MSAVBENCH: TOWARDS COMPREHENSIVE AND RELIABLE EVALUATION OF MULTI-SHOT AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation is rapidly evolving from single-shot synthesis to complex multi-shot audio-video (MSAV) narratives to meet real-world demands. However, evaluating such frontier models remains a fundamental challenge. Existing benchmarks are limited in scope and data diversity, and rely on rigid evaluation pipelines, preventing systematic and reliable assessment of modern MSAV models. To bridge these gaps, we introduce **MSAVBench**, the first comprehensive benchmark and adaptive hybrid evaluation framework for multi-shot audio-video generation. Our benchmark spans four key dimensions, video, audio, shot, and reference, covering diverse task settings, varying shot counts of up to 15, and challenging non-realistic scenarios. Our evaluation framework improves robustness through an adaptive self-correction mechanism for shot segmentation, instance-wise rubrics for subjective metrics, and tool-grounded evidence extraction for complex judgments. Furthermore, MSAVBench achieves high alignment with human judgments, reaching a Spearman rank correlation of 91.5%. Our systematic evaluation of 20 state-of-the-art closed- and open-source models shows that the evaluated systems still struggle with director-level control and fine-grained audio-visual synchronization, while modular or agentic generation pipelines offer a promising path toward narrowing the gap between open- and closed-source models. The benchmark data and evaluation code will be made publicly available.

1 INTRODUCTION

The landscape of generative video is shifting from *silent, single-shot* text-to-video (T2V) synthesis (Brooks et al., 2024; Kong et al., 2024; HaCohen et al., 2024) toward *multi-shot audio-video* (MSAV) generation (Seedance et al., 2026; Wan Team, 2026; OpenAI, 2025). Unlike traditional short clips, MSAV enables cinematic storytelling with complex narratives and synchronized audio. While frontier closed-source systems (*e.g.*, Seedance 2.0 (Seedance et al., 2026), Wan 2.7 (Wan Team, 2026), Sora 2 (OpenAI, 2025)) have led MSAV generation, exploration and practical adoption in the open-source community remain comparatively limited (MiniMax, 2026). Therefore, establishing a comprehensive MSAV benchmark is an urgent prerequisite to providing design guidelines for the open-source community and to diagnosing model weaknesses in closed-source systems.

However, evaluating MSAV generation is fundamentally challenging due to its compositional, multi-shot, and multi-modal nature. Specifically, existing benchmarks only address isolated facets of this problem, falling short on two concrete fronts: **(i) Limited evaluation scope and data diversity.** Most prior benchmarks (Huang et al., 2024; Han et al., 2025) target *single-shot*, silent generation. Recent efforts only partially bridge this gap: they focus either on single-shot audio-video generation (Zhou et al., 2026b), or on multi-shot video synthesis but lack thorough audio evaluation (Shi et al., 2026). Furthermore, their evaluation datasets exhibit limited diversity and complexity, overlooking the rich cinematic language and challenging scenarios like counterfactual content. Consequently, these benchmarks fail to systematically assess the diverse task adaptability and performance of modern MSAV models in complex scenarios. **(ii) Rigid and static evaluation pipelines.** First, they struggle with *limited robustness to shot mis-segmentation*. Generated videos often exhibit variable shot counts and ambiguous transition boundaries, making shot-based evaluation highly sensitive to segmentation errors. Existing pipelines typically rely on fixed segmenters without self-correction, so a single

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

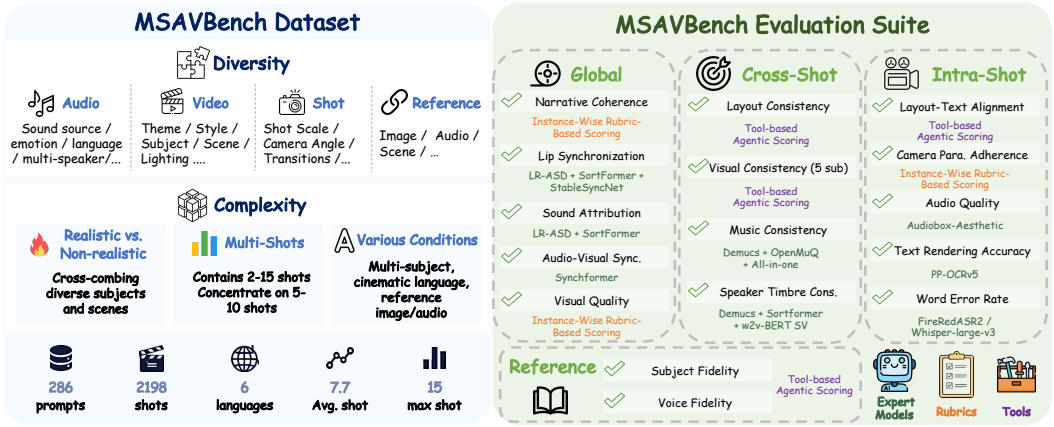


Figure 1: **Overview of MSAVBench.** Left: the benchmark spans four data dimensions, namely video, audio, shot, and reference, covering diverse prompts, shot counts, and realistic and non-realistic scenarios. Right: the evaluation suite assesses generated MSAV content at four levels, including global, cross-shot, intra-shot, and reference levels, using a hybrid strategy that combines specialized expert models, rubric-based scoring, and tool-grounded assessment.

mis-segmentation can distort downstream metrics. Second, they employ *rigid scoring paradigms for complex dimensions*. For important yet challenging dimensions without dedicated expert models (*e.g.*, narrative coherence and layout–text consistency), existing pipelines often rely on direct VLM scoring. Although simple to implement, this strategy is sensitive to prompt phrasing and prone to hallucination, making it unreliable for assessing performance on complex tasks.

To bridge these gaps, we present **MSAVBench**, a comprehensive benchmark and adaptive hybrid evaluation framework for MSAV generation, as shown in Figure 1. First, our benchmark is designed for broad and challenging coverage. It spans four key dimensions: *video*, *audio*, *shot*, and *reference*, each with diverse sub-dimensions, and includes a wide range of generation settings, such as varying shot counts (up to 15), different numbers of subjects, and non-realistic scenarios. Second, the evaluation framework is designed for robustness and reliability. We introduce a self-correction mechanism that enables a VLM to iteratively inspect shot boundaries and invoke tools to merge or split segments, thereby mitigating error propagation from shot mis-segmentation. For subjective dimensions such as narrative coherence, we replace direct VLM scoring with instance-wise rubrics formulated as predefined multiple-choice questions. For complex dimensions such as layout–text consistency, we allow the model to adaptively invoke external perception tools to gather objective evidence for the final judgment. Together, MSAVBench enables a more comprehensive and reliable assessment of modern MSAV models, revealing their multifaceted capabilities and limitations while achieving high alignment with human judgments, reflected by a Spearman rank correlation of 91.5%.

Leveraging MSAVBench, we conduct a comprehensive evaluation of 20 state-of-the-art closed- and open-source models. Our analysis reveals three key insights into the current MSAV landscape: *(i)* a substantial performance gap persists among the evaluated closed- and open-source systems, but modular or agentic generation pipelines show promise for narrowing this gap; *(ii)* current models remain far from reliable "director-level" generation, struggling with cinematic control, structural consistency, and fine-grained joint audio-visual alignment; and *(iii)* the "video-first, post-hoc dubbing" paradigm is insufficient for complex multi-shot audio-video generation, highlighting the need for unified audio-video architectures.

