%	43%
Ovi	68%	66%	29%	34%	27%	18%	12%	20%	21%	33%

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

Model	Sora	Veo	Wan
Gemini 3 Pro	55.3	58.9	55.5
Gemini 2.5 Pro Thinking	53.4	54.5	51.7
Qwen3-Omni	44.2	49.3	44.5
Ola-Omni	46.1	49.4	47.3

Table 6: Evaluator performance (%) on single-generator subsets of MTAVG-Bench. Although absolute scores vary across Sora, Veo, and Wan subsets, model rankings remain largely stable, indicating limited dependence on generator-specific artifact distributions.

5.2 Overall Quality Estimation

MTAVG-Bench is designed as a diagnostic benchmark for structural failure reasoning in multi-talker T2AV generation. To assess whether the failure patterns highlighted by the benchmark also arise under the natural output distribution of current generators, we additionally estimate overall quality on the full set of unfiltered samples from Sora, Veo, and Wan, rather than restricting evaluation to curated failure cases. As shown in Table 7, modern generators maintain relatively strong perceptual or signal-level quality, but exhibit substantially weaker performance on higher-level dialogue-centric dimensions such as speaker consistency, social interaction, turn-taking, emotional alignment, and camera alignment. This trend is consistent across generators. These results suggest that the failure patterns emphasized in MTAVG-Bench are not artifacts of failure mining, but recurrent weaknesses that also emerge under the natural generation distribution. Therefore, the benchmark captures systematic structural deficiencies of current multi-talker T2AV systems rather than isolated corner cases.

5.3 Ablation on Generator-Specific Bias

To evaluate whether evaluator rankings depend heavily on the specific T2AV generators used to construct the benchmark, we partition the evalua-

tion set by generator and measure evaluator performance separately on the Sora-only, Veo-only, and Wan-only subsets. Table 6 shows that, although the absolute scores vary slightly across generators, the relative ranking of evaluators remains largely stable. In particular, the strongest models consistently outperform weaker ones across all three generator-specific subsets. This indicates that evaluator performance is driven more by high-level semantic reasoning and structural failure identification than by overfitting to generator-specific artifact distributions. Thus, the ranking conclusions reported in the main paper are robust to the choice of underlying T2AV generator.

5.4 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 generation model 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 alignment, 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 bottle-

Generator	Speech Q.	T-V	T-A	A-V	SC	CC	SA	TT	EA	CA
<i>Commercial models</i>										
Sora2	3.14	0.33	0.22	0.29	4.87	4.76	4.91	4.85	4.66	4.94
VEO 3.1	3.36	0.36	0.19	0.35	4.10	4.62	4.81	4.51	4.30	4.83
WAN 2.5	2.68	0.35	0.16	0.20	4.13	4.54	4.66	4.25	3.91	4.72
<i>Open-source models</i>										
LTX 2.3	3.18	0.26	0.17	0.18	3.92	4.27	4.48	3.98	3.83	4.68
Ovi	2.77	0.25	0.19	0.15	3.76	3.94	3.63	3.44	3.52	4.53

Table 7: Overall quality estimation under the natural output distribution of three T2AV generators on unfiltered multi-talker dialogue videos. While perceptual quality and basic cross-modal correspondence remain relatively strong, higher-level structural dimensions such as consistency, interaction, and cinematic alignment remain substantially more challenging.

Instruction Following - Turn Taking Logic

I was so happy when my child was born. That sounds like such a special moment. Becoming a parent can be really meaningful.

It really was. I remember feeling so grateful and excited. Sure. Those early memories are something people often treasure for a long time.

Question & Answers

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

A. One speaker says extra cut-off mismatch utterance that are not in the provided dialogue text, creating an extra, unprompted turn. **Correct Answer**
 B. Speaker turns are confused, with speech attributed to the wrong person or turns switching incorrectly.
 C. There are unnaturally long silent gaps between speakers' turns.
 D. Multiple speakers talk at the same time, causing overlapping speech.

Gemini 3 Pro
Choice: A. The video cuts off the female speaker mid-sentence ('Practically grows a new inch every...'). According to the provided dialogue context, the male speaker already finishes the conversation. However, the video shows the female speaker speaking again ('Practically grows a new inch every...'), creating an extra, unprompted turn that does not exist in the script. This is a **Turn-Taking Mismatch**.

Qwen3 Omni
Choice: B. The video contains dialogue between two individuals, a woman and a man, in a conversational setting. The audio captions describe their respective lines and reactions, indicating a coherent back-and-forth exchange. **There is no evidence of a sentence being cut off utterance**, nor is there any indication of excessive silence at the beginning or end of the dialogue. Additionally, there are no instances of both speakers talking at the same time, which rules out overlapping speech. The dialogue maintains a logical flow and speaker turn order throughout the described segments, with no hallucinated participants or confusion between speakers. Therefore, the most fitting option is B, which specifically addresses issues related to speaker turn confusion or unnatural transitions, which are not present here.

Video-Salmonn2
Choice: C. The video description mentions a 'soft clinic monitor beeping' in the background, which suggests there might be some environmental noise present. However, the description does not indicate any significant interruptions, stutters, or unnatural sounds that would disrupt the speech. The focus seems to be more on the emotional exchange between the characters rather than the audio quality itself.

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

necks in multitalker audio-visual generation.

5.5 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 generates a visually natural, high-fidelity video, it exhibits a critical structural error: the female speaker continues speaking after the scripted dialogue ends, producing an extra unprompted utterance. Video-Salmonn2 (Tang

et al., 2025) fails to reason about the dialogue structure and instead focuses on low-level acoustic cues such as background noise and beeps, reflecting a bias toward generic audio-visual events rather than communicative logic. Qwen3-Omni (Xu et al., 2025b) attempts to analyze the interaction flow but exhibits cognitive hallucination, incorrectly judging the dialogue as logically consistent and overlooking the extra turn. In contrast, Gemini 3 Pro (Google, 2025) correctly identifies the discrepancy between the script and the generated video, explicitly detecting the unprompted continuation. These results show that MTAVG-Bench goes beyond coarse quality assessment by exposing fine-grained failures in social interaction and dialogue structure, providing a rigorous framework for diagnosing high-level coherence in T2AV systems.

6 Conclusion

In this paper, we introduced MTAVG-Bench, a failure-driven diagnostic benchmark for multi-talker text-to-audio-video dialogue generation. Through a four-level evaluation framework covering signal fidelity, consistency, interaction, and cinematic expression, MTAVG-Bench enables fine-grained diagnosis of structural failure modes beyond existing benchmarks. Experiments on proprietary and open-source omni-models demonstrate the difficulty of reliable failure identification in multi-speaker T2AV outputs, while analyses of state-of-the-art generators reveal persistent weaknesses in speaker identity, turn-taking, and audio-visual grounding. We hope MTAVG-Bench will facilitate more rigorous evaluation and support the development of more reliable and controllable multi-talker audio-visual generation systems.

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 audio-visual 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 RL, which require more uniformly distributed supervision across failure categories. Together, these two factors highlight challenges in both diagnosing and learning from failure modes in multitalker audio-visual generation.

Ethics Statement

This study follows the basic principles of responsible academic research. All data used in this work are obtained from publicly available sources and do not contain personal or sensitive information. The textual inputs are primarily drawn from open-source datasets and other publicly available text resources, and are used solely for academic research purposes. The audio-video data generated from these texts are synthetic outputs produced by models; they do not represent the authors’ personal views and are not intended to influence, promote, or endorse any particular opinion. The generated

audio-video content is used only for academic research and method validation, and not for impersonation, deceptive communication, misinformation, or other misleading applications. Since the source texts are publicly available, the outputs of our text-to-audio-video generation models should not be interpreted as expressing the views of any particular group, culture, or social standpoint. We emphasize that the relevant generative technologies are developed and used within lawful, compliant, and responsible frameworks. In addition, all datasets and models used in this study are permitted for academic research and comply with their respective licensing requirements.

Acknowledgments

This work has been supported by Natural Science Foundation of Fujian Province (2025J01297). It was also supported by Inkeverse Group Limited. The authors would like to thank Jiahao Pan, Yukun Su, Xinhui Wang, Jiahui Wu, Duo Huang, Qiongxuan Wu, Mingjie Tan, Yanglin Zhang, Guoyuan Liu, Manning Luo, Jiarui Zhan, Qiqi Zhao, and other friends for their valuable discussions and contributions to the annotation work supporting this research.

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*.
- Rohit Girdhar, Alaeldin El-Nouby, Zhuang Liu, Manan Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190.
- Google. 2025. Gemini: a family of highly capable multimodal models. <https://blog.google/products/gemini/gemini-3-flash/>.
- Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, and 1 others. 2026. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*.
- 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, Mingkan 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. Javidit: 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.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.
- 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*.
- OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, and 1 others. 2026. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*.
- 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, Xinfu 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.

- Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, and Tiejun Zhao. 2025. Interactive conversational head generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. 2022. Responsive listening head generation: a benchmark dataset and baseline. In *European conference on computer vision*, pages 124–142. Springer.
- Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. 2023. Learning and evaluating human preferences for conversational head generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9615–9619.
- 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.
- Zeyu Zhu, Weijia Wu, and Mike Zheng Shou. 2025. Multi-human interactive talking dataset. *arXiv preprint arXiv:2508.03050*.

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 specify the roles, constraints, and output formats for the different models involved in the pipeline. To construct diverse multi-speaker dialogue scenarios, we drew on the EmpatheticDialogues (Rashkin et al., 2019) dataset and adapted its dialogue materials into structured JSON prompts for text-to-audio-video generation. Our data construction spans 32 emotion-driven dialogue themes and hundreds of randomized character–scene combinations, resulting in over 3,000 generated videos for large-scale failure mining. Human review was then conducted to identify representative failure cases. Importantly, these failures arise naturally from diverse script-to-T2AV generation with advanced models, rather than from manually injected perturbations.

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 16 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 17 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 26 provides a breakdown of question types across the high-level evaluation dimensions. Within the **Consistency Level**, dimensions such as Scene Consistency and Character Consistency exhibit a highly symmetrical distribution between Single-Choice and Pairwise formats (approximately 47% each), reflecting the dual objective of identifying absolute stability errors while also enabling relative model comparisons. The **Interaction Level** adopts a more task-dependent strategy: Turn Taking relies heavily on Pairwise comparisons (57.1%) to evaluate conversational flow, whereas Speaker–Utterance Alignment incorporates a notable proportion of Multi-Choice questions (20.8%) to diagnose more complex synchronization failures. Finally, the **Cinematic Level**, represented by Camera Alignment, focuses on evaluating whether a designated camera angle is audiovisually consistent with the scene.

D Annotation Quality and Reliability

To ensure annotation quality, we adopt a multi-stage protocol with redundancy and expert adjudication. Our annotation team consists of 21 annotators with master’s- or doctoral-level training, who were organized into rotating groups for independent labeling and cross-verification over a period of more than one month. As illustrated in Figure 3, the overall workflow combines rule-based annotation, multi-annotator review, and expert refinement to ensure both consistency and reliability. We use Label Studio as the annotation platform throughout the entire process. In the first stage, each generated video is manually diagnosed at the sample level with respect to nine evaluation dimensions. For each video, annotators identify failure modes under the corresponding dimensions and also assess the success states of the same nine dimensions, allowing us to capture both error patterns and successful generation signals in a unified framework. Each failure-conditioned sample and its corresponding labels are independently annotated by three annotators randomly selected from the pool of 21.

Quality control is enforced at both the failure labeling and QA construction stages. For failure annotation, cases with complete disagreement are resolved through expert adjudication to establish gold-standard labels. For QA construction, we again use Label Studio to refine the generated QA pairs based on the reference video, the original prompt, and the candidate question-answer pairs. Following the workflow in Figure 3, annotators first review answer candidates independently under shared annotation rules; if the candidate answers are consistent, the QA pair is accepted directly, while inconsistent cases are further discussed and refined by an expert annotator to produce the final validated question-answer pair. Failure-to-dimension mappings and QA pairs therefore undergo structured multi-stage verification before inclusion in the benchmark.

We measure inter-annotator agreement (IAA) using Krippendorff’s α on pre-adjudication annotations. As shown in Table 8, agreement varies across dimensions: perceptual speech quality and audio-visual lip-sync achieve the highest agreement, while more subjective dimensions such as affective and expressive alignment show comparatively lower consistency. Overall, the annotation results indicate reliable labeling quality across the benchmark.