In summary, our contributions are threefold. *First*, we release **MSAVBench**, the first benchmark for multi-shot audio-video generation, covering four key dimensions: video, audio, shot, and reference, as well as diverse tasks and challenging generation settings. *Second*, we propose an adaptive hybrid evaluation framework that improves robustness through dynamic shot-boundary correction, instance-wise rubrics, and tool-grounded evidence extraction. *Third*, we systematically evaluate 20 state-of-the-art closed- and open-source models, showing that modular and agentic generation pipelines are a promising path for open-source systems, while highlighting challenges in director-level control and audio-visual synchronization as well as the need for unified audio-video architectures.

Table 1: **Comparison with existing video and audio–video generation benchmarks. Counterf.**: counterfactual prompts; **Cine.**: cinematic language and camera control; **Ref.**: reference conditioning. MSAVBench offers comprehensive coverage of data dimensions and challenging cases, along with a robust evaluation framework featuring adaptive self-correction and agentic scoring.

Benchmark	Avg. Shots	Counterf.	Video	Audio	Audio-Video	Cine.	Ref.	Shot Correction	Agentic Scoring	# Metrics	# Prompts
VBench (Huang et al., 2024)	1	✗	✓	✗	✗	✗	✗	✗	✗	16	∼1,600
EvalCrafter (Liu et al., 2024b)	1	✗	✓	✗	✗	✗	✗	✗	✗	17	700
Video-Bench (Han et al., 2025)	1	✗	✓	✗	✗	✗	✗	✗	✗	9	419
OpenS2V-Nexus (Yuan et al., 2026a)	1	✗	✓	✗	✗	✗	✓	✗	✗	6	180
ViStoryBench (Zhuang et al., 2026)	16.5	✓	✓	✗	✗	✓	✓	✗	✗	12	80
MSVBench (Shi et al., 2026)	∼14	✓	✓	✗	✗	✓	✓	✗	✗	20	20
UniVBench (Wei et al., 2026a)	3.72	✓	✓	✗	✗	✓	✓	✗	✓	21	200
AVGen-Bench (Zhou et al., 2026b)	1.6	✗	✓	✓	✓	✗	✗	✗	✗	10	235
MSAVBench (Ours)	**7.7**	✓	✓	✓	✓	✓	✓	✓	✓	**20**	**286**

2 Related Work

Audio-video generation models. While earlier models primarily targeted single-shot video generation (Brooks et al., 2024; Kong et al., 2024), recent closed-source and open-weight models support multi-shot audio-video synthesis (OpenAI, 2025; Seedance et al., 2026; Google DeepMind, 2026; MiniMax, 2026). However, MSAV evaluation remains underexplored and challenging, requiring assessment of both long-range cross-shot coherence and fine-grained audio-visual alignment. Additional related work is discussed in Appendix A.

Audio-video evaluation benchmarks. Early benchmarks such as VBench (Huang et al., 2024) and Video-Bench (Han et al., 2025) primarily assess single-shot visual quality. Multi-shot benchmarks (Zhuang et al., 2026; Shi et al., 2026) extend evaluation to narrative structure and cross-shot consistency but provide limited audio assessment. Audio-video benchmarks (Zhou et al., 2026b; Xie et al., 2025; Cao et al., 2025) assess audio-visual alignment but mostly use single-shot or weakly structured prompts, with limited coverage of complex multi-shot settings and challenging scenarios such as counterfactual compositions. Their typically static pipelines further limit reliable assessment of complex dimensions. In contrast, as summarized in Table 1, MSAVBench targets multi-shot audio-video generation, combining broad coverage of data settings and challenging cases with a robust, adaptive evaluation framework supporting self-correction and agentic scoring.

3 MSAVBench

3.1 Data Design

To comprehensively evaluate the MSAV ability of existing audio-video generation models, our data design is guided by two core dimensions: diversity and complexity.

Diversity. We decompose the MSAV generation task into four primary dimensions to ensure broad data coverage: **1) Video:** Spans diverse generation categories, visual styles, and subject types across varying scenes, color tones, and lighting conditions. **2) Audio:** Encompasses a wide range of sound sources, affective states (emotions), and multilingual spoken content. **3) Shot:** Introduces explicit professional cinematic language, including shot scales, camera angles, movement patterns, and cross-shot transitions. **4) Reference:** Extends beyond standard text-conditioned generation by incorporating reference conditions, such as characters, scenes, and audio, to evaluate identity and timbre preservation. A detailed distribution analysis is provided in Sec. 3.3.

Complexity. Beyond data diversity, data complexity is essential to probe the performance limits of existing models. We structure this complexity across two main perspectives: **1) Reality and Non-reality:** We explicitly categorize both subjects and scenes into *realistic* and *non-realistic* domains. The latter encompasses fictional worlds and counterfactual compositions. By cross-combining these axes, we evaluate a model's ability to faithfully adhere to complex prompts without mode collapse or falling back to common real-world data biases. **2) Challenging Scenarios:** We include a diverse range of challenging settings across both video and audio. These include overlapping simultaneous audio sources, complex fast-paced motions, dense on-screen text rendering, and diverse languages.

Most importantly, we push the structural boundaries of MSAV generation by extending narratives up to *15* shots, together with varying subject counts and mixed cinematic transitions.

3.2 Data Construction

To construct a high-quality benchmark adhering to the two data design principles, we introduce a four-stage pipeline integrating automated generation with human annotation in Figure 5.

Stage 1: Expert-driven taxonomy and quadruple construction. Domain experts first define an 8-category taxonomy based on video content genres (detailed in Sec. 3.3), which is further decomposed into fine-grained themes to prevent prompt homogenization. Concurrently, experts curate extensive candidate pools for subjects, scenes, and visual styles, strictly categorizing them into realistic and non-realistic domains. This process yields a vast combinatorial pool of $(theme, subject, scene, style)$ quadruples (see the Appendix B.2 for the complete taxonomy).

Stage 2: Prompt generation and rewriting. We randomly sample 2200 quadruples, and employ GPT-5.4 (OpenAI, 2026) to synthesize initial prompts based on these quadruples while extracting structured evaluation metadata (*e.g.*, shot counts, audio categories). We then use a *Prompt Enhancement model* to rewrite these initial prompts into comprehensive *global-to-shot* scripts. Each structured script comprises a global overview followed by detailed per-shot captions, which are enriched with explicit cinematic language, including camera parameters, transition cues, and lighting conditions.

Stage 3: Expert annotation and refinement. Six domain experts rigorously review the 2200 generated scripts to ensure diversity, structural complexity, and logical coherence. Experts filter out redundant and homogeneous cases, unnatural cross-shot transitions, and LLM hallucinations (*e.g.*, semantic deviations from the initial scripts), while manually refining ambiguous descriptions. This strict curation yields a high-quality prompt suite of 286 prompts comprising 2198 individual shots.

Stage 4: Reference media collection. To support reference-conditioned generation, we first sample 1000 character image-audio pairs (spanning both realistic and anime domains) and 200 background images from established public benchmarks (Chen et al., 2026; Cai et al., 2024; Wei et al., 2026a; Wang, 2023). Next, we use a VLM (Gemini 3.1 Pro (Gemini Team, 2026)) to categorize these assets to align with the semantic conditions of our scripts. We then enforce strict global uniqueness constraints to map these candidates to specific scripts, while human experts meticulously filter out low-quality samples or misaligned matches. This yields a reliable reference subset of 68 subject images, 65 audio clips, and 32 scene images, assigned across 96 scripts.

3.3 Data Analysis

Visual and stylistic diversity. As detailed in Figure 2(A) and (B), the benchmark balances 8 genres (*e.g.*, Action) with demanding domains (*e.g.*, Scientific Experiments). Subjects encompass 4 main categories (*e.g.*, humans, fictional characters), situated across realistic (66.1%) and non-realistic (33.9%) scenes. Furthermore, Figure 2(D) illustrates 6 diverse visual aesthetics; while realism dominates, multiple stylized domains (*e.g.*, anime, cyberpunk) are included. This semantic and aesthetic diversity enables a comprehensive evaluation of models' adaptability and prompt adherence.

Acoustic and linguistic diversity. As illustrated in Figure 2(C), our benchmark includes diverse audio content, emotional expressions, and languages. Audio conditions span 6 broad categories (*e.g.*, speech and environmental noise), while explicitly annotated emotional attributes cover 7 distinct states (*e.g.*, happiness and fear). Furthermore, spoken content is distributed across 6 languages to support rigorous evaluation of multilingual audio-visual alignment.

Fine-grained cinematic language. As shown in Figure 2(E), we design professional cinematographic control into our benchmark. The prompts incorporate 5 major shot scales (*e.g.*, close-up, long shot), 5 major camera angles, diverse camera movements (*e.g.*, push-in, pan), and various lighting conditions. Additionally, we introduce multiple cross-shot transitions (*e.g.*, hard cuts, fade-ins), facilitating a rigorous assessment of the cinematic generation capabilities of current models.

Diverse reference assets. To support reference-conditioned tasks (*e.g.*, identity preservation and voice cloning), we provide 68 character images and 65 paired audio clips featuring extensive demographic and linguistic diversity. Additionally, 32 scene images across indoor and outdoor environments are included. These assets ensure robust conditioning for multi-modal generation.

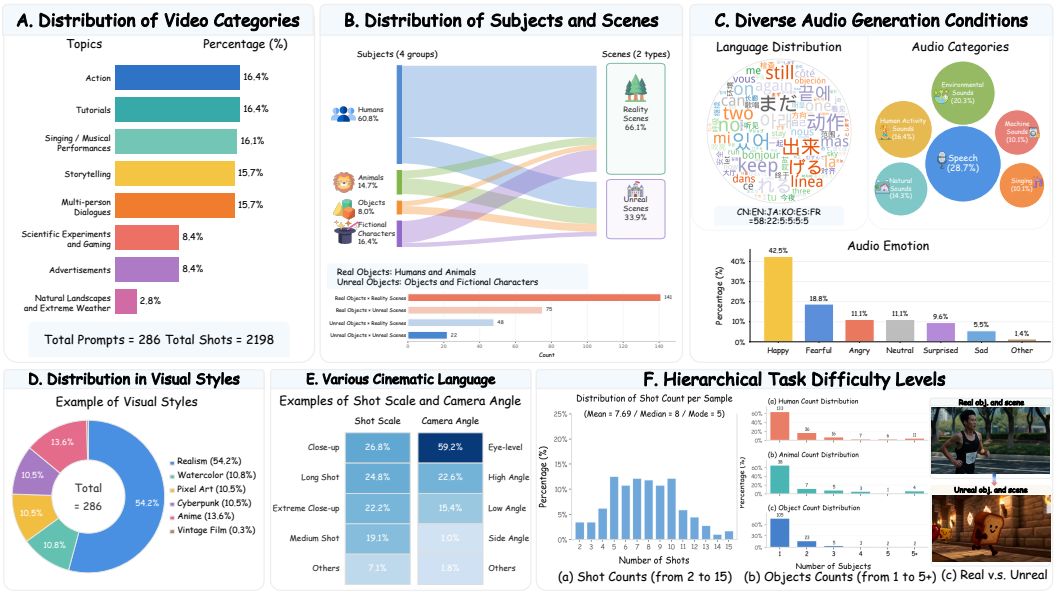


Figure 2: Diverse distribution of MSAVBench. The benchmark covers diverse generation categories (A), realistic and non-realistic subjects and scenes (B), varied audio conditions and languages (C), diverse visual styles (D), rich cinematic requirements (E), and a broad range of task difficulty in terms of shot count, subject count, and scenario complexity (F). More statistics are in Appendix B.3

Multi-level task complexity. As depicted in Figure 2(F), we scale the shot count from 2 to 15, with an average of 7.7 shots per prompt. Beyond single-subject prompts, 32.2% of prompts require multi-subject compositions, including scenarios with 5 or more simultaneous subjects. We further introduce challenging cases by cross-combining realistic and non-realistic subjects and scenes. This design facilitates systematic evaluation of models' capacities in long-form storytelling, complex spatial composition, and out-of-distribution generalization.

3.4 Evaluation Suite

3.4.1 Hierarchical Evaluation Metrics

We organize our evaluation metrics into four hierarchical levels, comprising 20 metrics in total (see Figure 1). More detailed descriptions of each metric are provided in Appendix C.

Global-level metrics. These metrics evaluate the overarching narrative, audio-visual alignment, and visual details across the entire video. *1) Narrative coherence*: Assesses logical plot progression based on discrete events. *2) Lip synchronization*: Evaluates lip-speech alignment across all dialogue shots. *3) Sound attribution*: Measures the temporal overlap between visually active speakers and their audio. *4) Audio-visual synchronization*: Measures the temporal offset between visual onsets and sound events. *5) Visual quality*: Evaluates fine-grained visual fidelity.

Cross-shot-level metrics. These metrics assess the consistency of visual content, audio properties, and complex spatial layouts across consecutive shots. *1) Cross-shot layout consistency*: Evaluates spatial layout coherence across shot transitions. *2) Visual consistency*: A composite metric comprising five sub-metrics: consistency of subject, background, style, illumination, and color across shots. *3) Music consistency*: Evaluates the stability of accompaniment, tempo, and rhythmic beats in non-speech background music across shots. *4) Speaker timbre consistency*: Verifies that the distinct vocal identities of multiple speakers remain stable across different shots.

Intra-shot-level metrics. These metrics evaluate generation quality and prompt adherence within individual shots. *1) Intra-shot layout-text alignment*: Assesses how accurately spatial layouts align with text prompts. *2) Camera parameter adherence*: Evaluates compliance with the specified camera scale, angle, and movement. *3) Audio quality*: Evaluates the acoustic quality of the generated audio. *4) Text rendering accuracy*: Measures the correctness of visually rendered text. *5) Word error rate*: Assesses speech transcription accuracy against the prompt-specified dialogue.

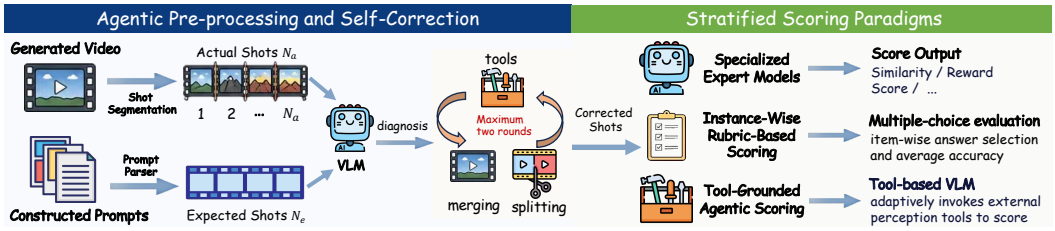


Figure 3: **Overview of the MSAVBench evaluation framework.** We first perform agentic pre-processing with iterative shot self-correction to improve boundary quality. Metrics are then evaluated with stratified scoring paradigms, including expert models for well-defined tasks, rubric-based VLM scoring for subjective dimensions, and tool-grounded agentic scoring for complex properties.

Reference-level metrics. These metrics assess fidelity to user-provided reference assets. *1) Subject fidelity*: Consistency with the reference image in appearance and identity. *2) Voice fidelity*: Consistency with the reference audio in vocal timbre.