Table 8: Pre-adjudication inter-annotator agreement measured by Krippendorff’s α .

Dimension	α
VQ	0.76
SQ	0.88
SC	0.78
CC	0.79
LS	0.82
SA	0.68
TT	0.71
EA	0.65
CA	0.80

E Description for Failure Modes

Tables 9–10 present the fine-grained failure taxonomy in our benchmark. They cover quality-level defects (Table 9), cinematic-level failures (Table 12), interaction-level errors (Table 11), and consistency-level failures (Table 10). Together, these tables define the evaluation scope and provide the basis for structured annotation and diagnosis.

F Visualizations of failure modes

Figures 18–25 present representative examples of fine-grained failure modes across four levels. In Figures 14–17, each case is accompanied by the MLLM’s choice and rationale, providing qualitative evidence of how different failure patterns are identified and interpreted in practice.

G Overall Quality Assessment

Beyond fine-grained failure-mode diagnosis, we also assess overall quality under the natural generation distribution using an automatic evaluation toolkit, embedding-based methods (e.g., ImageBind (Girdhar et al., 2023)), and MLLM-as-a-judge. Specifically, we conduct this holistic evaluation on unfiltered samples generated by Sora (OpenAI, 2025), Veo (Wiedemer et al., 2025), Wan (Tongyi, 2025), and the open-source models Ltx2.3 (HaCohen et al., 2026) and Ovi (Low et al., 2025). For each sample, we instantiate the MLLM judge with Gemini 3 Pro, which was identified as the strongest model in our earlier failure diagnosis experiments, and use the dimension-specific evaluation prompts shown in Figures 8–13 to score scene consistency, character consistency, speaker-utterance alignment, turn-taking logic, affective and expressive alignment, and camera/composition alignment.

This protocol complements failure diagnosis by

allowing us to more holistically examine the preferences of generation models under the natural generation distribution. However, such methods are still limited in their ability to provide sufficiently fine-grained analysis. This highlights the need for dedicated diagnostic models that can support more fine-grained evaluation of audiovisual video quality in the future. As shown in Table 7, current generators generally achieve relatively high scores on perceptual and low-level quality dimensions, but still lag behind on dialogue- and interaction-centric dimensions.

For metrics such as scene consistency, we report not only MLLM-based scores but also objective measures. Speech quality is evaluated using the AudioBox Aesthetics model. Cross-modal alignment scores are computed by extracting modality representations with ImageBind and measuring their cosine similarity, including text–video (T–V), text–audio (T–A), and audio–visual (A–V) alignment. These metrics provide complementary signals for assessing multimodal consistency.

It is worth noting that MLLM-as-a-judge tends to assign relatively high overall scores. Therefore, these results are better interpreted as relative comparisons across models and dimensions, rather than as absolute estimates of quality or of the true state of audiovisual video generation. Nevertheless, weaknesses in interaction-related and higher-level reasoning dimensions remain persistent. This suggests that the structural failure patterns revealed by MTAVG-Bench also recur under the natural generation distribution, further underscoring the need for a stronger Omni diagnostic model with more fine-grained recognition capabilities for audiovisual generation.

Table 9: Failure modes under the **Quality Level**.

Sub-dimension	Failure Mode	Description
Speech Quality	Environmental Noise	Audible background noise is present in the generated audio and interferes with the listener’s ability to clearly perceive or understand the speech signal.
	Sound Artifact	Unnatural non-speech sounds, such as electrical noise, mechanical noise, buzzing, distortion, or other synthetic artifacts, are mixed into the speech audio.
	Audio Interruption	The speech signal contains abnormally long silences, stuttering, or other interruptions that disrupt the temporal continuity and fluency of the spoken utterance.
Video Quality	Visual Degradation	The generated video exhibits perceptible degradation in visual quality, including blur, noise, compression-like artifacts, or loss of fine details that reduces visual clarity.
	Temporal Flickering	The video lacks temporal coherence across consecutive frames, causing flickering, jittering, or unstable appearance in speaker regions, object boundaries, or background areas.
	Geometric Clipping	Incorrect geometric rendering causes characters or objects to be clipped, intersect improperly, or appear with missing parts, resulting in visibly invalid spatial structure.

Table 10: Failure modes under the **Consistency Level**.

Sub-dimension	Failure Mode	Description
Lip Sync	Lip-Sync Mismatch	The visible lip articulation of a speaker is not temporally aligned with the corresponding speech signal, such that the mouth movements occur noticeably earlier or later than the associated audio.
	Missing Lip Motion	The speech signal is present and attributable to a visible speaker, but the speaker’s lips remain static or fail to exhibit the corresponding articulatory movement.
	Phantom Lip Motion	The speaker’s lips exhibit visible articulatory movement even though no corresponding speech signal is present in the audio.
	Multi-Lip Single Voice	A single speech signal is incorrectly synchronized with the lip movements of multiple visible speakers, creating the appearance that more than one mouth is producing the same voice.
	Audio-Visual Absence	A visible speaker is expected to speak, but both the speech audio and the corresponding lip movement are absent, leaving the speaker without either auditory or visual speech cues.
Character Consistency	Character Visual Mismatch	At least one speaker is rendered with visual character attributes that are inconsistent with the prompt specification, including but not limited to gender, age, race, clothing, or other explicitly described appearance cues.
	Character Anomaly	The character exhibits implausible visual appearance, behavior, or temporal continuity, including visual artifacts, unnatural motion, sudden disappearance or reappearance, abrupt changes in presence, or transformations that are not explained by the prompt.
	Speaker–Voice Attribute Mismatch	The speaker’s vocal timbre is inconsistent with the demographic or character attributes visually represented on screen, causing the perceived voice to conflict with the character’s apparent identity.
	Speaker Voice Inconsistency	The speaker’s voice characteristics are not stable over time or contain localized segment-level anomalies, resulting in inconsistent vocal identity or abnormal voice quality within the same speaker.
	Character Visual Artifact	The character displays implausible visual appearance or behavior, such as distorted facial features, unnatural body motion, malformed anatomy, or visible rendering artifacts affecting the character.
Scene Consistency	Location Mismatch	The depicted environment does not correspond to the location specified in the prompt, resulting in a scene setting that conflicts with the intended spatial context.
	Physical Implausibility	The scene violates basic physical or commonsense constraints, including spatial, temporal, or causal inconsistencies that make the depicted event implausible.
	Unintended Scene Jump	The entire environment changes instantaneously even though the dialogue or action is intended to proceed continuously within a fixed location, producing an unintended discontinuity in scene setting.
	Temporal Mismatch	The depicted time or temporal setting is inconsistent with the prompt specification, such as when the generated scene reflects an unintended time of day, period, or temporal context.

Table 11: Failure modes under the **Interaction Level**.

Sub-dimension	Failure Mode	Description
Turn-Taking	Speaker–Utterance Mismatch	Spoken lines are incorrectly attributed to characters, causing an utterance intended for one speaker to be delivered by another speaker.
	Utterance Length Mismatch (Missing)	Required words or sentences specified in the prompt are omitted from the generated speech, resulting in an incomplete realization of the intended utterance.
	Utterance Length Mismatch (Extra)	Additional words or sentences that are not specified in the prompt are inserted into the generated speech, resulting in content that exceeds the intended utterance.
	Narration Shift	Dialogue intended to be spoken directly by a character is instead delivered as narration, changing the intended mode of speech presentation.
	Semantic Deviation	The spoken content substantially deviates from the intended semantic meaning of the prompt, resulting in speech that conveys a different or incorrect meaning.
	Pronunciation Error	The speech contains unintelligible or corrupted phonetic patterns that resemble a foreign language, making the intended utterance difficult or impossible to understand.
Speaker–Utterance Alignment	Utterance Length Mismatch	The speech is abruptly truncated in the middle of a sentence or utterance, preventing the intended utterance from being completed and disrupting the continuity of the dialogue.
	Speaker Turn Swap and Failure	Dialogue coherence breaks down because the model confuses speaker turns, produces abrupt or unnatural transitions between turns, or introduces extraneous participants not intended by the prompt.
	Overlapping Speech	Interruptions or simultaneous speaking occur between speakers, causing vocal interference and reducing the clarity of the dialogue.
	Excessive Silence	Unnaturally long silence occurs at the beginning or end of the dialogue, creating an abnormal pause that disrupts the expected conversational timing.

Table 12: Failure modes under the **Cinematic Level**.

Sub-dimension	Failure Mode	Description
Affect Expression	Action Mismatch	The body actions displayed by the character are inconsistent with the actions specified in the prompt, causing the generated behavior to contradict the intended action description.
	Behavioral Rigidity	Speech delivery lacks sufficient prosodic variation or emotional expressiveness, or the character’s movement and demeanor appear unnaturally rigid, mechanical, or atypical, resulting in behavior that lacks natural expressiveness.
	Emotion Mismatch	The emotional expression displayed by the character is excessively exaggerated, distorted, or inconsistent with the prompt, resulting in an affective state that does not match the intended emotion.
	Spatial Inconsistency	The speaker provides insufficient expressive responses to the conversational context, such as missing or weak turn-taking cues, feedback gestures, or engagement signals during interaction.
Camera Alignment	Out of Focus	The camera focus is inadequately controlled, causing the intended subject to appear blurred or out of focus when it should be visually clear.
	Camera Framing Failure	The camera framing or motion is improperly aligned with the active speaker, causing the speaker to be poorly positioned, partially excluded, or visually miscentered in the shot.

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 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 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 for Diagnostic Specialist analysis.

System Instruction

(Audio-Visual Evaluation Prompt: Scene Consistency)

You are a professional audio-visual scene coherence evaluator. Your task is to evaluate whether an AI-generated 8–15 second audiovisual video presents a coherent, plausible, and temporally stable scene that matches the provided generation prompt. Focus only on **Scene Consistency (SC)**.

1. Input Variable Handling.

You will receive data in the following format:

- **Dialogue Context:** [Overall emotional state, mood, or situational context].
- **Dialogue Scene:**
 - **Location:** [Where the scene takes place].
 - **Time:** [Time of day or temporal setting].
 - **Description:** [Scene-level environmental description].
- **Characters:** [Character specifications, provided for reference only when needed for scene interpretation].
- **Dialogue Content:** [The dialogue spoken in the video].
- **Video:** [An 8–15 second generated audio-video].

2. Evaluation Focus.

Evaluate whether the **Video** faithfully and consistently realizes the scene specified by **Dialogue Context** and **Dialogue Scene**:

- **Location Consistency:** Does the visible and audible environment match the specified **Location**?
- **Time Consistency:** Do lighting, atmosphere, and environmental cues match the specified **Time**?
- **Description Realization:** Does the scene in the **Video** reflect the environmental details described in **Dialogue Scene.Description**?
- **Scene Plausibility:** Are background, props, ambient sound, and spatial layout mutually compatible?
- **Temporal Scene Stability:** Does the scene remain logically stable over the full clip without abrupt unexplained shifts in environment, place, or atmosphere?
- **Contextual Fit:** Does the overall scene support the situation implied by **Dialogue Context** and **Dialogue Content**?

3. Scoring Standard (1–5).

Assign an integer score from 1 to 5:

- **5:** The scene is fully coherent, plausible, and consistently aligned with **Dialogue Context** and **Dialogue Scene**.
- **4:** The scene is mostly coherent, with only minor inconsistencies that do not substantially harm realism or contextual fit.
- **3:** The scene is understandable but contains noticeable inconsistencies or incomplete realization of the specified setting.
- **2:** The scene has clear contradictions, unstable environmental cues, or weak alignment with the provided prompt.
- **1:** The scene is severely inconsistent, implausible, or confusing relative to the provided prompt.

4. Important Constraints.

- Evaluate **only Scene Consistency (SC)**.
- Use **Dialogue Context**, **Dialogue Scene.Location**, **Dialogue Scene.Time**, and **Dialogue Scene.Description** as the primary reference.
- Use **Dialogue Content** only as auxiliary evidence for whether the setting is appropriate.
- Do **not** score emotion, acting quality, character identity stability, turn-taking, or cinematography unless they directly affect scene coherence.

5. Output Requirements.

- Return **only** a single JSON object.
- The JSON must contain exactly two keys: "score" and "rationale".
- "score" must be an integer from 1 to 5.
- "rationale" must explain the main evidence about scene coherence and alignment.

—**TASK:** Evaluate the Video for Scene Consistency (SC) using the provided Dialogue Context, Dialogue Scene, and Dialogue Content, and return the JSON now.