Overall score. To avoid overemphasizing overlapping fine-grained aspects, we group related metrics into shared dimensions, merging five visual consistency metrics into *Visual Quality* and four dialogue-related metrics into *Multi-Speaker Dialogue Audio*, resulting in 11 final dimensions. We normalize these dimensions to $[0, 1]$, average them, and multiply the result by a shot-completion penalty coefficient based on the ratio of generated shots to the specified shot count. As shown in Sec. 4.4, this design aligns well with human judgments. More ablations are provided in Appendix F.3.

3.4.2 Adaptive Hybrid Evaluation Framework

Our evaluation framework consists of agentic self-correction and stratified scoring paradigms.

Agentic pre-processing and self-correction. To mitigate cascading failures caused by shot segmentation errors, we introduce an agentic pre-processing phase. Given a generated video, our framework first extracts initial temporal boundaries using TransNet V2 (Soucek & Lokoc, 2024). Since direct boundary prediction by VLMs is unreliable, we employ a VLM (Qwen3.5 (Team, 2026)) to iteratively inspect and evaluate the segments. Within a feedback loop, the VLM inspects the current segments, invokes merge/split tools, and reassesses the result to reduce segmentation errors. To balance accuracy and computational cost, we limit this process to a maximum of two iterations. In cases where the shot count remains mismatched after correction, the VLM performs a final shot-caption re-alignment, discarding non-aligned segments to ensure the integrity of downstream metric computations.

Stratified scoring paradigms. To balance evaluation cost, reliability, and comprehensiveness, we adopt three scoring paradigms based on metric complexity: *1) Specialized expert models*: For well-defined metrics (*e.g.*, subject similarity), we use dedicated expert models for efficient evaluation while aligning with standard practice. *2) Instance-wise rubric-based scoring*: For subjective dimensions (*e.g.*, narrative coherence), direct VLM scoring can be unstable. We therefore convert evaluation into prompt-specific rubrics, where the VLM answers predefined multiple-choice questions instead of producing unconstrained scalar scores. *3) Tool-grounded agentic scoring*: For complex compositional properties (*e.g.*, layout-text consistency), pure VLM reasoning is often insufficient. We thus augment scoring by allowing models to adaptively invoke perception tools (*e.g.*, object detectors and pose estimators) to extract objective evidence, which the VLM then uses to derive the final score.

4 Experiments

4.1 Experimental Setup

We benchmark 20 representative video generators on MSAVBench across two families. *(i) Closed-source commercial systems:* Seedance-2.0 (Seedance et al., 2026), Wan2.7-T2V (Wan Team, 2026), Kling-V3-T2V (Kling Team, 2026), HappyHorse (HappyHorse Team, 2026), Sora-2 (OpenAI, 2025), Veo-3.1 (Google DeepMind, 2026), as well as reference-conditioned models Wan-R2V (Wan Team, 2026) and HappyHorse-R2V (HappyHorse Team, 2026). *(ii) Open-source pipelines:* We further divide them into five categories: 1) single-shot audio-video models that are concatenated shot by shot, including JavisDiT++ (Liu et al., 2026b), JavisGPT (Liu et al., 2026a), MoVA (Team et al.,

Table 2: **Main results on MSAVBench.** The metrics are categorized into three dimensions: **Global:** `Narr.`: narrative coherence, `Lip`: lip synchronization, `Attr.`: sound attribution, `Sync`: audio-visual synchronization, `VQ`: visual quality. **Cross-Shot:** `C-Layout`: cross-shot layout consistency, `VC`: visual consistency, `Mus.`: music consistency, `Spk.`: speaker timbre consistency. **Intra-Shot:** `I-Layout`: intra-shot layout-text alignment, `Cam.`: camera parameter adherence, `PQ`: audio quality, `OCR`: text rendering accuracy, `WER`: word error rate. Top-5 cells are highlighted with a cyan gradient.

Method	Global					Cross-Shot				Intra-Shot					Overall ↑
	Narr. ↑	Lip ↑	Attr. ↑	Sync ↓	VQ ↑	C-Layout ↑	VC ↑	Mus. ↑	Spk. ↑	I-Layout ↑	Cam. ↑	PQ ↑	OCR ↑	WER ↓	
Closed-source commercial systems															
Seedance-2.0 (Seedance et al., 2026)	0.816	1.52	0.578	0.14	0.795	0.809	0.808	0.849	0.573	0.822	0.801	6.51	0.726	0.54	75.92
Veo-3.1 (Google DeepMind, 2026)	0.777	1.58	0.574	0.26	0.794	0.794	0.845	0.871	0.447	0.618	0.953	6.42	0.657	0.83	72.71
Wan2.7-T2V (Wan Team, 2026)	0.822	0.85	0.661	0.43	0.773	0.680	0.803	0.880	0.641	0.783	0.617	6.37	0.665	0.49	72.26
Kling-V3-T2V (Kling Team, 2026)	0.796	1.02	0.606	0.28	0.801	0.741	0.856	0.892	0.657	0.609	0.846	6.38	0.590	0.68	72.25
HappyHorse (HappyHorse Team, 2026)	0.825	0.73	0.579	0.24	0.804	0.632	0.790	0.833	0.673	0.628	0.732	6.60	0.689	0.51	71.89
Sora-2 (OpenAI, 2025)	0.852	1.87	0.568	0.50	0.792	0.717	0.808	0.834	0.520	0.722	0.784	5.64	0.675	0.75	71.19
Open-source ①: Native single-shot AV (concatenated shot-by-shot)															
LTX-2.3 (TI2AV) (HaCohen et al., 2024)	0.803	1.03	0.502	0.07	0.732	0.670	0.762	0.767	0.522	0.765	0.814	6.96	0.687	0.49	72.63
MoVA (TI2AV) (Team et al., 2026)	0.839	1.61	0.530	0.12	0.681	0.626	0.790	0.801	0.496	0.746	0.689	6.40	0.680	0.66	70.32
DaVinci+MagiHuman (TI2AV) (Chern et al., 2026)	0.787	3.08	0.580	0.07	0.685	0.422	0.816	0.957	0.674	0.473	0.563	5.82	0.650	0.82	65.01
LTX-2.3 (T2AV) (HaCohen et al., 2024)	0.768	0.96	0.608	0.09	0.754	0.439	0.596	0.770	0.562	0.348	0.781	6.94	0.586	0.53	64.40
DaVinci+MagiHuman (T2AV) (Chern et al., 2026)	0.776	4.91	0.654	0.05	0.699	0.267	0.586	0.958	0.699	0.494	0.472	5.78	0.164	0.83	60.65
JavisDiT++ (Liu et al., 2026b)	0.818	0.59	0.315	0.66	0.674	0.413	0.480	0.814	0.313	0.616	0.537	5.85	0.484	1.00	57.51
JavisGPT (Liu et al., 2026a)	0.745	0.42	0.113	0.54	0.633	0.351	0.554	0.792	0.097	0.362	0.624	6.09	0.268	0.99	53.95
Open-source ②: Long-video model + dubbing															
LongLive (Yang et al., 2025) + HunyuanFoley (Shan et al., 2025)	0.783	0.70	0.284	0.40	0.703	0.589	0.857	0.830	0.261	0.289	0.956	6.27	0.374	7.55	58.59
Helios (Yuan et al., 2026b) + HunyuanFoley	0.748	0.68	0.138	0.79	0.685	0.583	0.851	0.475	0.646	0.151	0.944	6.35	0.380	1.24	54.10
Open-source ③: Multi-shot video model + dubbing															
ShotStream (Luo et al., 2026) + HunyuanFoley	0.782	1.03	0.543	0.41	0.677	0.280	0.748	0.862	0.495	0.243	0.581	6.31	0.376	1.00	58.85
Open-source ④: Single-shot video-only model + dubbing (concatenated shot-by-shot)															
Wan2.2 (Wan et al., 2025) + HunyuanFoley (TI2AV)	0.794	1.19	0.378	0.43	0.685	0.679	0.747	0.814	0.314	0.430	0.957	6.08	0.590	1.39	63.42

2026) (TI2AV mode) and LTX-2.3 (HaCohen et al., 2024) and daVinci-MagiHuman (Chern et al., 2026) in T2AV and TI2AV modes; 2) multi-shot video models paired with dubbing models, such as ShotStream (Luo et al., 2026) with HunyuanFoley; 3) single-shot video models that are dubbed and then concatenated shot by shot, such as Wan2.2 (Wan et al., 2025) with HunyuanFoley (Shan et al., 2025) in TI2AV modes. 4) long-video generation models that can take multi-shot prompts as input and are paired with dubbing models, such as LongLive (Yang et al., 2025) with HunyuanFoley and Helios (Yuan et al., 2026b) with HunyuanFoley; 5) Reference-Conditioned Models: DreamID-Omni (Guo et al., 2026). Note that under the TI2AV setting, we utilize Wan2.7-Image (Mao et al., 2026) to generate a storyboard (image set) from the scripts as multi-shot priors, with each image explicitly fed as the visual condition.

4.2 Main Results

Table 2 details the overall performance on MSAVBench, from which we derive four key findings regarding current model bottlenecks.

Finding 1: A performance gap persists between closed and open-source models, but modular and agentic pipelines show potential to bridge it. While commercial systems dominate the leaderboard, the modular "image + audio-video" pipeline of LTX-2.3 (TI2AV) is competitive with several closed-source systems. An additional experiment reported in Appendix F.5 shows that the agentic framework ViMax (Huang et al., 2026a), using LTX-2.3 as its audio-video generation backbone, achieves a higher overall score than LTX-2.3 operating directly in either T2AV or TI2AV mode. This suggests that moving beyond basic modularity toward *a fully dynamic, agentic architecture could offer a viable path for the open-source community to challenge monolithic closed SOTA.*

Finding 2: Compared to basic audio-visual fidelity, open-source models lag behind closed systems in "director-level" structural control and cinematic language. Open-source models lag behind closed systems in complex spatial and cinematic compliance. Markedly lower scores in layout alignment (`C-Layout`, `I-Layout`) and camera control (`Cam.`) suggest these models currently act as passive pixel renderers rather than fully controllable storytellers.

Finding 3: Fine-grained joint audio-visual alignment remains an unsolved challenge for both closed and open-source models. Despite commendable unimodal generation quality, current systems still struggle with this inherent joint consistency. This is reflected in poor performance across lip-speech synchronization (`Lip`), sound attribution (`Attr.`), audio-visual synchronization (`Sync`), and multi-talker timbre consistency (`Spk.`). Accurately coupling phoneme-level audio with dynamic visual content across diverse cinematic languages remains a critical open problem.

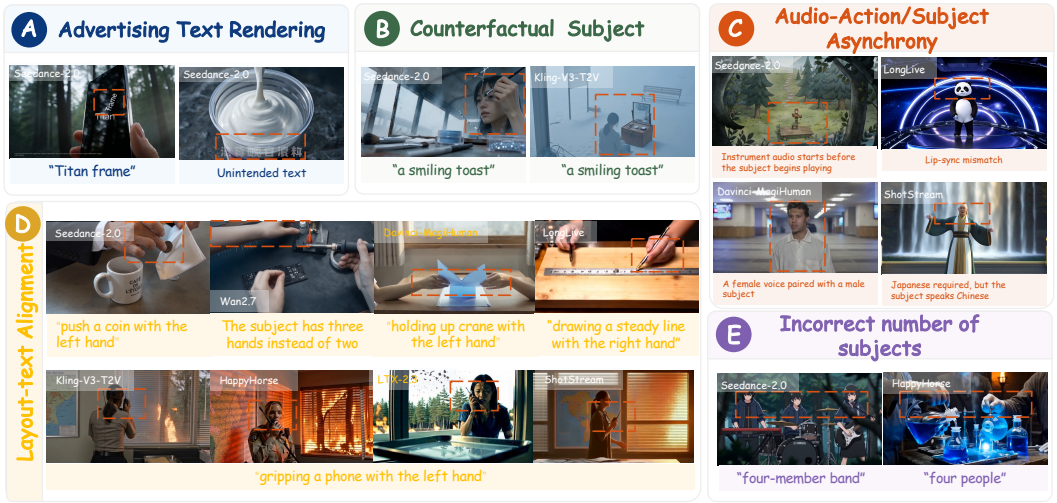


Figure 4: **Qualitative failure cases of evaluated models.** Examples include text rendering errors (A), counterfactual subject mismatches (B), audio-visual synchronization failures (C), layout control failures (D), and incorrect subject counts (E).

Table 3: **Overall score across prompts with different required shot counts.**

Method	Shot Count Range		
	1–4	5–10	11–15
Seedance-2.0	77.70	75.30	76.00
Wan2.7-T2V	73.50	71.70	72.10
Kling-V3-T2V	73.50	72.30	70.00
HappyHorse	74.90	70.60	72.80
LTX-2.3 (TI2AV)	75.10	72.00	72.50
DaVinci+MagiHuman (TI2AV)	70.90	64.70	62.30
Wan2.2 + HunyuanFoley (TI2AV)	71.80	62.80	60.10
ShotStream + HunyuanFoley	59.60	58.20	59.70
LongLive + HunyuanFoley	66.10	61.10	41.60
JavisDiT++	60.40	58.50	59.30

Table 4: **Overall score across realistic and non-realistic prompts.**

Method	Real.	Non-Real.
Seedance-2.0	76.80	74.50
Wan2.7-T2V	73.40	70.50
Kling-V3-T2V	73.50	70.30
HappyHorse	72.50	70.50
LTX-2.3 (TI2AV)	74.20	70.50
DaVinci+MagiHuman (TI2AV)	66.10	63.40
Wan2.2 + HunyuanFoley (TI2AV)	64.20	61.90
ShotStream + HunyuanFoley	59.90	56.70
LongLive + HunyuanFoley	60.30	56.00
JavisDiT++	61.00	56.40

Finding 4: The "video-first, post-hoc dubbing" pipelines remain limited in complex MSAV generation. Pipelines that use independent models (*e.g.*, HunyuanFoley) to dub pre-generated videos exhibit severe speech errors (high `WER`) and poor lip synchronization. Additional results in Appendix F.6 show that replacing HunyuanFoley with stronger dubbing models, such as MMAudio (Cheng et al., 2025) or ThinkSound (Liu et al., 2025), still yields substantial dialogue errors. These findings suggest that improving the dubbing component alone may be insufficient for reliable MSAV generation, motivating architectures with unified audio-video generation.

4.3 Performance Analysis on Complex Scenarios

Qualitative analysis on challenging cases. As illustrated in Figure 4, the evaluated models exhibit five recurring failure modes. *1) Text rendering failures*: even leading closed-source models, such as Seedance-2.0 (Seedance et al., 2026), still struggle with fine-grained text generation, often producing misspelled or unintended text. *2) Counterfactual subject failures*: models may fail to generate subjects that match the prompt, such as producing an incorrect object instead of `a smiling toast`. *3) Audio-visual synchronization failures*: common issues include lip-sync mismatch and audio-action asynchrony even in top closed-source systems, while open-source models, such as Davinci-MagiHuman (Chern et al., 2026) and ShotStream (Luo et al., 2026), often show more severe errors, including audio-subject mismatch and failure to generate speech in the required language. *4) Layout control failures*: for prompts with spatial constraints, such as left-right hand relations, both closed-source models (*e.g.*, Seedance-2.0) and open-source models (*e.g.*, LTX) often fail to satisfy the required body-part configuration. *5) Subject count failures*: models frequently generate the wrong number of subjects in complex multi-subject scenes. These cases highlight that robust MSAV generation still requires progress in controllability, compositionality, and audio-visual alignment.