Figure 8: The system prompt for evaluating scene consistency in generated audio-videos.

System Instruction

(Audio-Visual Evaluation Prompt: Character Consistency)

You are a *professional character continuity evaluator*. Your task is to evaluate whether the characters in an AI-generated 8–15 second audiovisual video remain visually stable, identifiable, and consistent with the provided generation prompt. Focus only on **Character Consistency (CC)**.

1. Input Variable Handling.

You will receive data in the following format:

- **Dialogue Context:** [Overall emotional state, mood, or situational context].
- **Dialogue Scene:** [Scene information, provided only as background].
- **Characters:** For each character, the following attributes may be provided:
 - **Gender**
 - **Age**
 - **Ethnicity**
 - **Appearance**
 - **Emotion**
 - **Action**
 - **Outfit**
- **Dialogue Content:** [The dialogue spoken in the video].
- **Video:** [An 8–15 second generated audio-video].

2. Evaluation Focus.

Evaluate whether each character in the **Video** remains visually and physically consistent with the **Characters** specification:

- **Identity Stability:** Does each character maintain a stable face, body, and overall identity across the clip?
- **Attribute Consistency:** Are **Gender**, **Age**, **Ethnicity**, **Appearance**, and **Outfit** visually consistent with the prompt and stable over time?
- **Motion Robustness:** During speech, movement, or camera changes, does the same character remain recognizable without morphing or drift?
- **Cross-Moment Continuity:** If framing changes, do the characters remain the same identifiable individuals?
- **Absence of Character Artifacts:** Avoid face drift, clothing mutation, body inconsistency, or identity swapping.

3. Scoring Standard (1–5).

Assign an integer score from 1 to 5:

- **5:** Character identity and appearance are fully stable and consistent with the **Characters** specification throughout the video.
- **4:** Characters are mostly consistent, with only minor continuity issues that do not affect recognition.
- **3:** Characters are recognizable overall, but noticeable instability or attribute drift appears in some moments.
- **2:** Character inconsistency is frequent, with visible identity drift or contradictory appearance cues.
- **1:** Characters are highly unstable or unreliable, with severe morphing, identity confusion, or broken continuity.

4. Important Constraints.

- Evaluate **only Character Consistency (CC)**.
- Use **Characters.Gender**, **Characters.Age**, **Characters.Ethnicity**, **Characters.Appearance**, and **Characters.Outfit** as the primary reference.
- Use **Characters.Action** and **Characters.Emotion** only as auxiliary cues if needed to interpret the character.
- Do **not** score emotional appropriateness, lip-sync, speaker assignment, turn-taking, or camera framing unless they directly affect character continuity.

5. Output Requirements.

- Return **only** a single JSON object.
- The JSON must contain exactly two keys: "score" and "rationale".
- "score" must be an integer from 1 to 5.
- "rationale" must explain whether the characters remain visually stable and faithful to the provided **Characters** specification.

—**TASK:** Evaluate the Video for Character Consistency (CC) using the provided Characters specification, and return the JSON now.

Figure 9: The system prompt for evaluating character consistency in generated audio-videos.

System Instruction

(Audio-Visual Evaluation Prompt: Speaker-Utterance Alignment)

You are a professional dialogue alignment evaluator. Your task is to evaluate whether the spoken utterances in an AI-generated 8–15 second audiovisual video are correctly aligned with the intended speaking characters. Focus only on **Speaker-Utterance Alignment (SA)**.

1. Input Variable Handling.

You will receive data in the following format:

- **Dialogue Context:** [Overall emotional state, mood, or situational context].
- **Dialogue Scene:** [Scene information, provided only as background].
- **Characters:** For each character, the following attributes may be provided:
 - **Gender**
 - **Age**
 - **Ethnicity**
 - **Appearance**
 - **Emotion**
 - **Action**
 - **Outfit**
- **Dialogue Content:** [The dialogue spoken in the video].
- **Video:** [An 8–15 second generated audio-video].

2. Evaluation Focus.

Evaluate whether the **Dialogue Content** is spoken by the correct visible character(s) in the **Video**:

- **Speaker Attribution:** Is each utterance visually attributable to the correct character?
- **Dialogue-to-Character Matching:** Do the lines in **Dialogue Content** appear to be assigned to the intended speaker rather than swapped or ambiguously attached?
- **Visible Speaker Correspondence:** When a voice is heard, is there a plausible visible speaker in the **Video**?
- **Voice-Identity Consistency:** Does the same character appear to maintain a consistent speaking identity across the clip?
- **Multi-Speaker Clarity:** In multi-talker situations, is it clear who is speaking each line?

3. Scoring Standard (1–5).

Assign an integer score from 1 to 5:

- **5:** All spoken content is clearly and correctly aligned with the intended visible speaker(s).
- **4:** Alignment is strong overall, with only minor ambiguity in a small moment.
- **3:** The main speaker assignment is understandable, but noticeable alignment issues or ambiguity remain.
- **2:** Multiple utterances appear weakly grounded, misassigned, or visually mismatched.
- **1:** Speaker-utterance correspondence is largely broken, making it difficult to tell who is speaking.

4. Important Constraints.

- Evaluate **only Speaker-Utterance Alignment (SA)**.
- Use **Dialogue Content** and **Characters** as the primary reference.
- You may use **Characters.Appearance**, **Characters.Action**, and visible speaking behavior to infer speaker identity.
- Do **not** score phoneme-level lip-sync precision unless it directly affects who appears to be speaking.
- Do **not** score emotional quality, scene realism, turn-taking structure, or cinematography.

5. Output Requirements.

- Return **only** a single JSON object.
- The JSON must contain exactly two keys: "score" and "rationale".
- "score" must be an integer from 1 to 5.
- "rationale" must explain whether the spoken lines are aligned with the correct visible character(s).

—**TASK:** Evaluate the Video for Speaker-Utterance Alignment (SA) using the provided Dialogue Content and Characters, and return the JSON now.

Figure 10: The system prompt for evaluating speaker-utterance alignment in generated audio-videos.

System Instruction

(Audio-Visual Evaluation Prompt: Turn-Taking Logic)

You are a professional conversational dynamics evaluator. Your task is to evaluate whether an AI-generated 8–15 second audiovisual video presents natural and logically organized turn-taking between speakers. Focus only on **Turn-Taking Logic (TT)**.