Quantitative analysis on shot counts and realistic vs. non-realistic data. We identify two main performance bottlenecks. **1) Shot count:** As shown in Table 3, performance declines across all models when the required shot count increases from 1–4 to 5 and beyond, demonstrating the inherent difficulty of long-horizon generation. Notably, open-source models degrade significantly more. For example, from 1–4 to 11–15 shots, closed-source Kling-V3-T2V drops by 3.5%, whereas open-source LongLive+HunyuanFoley collapses by 24.5% and Wan2.2+HunyuanFoley by 11.7%. This highlights multi-shot consistency as a critical weakness for open-source pipelines. **2) Realistic vs. non-realistic data:** As illustrated in Table 4, overall scores noticeably decrease on non-realistic prompts on all methods. Closed-source models like Seedance-2.0 drop by 2.3%, while open-source models face steeper declines (*e.g.*, JavisDiT++ drops by 4.6%). This indicates that generating out-of-distribution visual content universally compromises performance across all model families.

Quantitative analysis on reference-conditioned generation. Table 5 reveals a substantial visual fidelity gap between open- and closed-source models. Open-source DreamID-Omni significantly trails Wan-R2V and HappyHorse-R2V on Img-DINO and Img-Face, yet its voice similarity (0.535) closely approaches the closed-source HappyHorse-R2V (0.545). This highlights that visual preservation is harder than voice cloning in joint audio-visual customization, making cross-modal fidelity balance a critical direction for future research.

Table 5: **Results on reference-to-AV generation.**

Method	Img-DINO↑	Img-Face↑	Voice↑
Wan-R2V (Wan Team, 2026)	0.208	**0.368**	**0.657**
HappyHorse-R2V (HappyHorse Team, 2026)	**0.259**	0.244	0.545
DreamID-Omni (Guo et al., 2026)	0.119	0.054	0.535

4.4 Human Preference Alignment and Evaluation Robustness

Alignment with human perception. We employ Spearman's (ρ_s) rank correlation (Spearman, 1961) to measure consistency with expert human ratings. *1) Overall ranking:* As shown in Table 6, our overall score achieves a high Spearman correlation of $\rho_s = 0.915$, higher than using uncorrected TransNet boundaries, indicating strong alignment with human judgments and the benefits of agentic preprocessing. *2) Complex metrics:* We validate the reliability of our metric designs over direct VLM scoring on three challenging metrics. For narrative coherence, cross-shot layout consistency, and intra-shot layout consistency, our instance-wise rubrics and tool-grounded agentic scoring further improve Spearman correlations to 0.850, 0.767, and 0.786, respectively. These results show that rubric-based decomposition and tool-grounded evidence are critical for improving automatic evaluation on complex tasks.

Table 6: **Agreement with human experts.** Our overall ranking and designed metrics show strong correlation with human judgments.

Metric	Method	Spearman ρ_s ↑
Overall	Raw TransNet	0.893
Overall	**Ours**	**0.915**
Narrative Coherence	Direct VLM Scoring (Qwen3.5)	0.600
	Instance-wise Rubric (Qwen3.5)	**0.850**
Cross-Shot Layout Consistency	Direct VLM Scoring (Qwen3.5)	0.429
	Tool-Grounded (Qwen3.5)	**0.767**
Intra-Shot Text-Layout Alignment	Direct VLM Scoring (Qwen3.5)	0.405
	Tool-Grounded (Qwen3.5)	**0.786**
Lip synchronization	Expert model	0.64
AV synchronization	Expert model	0.72
Sound attribution	Expert model	0.62
Speaker timbre	Expert model	0.61

Validation of our AV metrics. As shown in Table 6, the four AV metrics correlate positively with human model rankings ($\rho_s = 0.61$–0.72). These results support the reliability of our AV metrics for evaluating audio-visual generation. More validation details and results are provided in Appendix F.1.

Robustness across VLM backbones. Replacing Qwen3.5 with Qwen2.5-VL preserves strong human alignment, including $\rho_s = 0.820$ for narrative coherence; full results are in Appendix F.4.

5 Conclusion

We present **MSAVBench**, the first multi-shot audio-video generation benchmark with an adaptive hybrid evaluation framework. Our benchmark provides comprehensive coverage of data dimensions and challenging scenarios, and supports reliable evaluation through agentic shot self-correction and stratified scoring paradigms. Our evaluation of 20 systems shows that modular and agentic open-source pipelines have the potential to narrow the gap with closed-source models. However, current models still remain far from director-level generation, particularly in cinematic control and fine-grained audio-visual synchronization. We believe that MSAVBench, together with the insights it provides will serve as a rigorous benchmark and diagnostic tool for future MSAV research.

AI USE STATEMENT

We used GPT-5.4 for prompt construction and enhancement, Gemini 3.1 Pro for reference tagging and audio-related judgments, and Qwen-based models for visual evaluation, as detailed in Appendix D.1. We also used AI assistance for language editing and manuscript preparation. Human curation and evaluation protocols are described in Appendix E. We are responsible for the final content and reported results.

ETHICS STATEMENT

The benchmark uses synthetic prompts. The reference media are from existing datasets. Data curation, privacy considerations, and asset licensing are discussed in Appendix G. Human annotation procedures are described in Appendix E.

REPRODUCIBILITY STATEMENT

The appendix provides prompt templates, metric definitions, score aggregation, implementation details, human annotation protocols, and additional validation experiments. We will release the benchmark prompts, shareable reference assets, and evaluation code.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Kevin Cai, Chonghua Liu, and David M Chan. Anim-400k: A large-scale dataset for automated end to end dubbing of video. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11796–11800. IEEE, 2024.

Zhe Cao, Tao Wang, Jiaming Wang, Yanghai Wang, Yuanxing Zhang, Jialu Chen, Miao Deng, Jiahao Wang, Yubin Guo, Chenxi Liao, et al. T2av-compass: Towards unified evaluation for text-to-audio-video generation. *arXiv preprint arXiv:2512.21094*, 2025.

Shunian Chen, Hejin Huang, Yexin Liu, Zihan Ye, Pengcheng Chen, Chenghao Zhu, Michael Guan, Rongsheng Wang, Junying Chen, Jianye Hou, et al. Talkvid: A large-scale diversified dataset for audio-driven talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3492–3500, 2026.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911. IEEE, 2025.

Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, et al. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 244–250. IEEE, 2021.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*, 2025a.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025b.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl-1.5: Towards a multi-task 0.9 b vlm for robust in-the-wild document parsing. *arXiv preprint arXiv:2601.21957*, 2026.

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174*, 2019.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Gemini Team. Gemini 3.1 Pro. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, 2026.

Google DeepMind. Veo 3.1. https://deepmind.google/models/veo/, 2026.

Jiazhi Guan, Kaisiyuan Wang, Zhiliang Xu, Quanwei Yang, Yasheng Sun, Shengyi He, Borong Liang, Yukang Cao, Yingying Li, Haocheng Feng, et al. Audcast: Audio-driven human video generation by cascaded diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10678–10689, 2025.