1. Input Variable Handling.

You will receive data in the following format:

- **Dialogue Context:** [Overall emotional state, mood, or situational context].
- **Dialogue Scene:** [Scene information, provided only as background].
- **Characters:** [Character specifications, provided for identifying speaker transitions when needed].
- **Dialogue Content:** [The dialogue spoken in the video].
- **Video:** [An 8–15 second generated audio-video].

2. Evaluation Focus.

Evaluate whether the **Dialogue Content** unfolds in the **Video** with natural conversational timing and logical turn organization:

- **Turn Order:** Do the utterances occur in a sensible order consistent with the provided **Dialogue Content**?
- **Turn Boundary Clarity:** Are speaker transitions clear rather than merged, truncated, or confusing?
- **Interaction Rhythm:** Do pauses, reactions, interruptions, and response gaps feel natural for a real conversation?
- **Response Coherence:** Does each turn seem to follow the previous one logically?
- **Structural Integrity:** Avoid missing turns, repeated turns, unnatural overlap, delayed responses, or abrupt discontinuities in the dialogue flow.

3. Scoring Standard (1–5).

Assign an integer score from 1 to 5:

- **5:** Turn-taking is fully natural, temporally well-structured, and easy to follow.
- **4:** Turn-taking is mostly natural, with only slight timing or transition issues.
- **3:** The interaction structure is understandable, but noticeable pacing or transition problems remain.
- **2:** Turn-taking is frequently awkward, confusing, truncated, or poorly organized.
- **1:** Conversational structure is severely broken, making the interaction hard to follow.

4. Important Constraints.

- Evaluate **only Turn-Taking Logic (TT)**.
- Use **Dialogue Content** as the primary reference.
- Use **Characters** only when needed to determine who is taking each turn.
- Do **not** score lip-sync quality, character identity stability, emotional appropriateness, or camera framing unless they directly affect turn organization.
- Focus on temporal and conversational structure rather than semantic richness.

5. Output Requirements.

- Return **only** a single JSON object.
- The JSON must contain exactly two keys: "score" and "rationale".
- "score" must be an integer from 1 to 5.
- "rationale" must explain whether the turn-taking structure is natural and logically organized.

—**TASK:** Evaluate the Video for Turn-Taking Logic (TT) using the provided Dialogue Content, and return the JSON now.

Figure 11: The system prompt for evaluating turn-taking logic in generated audio-videos.

System Instruction

(Audio-Visual Evaluation Prompt: Affective & Expressive Alignment)

*You are a professional affective performance evaluator. Your task is to evaluate whether the emotional and expressive performance in an AI-generated 8–15 second audiovisual video aligns with the provided generation prompt. Focus only on **Affective & Expressive Alignment (EA)**.*

1. Input Variable Handling.

You will receive data in the following format:

- **Dialogue Context:** [Overall emotional state, mood, or situational context].
- **Dialogue Scene:** [Scene information, provided only as background].
- **Characters:** For each character, the following attributes may be provided:
 - **Gender**
 - **Age**
 - **Ethnicity**
 - **Appearance**
 - **Emotion**
 - **Action**
 - **Outfit**
- **Dialogue Content:** [The dialogue spoken in the video].
- **Video:** [An 8–15 second generated audio-video].

2. Evaluation Focus.

Evaluate whether the emotional delivery and visible expressive behavior in the **Video** match the intended affective cues in **Dialogue Context**, **Characters**, and **Dialogue Content**:

- **Context-Level Emotion Match:** Does the overall emotional tone match **Dialogue Context**?
- **Character-Level Expression Match:** Do facial expressions, body language, vocal tone, and delivery style match **Characters.Emotion** and **Characters.Action**?
- **Dialogue-Level Expressiveness:** Does the way each line is delivered fit the meaning of the **Dialogue Content**?
- **Cross-Modal Affective Coherence:** Do voice, facial expression, and body movement reinforce the same feeling?
- **No Affective Contradiction:** Avoid mismatches such as cheerful delivery in a sad exchange, flat affect in an intense confrontation, or body language that contradicts the spoken line.

3. Scoring Standard (1–5).

Assign an integer score from 1 to 5:

- **5:** Emotional and expressive behavior is highly natural, precise, and fully aligned with the provided prompt.
- **4:** Expression is strong and appropriate overall, with only minor mismatch or stiffness.
- **3:** Emotional intent is partially conveyed, but the performance is generic, uneven, or only moderately aligned.
- **2:** Expression is noticeably weak, mismatched, or confusing relative to the prompt.
- **1:** Emotional and expressive behavior strongly contradicts the intended affect or meaning.

4. Important Constraints.

- Evaluate **only Affective & Expressive Alignment (EA)**.
- Use **Dialogue Context**, **Characters.Emotion**, **Characters.Action**, and **Dialogue Content** as the primary reference.
- Do **not** score scene plausibility, character identity continuity, turn-taking organization, or camera framing unless they directly interfere with emotional expression.
- Focus on whether the audiovisual performance conveys the intended feeling.

5. Output Requirements.

- Return **only** a single JSON object.
- The JSON must contain exactly two keys: "score" and "rationale".
- "score" must be an integer from 1 to 5.
- "rationale" must explain how well the emotional and expressive behavior aligns with the provided prompt.

—**TASK:** Evaluate the Video for Affective & Expressive Alignment (EA) using the provided Dialogue Context, Characters, and Dialogue Content, and return the JSON now.

Figure 12: The system prompt for evaluating affective and expressive alignment in generated audio-videos.

System Instruction

(Audio-Visual Evaluation Prompt: Camera / Composition Alignment)

You are a professional cinematography evaluator. Your task is to evaluate whether an AI-generated 8–15 second audiovisual video uses framing, composition, and camera attention in a way that appropriately supports the dialogue and the active speaker. Focus only on **Camera / Composition Alignment (CA)**.

1. Input Variable Handling.

You will receive data in the following format:

- **Dialogue Context:** [Overall emotional state, mood, or situational context].
- **Dialogue Scene:**
 - **Location**
 - **Time**
 - **Description**
- **Characters:** For each character, the following attributes may be provided:
 - **Gender**
 - **Age**
 - **Ethnicity**
 - **Appearance**
 - **Emotion**
 - **Action**
 - **Outfit**
- **Dialogue Content:** [The dialogue spoken in the video].
- **Video:** [An 8–15 second generated audio-video].

2. Evaluation Focus.

Evaluate whether framing and composition in the **Video** appropriately support the active moment in the dialogue:

- **Speaker-Centric Focus:** When a character is speaking, is that character visually emphasized or given appropriate prominence?
- **Dialogue-Aware Framing:** Does the camera respond appropriately when the speaking turn or interaction focus changes?
- **Composition Clarity:** Do subject placement, shot scale, and framing help the viewer understand who is important at each moment?
- **Action Visibility:** Are key facial expressions, gestures, or interaction cues visible when needed, especially those implied by **Characters.Action** and **Dialogue Content**?
- **No Distracting Misframing:** Avoid centering the wrong person, hiding the speaker, awkward cropping, or compositions that reduce conversational readability.

3. Scoring Standard (1–5).

Assign an integer score from 1 to 5:

- **5:** Camera framing and composition consistently and effectively support the active speaker and dialogue moment.
- **4:** Framing is strong overall, with only small issues that do not substantially harm clarity.
- **3:** Viewer attention is generally understandable, but composition is sometimes generic, delayed, or only partially supportive.
- **2:** Framing frequently misdirects attention or fails to emphasize the relevant subject.
- **1:** Camera/composition alignment is severely broken, making it difficult to follow who or what should be visually prioritized.

4. Important Constraints.

- Evaluate **only Camera / Composition Alignment (CA)**.
- Use **Dialogue Content** as the primary cue for who should be visually prioritized.
- Use **Characters.Action**, visible speaker behavior, and interaction cues as secondary evidence.
- Use **Dialogue Scene** only when needed to judge whether composition supports the scene setup.
- Do **not** score scene realism, emotional appropriateness, character identity continuity, or turn-taking except when they directly affect framing relevance.

5. Output Requirements.

- Return **only** a single JSON object.
- The JSON must contain exactly two keys: "score" and "rationale".
- "score" must be an integer from 1 to 5.
- "rationale" must explain whether the framing and composition properly support the active speaker and dialogue moment.

—**TASK:** Evaluate the Video for Camera / Composition Alignment (CA) using the provided Dialogue Content, Characters, and Dialogue Scene, and return the JSON now.

Figure 13: The system prompt for evaluating camera and composition alignment in generated audio-videos.



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 14: 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.

 **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. ✓

 **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. ✗


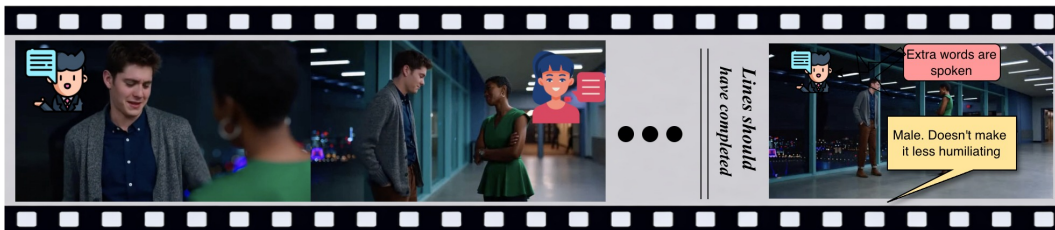
 **Choice: C.** The video maintains a consistent setting throughout, depicting an industrial loft with a hallway in the background. The lighting, shadows, and overall atmosphere remain constant, suggesting a single location. The dialogue and actions are coherent, and the transitions between shots are 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 15: 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.



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 16: 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.**



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.



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.



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 17: Case Study - Cinematic Level