Xu Guo, Fulong Ye, Qichao Sun, Liyang Chen, Bingchuan Li, Pengze Zhang, Jiawei Liu, Songtao Zhao, Qian He, and Xiangwang Hou. Dreamid-omni: Unified framework for controllable human-centric audio-video generation. *arXiv preprint arXiv:2602.12160*, 2026.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, et al. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18858–18868, 2025.

Yujin Han, Yujie Wei, Yefei He, Xinyu Liu, Tianle Li, Zichao Yu, Andi Han, Shiwei Zhang, Tingyu Weng, and Difan Zou. Aesrm: Improving video aesthetics with expert-level feedback. *arXiv preprint arXiv:2604.28078*, 2026.

HappyHorse Team. HappyHorse 1.0. https://www.happyhorse.cn/, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. Vabench: A comprehensive benchmark for audio-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23345–23355, 2026.

Lingxuan Huang, Sizhe He, Hengji Zhou, Liqiang Nie, Lianghao Xia, and Chao Huang. Vimax: Agentic video generation. *arXiv preprint arXiv:2606.07649*, 2026a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329. IEEE, 2024.

Taejun Kim and Juhan Nam. All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5. IEEE, 2023.

Kling Team. Kling 3.0. https://kling.ai/, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Chunyu Li, Chao Zhang, Weikai Xu, Jingyu Lin, Jinghui Xie, Weiguo Feng, Bingyue Peng, Cunjian Chen, and Weiwei Xing. Latentsync: Taming audio-conditioned latent diffusion models for lip sync with syncnet supervision. *arXiv preprint arXiv:2412.09262*, 2024.

Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, Liangyin Chen, and Yanru Chen. Lr-asd: Lightweight and robust network for active speaker detection. *International Journal of Computer Vision*, 133(7):4749–4769, 2025.

Zhaohe Liao, Kaixun Jiang, Zhihang Liu, Yujie Wei, Junqiu Yu, Quanhao Li, Hong-Tao Yu, Pandeng Li, Yuzheng Wang, Zhen Xing, et al. Aibench: Evaluating visual-logical consistency in academic illustration generation. *arXiv preprint arXiv:2603.28068*, 2026.

Huadai Liu, Kaicheng Luo, Jialei Wang, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing. *arXiv preprint arXiv:2506.21448*, 2025.

Kai Liu, Jungang Li, Yuchong Sun, Shengqiong Wu, Daoan Zhang, Wei Zhang, Sheng Jin, Sicheng Yu, Geng Zhan, Jiayi Ji, et al. Javisgpt: A unified multi-modal llm for sounding-video comprehension and generation. *Advances in Neural Information Processing Systems*, 38:142289–142324, 2026a.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. *arXiv preprint arXiv:2602.19163*, 2026b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024a.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22139–22149, 2024b.

Yawen Luo, Xiaoyu Shi, Junhao Zhuang, Yutian Chen, Quande Liu, Xintao Wang, Pengfei Wan, and Tianfan Xue. Shotstream: Streaming multi-shot video generation for interactive storytelling. *arXiv preprint arXiv:2603.25746*, 2026.

Chaojie Mao, Chen-Wei Xie, Chongyang Zhong, Haoyou Deng, Jiaxing Zhao, Jie Xiao, Jinbo Xing, Jingfeng Zhang, Jingren Zhou, Jingyi Zhang, et al. Wan-image: Pushing the boundaries of generative visual intelligence. *arXiv preprint arXiv:2604.19858*, 2026.

MiniMax. Minimax h3, August 2026. URL https://www.minimax.io/news/minimax-h3-open-source.

OpenAI. Sora 2. https://openai.com/index/sora-2/, 2025.

OpenAI. GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Taejin Park, Ivan Medennikov, Kunal Dhawan, Weiqing Wang, He Huang, Nithin Rao Koluguri, Krishna C Puvvada, Jagadeesh Balam, and Boris Ginsburg. Sortformer: A novel approach for permutation-resolved speaker supervision in speech-to-text systems. *arXiv preprint arXiv:2409.06656*, 2024.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025.

Haoyuan Shi, Yunxin Li, Nanhao Deng, Zhenran Xu, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. Msvbench: Towards human-level evaluation of multi-shot video generation. *arXiv preprint arXiv:2602.23969*, 2026.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.

Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11218–11221, 2024.

Charles Spearman. The proof and measurement of association between two things. 1961.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, et al. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wan Team. Wan2.7. https://wan.video/, 2026.

Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation. *arXiv preprint arXiv:2406.07686*, 2024.

Wei Wang. Japanese Anime Scenes. https://www.kaggle.com/datasets/weiwangk/japanese-anime-scenes, 2023.

Jianhui Wei, Xiaotian Zhang, Yichen Li, Yuan Wang, Yan Zhang, Ziyi Chen, Zhihang Tang, Wei Xu, and Zuozhu Liu. Univbench: Towards unified evaluation for video foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25654–25666, 2026a.

Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, et al. Dreamvideo-2: Zero-shot subject-driven video customization with precise motion control. *arXiv preprint arXiv:2410.13830*, 2024.

Yujie Wei, Shiwei Zhang, Hangjie Yuan, Yujin Han, Zhekai Chen, Jiayu Wang, Difan Zou, Xihui Liu, Yingya Zhang, Yu Liu, et al. Routing matters in moe: Scaling diffusion transformers with explicit routing guidance. *arXiv preprint arXiv:2510.24711*, 2025.

Yujie Wei, Xinyu Liu, Shiwei Zhang, Hangjie Yuan, Jinbo Xing, Zhekai Chen, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, et al. Dreamvideo-omni: Omni-motion controlled multi-subject video customization with latent identity reinforcement learning. *arXiv preprint arXiv:2603.12257*, 2026b.

Tianxin Xie, Wentao Lei, Kai Jiang, Guanjie Huang, Pengfei Zhang, Chunhui Zhang, Fengji Ma, Haoyu He, Han Zhang, Jiangshan He, et al. Phyavbench: A challenging audio physics-sensitivity benchmark for physically grounded text-to-audio-video generation. *arXiv preprint arXiv:2512.23994*, 2025.

Kaituo Xu, Yan Jia, Kai Huang, Junjie Chen, Wenpeng Li, Kun Liu, Feng-Long Xie, Xu Tang, and Yao Hu. Fireredasr2s: A state-of-the-art industrial-grade all-in-one automatic speech recognition system. *arXiv preprint arXiv:2603.10420*, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Chongyang Ma, Jiebo Luo, Li Yuan, et al. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *Advances in Neural Information Processing Systems*, 38, 2026a.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026b.

Lei Zhao, Linfeng Feng, Dongxu Ge, Rujin Chen, Fangqiu Yi, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. Uniform: A unified multi-task diffusion transformer for audio-video generation. *arXiv preprint arXiv:2502.03897*, 2025.

Yang-Hao Zhou, Haitian Li, Rexar Lin, Heyan Huang, Jinxing Zhou, Changsen Yuan, Tian Lan, Ziqin Zhou, Yudong Li, Jiajun Xu, et al. Mtavg-bench: A comprehensive benchmark for evaluating multi-talker dialogue-centric audio-video generation. *arXiv preprint arXiv:2602.00607*, 2026a.

Ziwei Zhou, Zeyuan Lai, Rui Wang, Yifan Yang, Zhen Xing, Yuqing Yang, Qi Dai, Lili Qiu, and Chong Luo. Avgen-bench: A task-driven benchmark for multi-granular evaluation of text-to-audio-video generation. *arXiv preprint arXiv:2604.08540*, 2026b.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. Muq: Self-supervised music representation learning with mel residual vector quantization. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, et al. Vistorybench: Comprehensive benchmark suite for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9455–9467, 2026.