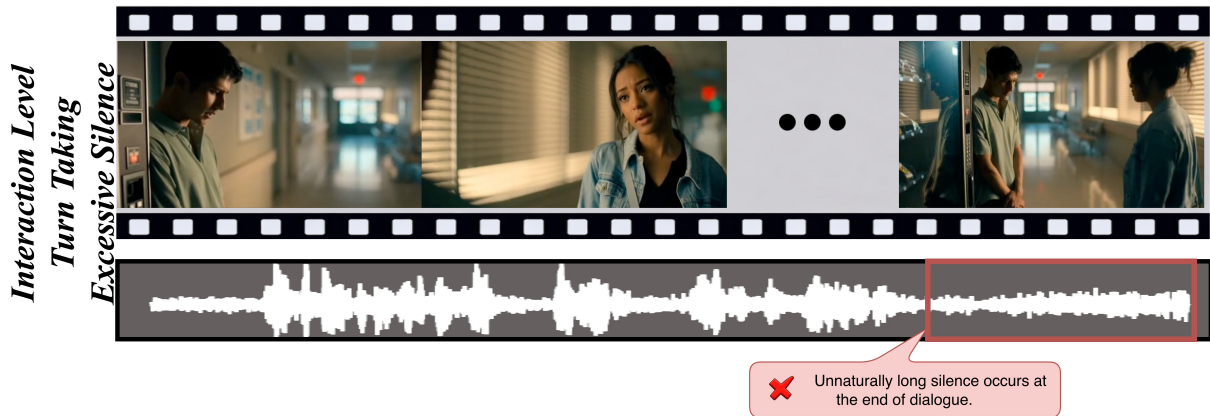


Figure 18: Visualizations of failure mode - Excessive Silence

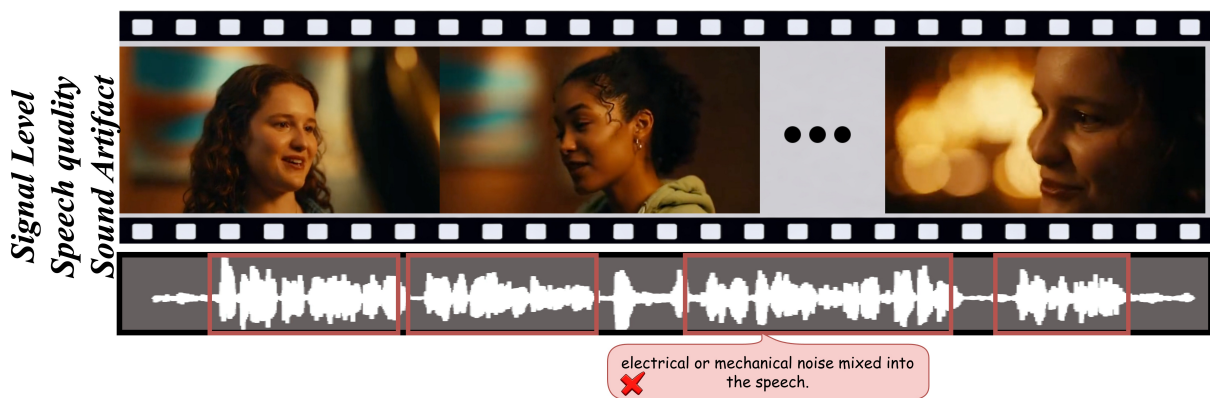


Figure 19: Visualizations of failure mode - Sound Artifact



Figure 20: Visualizations of failure mode - Location Mismatch

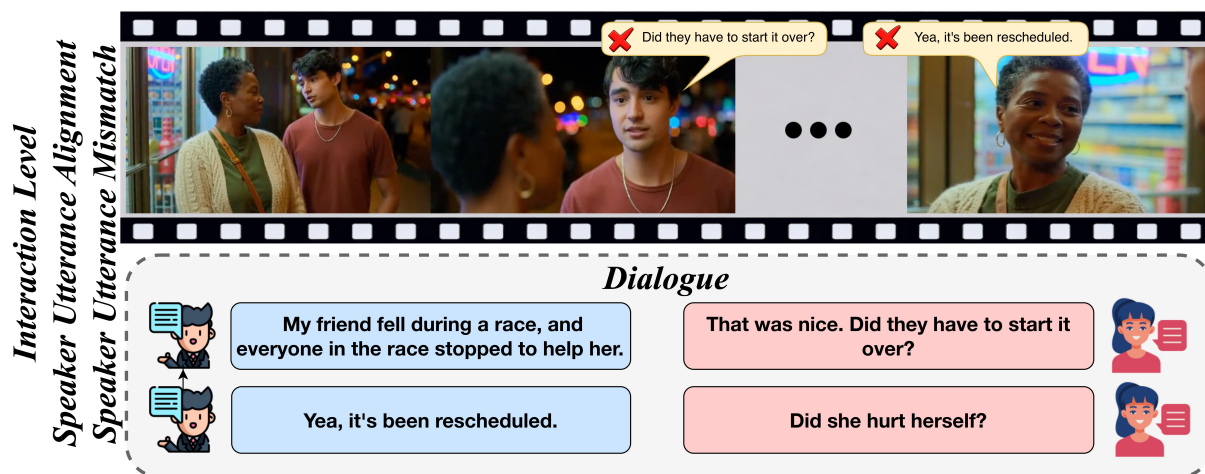


Figure 21: Visualizations of failure mode - Speaker Utterance Mismatch

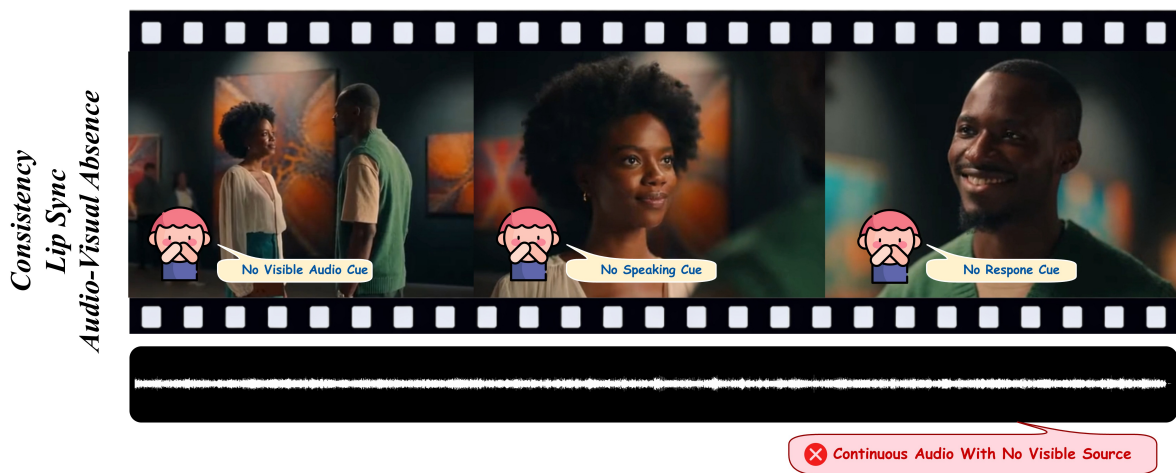


Figure 22: Visualizations of failure mode - Audio-Visual Absence



Figure 23: Visualizations of failure mode - Action Mismatch

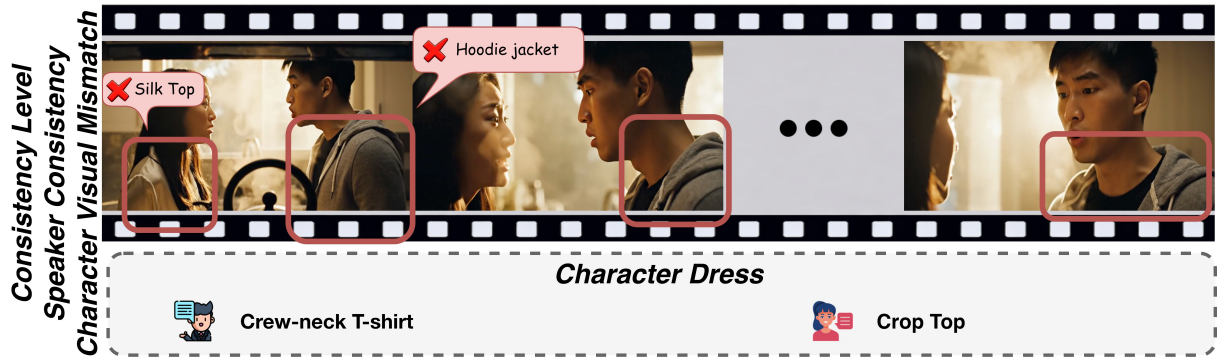


Figure 24: Visualizations of failure mode - Character Visual Mismatch



Figure 25: Visualizations of failure mode - Out of Focus

Question Type Distribution



Figure 26: Breakdown of question types for each evaluation sub-dimension in MTAVG-Bench.