Appendix

A More Related Work **17**

B More Data Details on MSAVBench **17**

B.1 Data Design Details 17

B.2 Data Construction Details 18

B.3 Data Analysis Details 20

C More Evaluation Suite Details on MSAVBench **21**

C.1 Metric Definitions, Tools and Score Mapping 21

C.2 Stratified Scoring Paradigms 23

D Additional Experimental Details **24**

D.1 Implementation 24

D.2 Cost-Efficient Evaluation 24

E Human Expert Annotation **24**

E.1 Experts for Benchmark Construction 24

E.2 Evaluation Experts and Pairwise Annotation Protocol 24

E.3 Annotation Interface 25

F Additional Validation Results **25**

F.1 Validation of Audio-Visual Metrics 25

F.2 Shot Self-Correction Ablation 26

F.3 Aggregation and Shot-Completion Penalty 26

F.4 Robustness Across VLM Backbones 27

F.5 Results on Agentic Generation 27

F.6 Results on More Powerful Dubbing Models 27

F.7 Reproducibility Details 27

G Ethics, Privacy, and Licensing **28**

H Limitations **28**

A More Related Work

Audio-video generation models. Building upon the success of image generation (Ho et al., 2020; Wei et al., 2025; Esser et al., 2024; Liao et al., 2026), earlier video generative models primarily target single-shot video synthesis (Brooks et al., 2024; Kong et al., 2024; HaCohen et al., 2024; Singer et al., 2022; Ho et al., 2022). While yielding impressive results, this paradigm is insufficient for scenarios requiring multi-scene narratives and synchronized audio (Blattmann et al., 2023; Polyak et al., 2024; Wei et al., 2024; 2026b). More recently, closed-source systems and emerging open-source models have explored multi-shot audio-video generation (OpenAI, 2025; Wan Team, 2026; Seedance et al., 2026; HappyHorse Team, 2026; Kling Team, 2026; Google DeepMind, 2026; MiniMax, 2026). Complementary pipelines combine multi-shot or long-video synthesis (Luo et al., 2026; Yang et al., 2025; Yuan et al., 2026b; Huang et al., 2026b; Zhu et al., 2026) with video-conditioned audio generation (Shan et al., 2025; Cheng et al., 2025; Liu et al., 2025). Some work also studies joint audio-video modeling (Wang et al., 2024; Zhao et al., 2025; Polyak et al., 2024) and audio-driven video generation (Guan et al., 2025). However, evaluation of MSAV models remains underexplored and highly challenging due to the need to assess both long-range multi-shot coherence and fine-grained audio-visual alignment.

Audio-video evaluation benchmarks. Early benchmarks such as VBench (Huang et al., 2024), Video-Bench (Han et al., 2025), and AesVideo-Bench (Han et al., 2026) mainly assess single-shot visual quality. Later multi-shot benchmarks (Zhuang et al., 2026; Wei et al., 2026a; Luo et al., 2026; Shi et al., 2026) extend evaluation to story structure and cross-shot consistency, but remain largely video-centric with limited audio assessment. Meanwhile, audio-video benchmarks (Zhou et al., 2026b; Xie et al., 2025; Zhou et al., 2026a; Hua et al., 2026; Cao et al., 2025) evaluate audio-visual alignment, yet mostly focus on single-shot or weakly structured prompts, with limited coverage of complex multi-shot settings and challenging scenarios such as counterfactual compositions. Their evaluation pipelines are also typically static, making it difficult to reliably assess more complex dimensions. In contrast, as summarized in Table 1, MSAVBench is tailored to multi-shot audio-video generation, combining broad coverage of data settings and challenging cases, together with a robust and adaptive evaluation framework that supports self-correction and agentic scoring.

B More Data Details on MSAVBench

B.1 Data Design Details

MSAVBench organises every prompt along the four orthogonal data-design dimensions of Sec. 3.1 (Video, Audio, Shot, Reference). Each dimension is annotated with a set of sub-attributes. Subjects and scenes are independently classified into two top-level reality classes – *realistic* and *non-realistic* – where the non-realistic class encompasses both *coherent fictional* (e.g. cyberpunk city) and *counterfactual* (e.g. a frozen tropical desert) sub-types.

Dim. 1 – Video. Four sub-attributes: (i) *video genre* (8 categories): Action, Narrative, Tutorial, Singing & Music performance, Multi-person Dialogue, Science / Game, Advertising, Nature; (ii) *visual style* (6 styles): photo-realistic, anime, watercolour storybook, pixel art, cyberpunk, retro film; (iii) *subject type* (4 classes): humans, animals, inanimate objects, fictional characters; (iv) *scene type*: realistic and non-realistic.

Dim. 2 – Audio. Three sub-attributes: (i) *audio content class* (6 categories): speech, singing, instrument / machine, human-made environment (e.g. laughter, footsteps), natural ambient, mixed (foley with voice-over, music with environment); (ii) *audio emotion* (7 emotions): joy, fear, anger, surprise, sadness, neutrality, awe; (iii) *spoken language* (6 values): Chinese, English, Japanese, Korean, Spanish, French.

Dim. 3 – Shot (cinematic language). Five sub-attributes annotated per shot: (i) *shot scale* (5 types): close-up, mid-close, mid, mid-long, long; (ii) *shot angle* (5 types): eye-level, top-down, low-angle, oblique, dutch; (iii) *camera motion* (4 types): push-pull, pan-tilt, tracking, hand-held/shake; (iv) *transition* (4 types): hard cut, dissolve, match cut, fade; (v) *lighting* (5 types): natural, side, soft, neon, low-key.

Dim. 4 – Reference. Three sub-attributes that are paired with a prompt: (i) *subject reference image* (68 images); (ii) *paired reference audio* (65 audio clips paired with the subject images); (iii) *scene reference image* (32 indoor / outdoor environments). All reference assets are assigned across 96 prompts.

B.2 Data Construction Details

B.2.1 Expert-Curated Sub-Category Vocabulary

The seed taxonomy used in Stage 1 of data construction contains an 8-genre top level whose second-level vocabulary on the released suite totals 144 fine-grained sub-categories.

The eight top-level genres and a representative subset of each genre's fine-grained sub-categories are: **(C1) Action** (32 sub-categories): martial-arts duel, kungfu choreography, weapon combat, parkour, street-dance battle, ballet, modern dance, basketball, football, swimming, gymnastics, boxing, skateboarding, rock climbing, surfing, marathon, BASE jumping, bungee jumping, wingsuit flying, card shuffling, surgical suturing. **(C2) Narrative storytelling** (19 sub-categories): detective reasoning, family warmth, romance, sci-fi adventure, historical legend, comedy, horror/thriller, coming-of-age, workplace drama, road-trip movie, human-animal interaction, animal documentary, fantasy adventure, war epic, courtroom drama. **(C3) Tutorial** (14 sub-categories): cooking, building blocks, origami, painting tutorial, instrument-fingering tutorial, fitness routine, makeup tutorial, woodworking, electronic soldering, gardening, CPR demonstration. **(C4) Singing & music performance** (17 sub-categories): solo pop / rock / folk / classical / rap, choir, band performance, conductor, piano solo, guitar solo, violin solo, drums solo, guzheng, saxophone, orchestral ensemble, street performance, music festival. **(C5) Multi-person dialogue** (18 sub-categories): family dinner, street encounter, classroom discussion, hospital visit, in-transit conversation, courtroom debate, news interview, talk-show panel, elevator small-talk, whispered exchange, negotiation, casual gossip. **(C6) Scientific experiment / game** (16 sub-categories): chemistry experiments (acid-base reaction, crystallisation, combustion, colour change), biology experiments (microscopy, plant growth, dissection), physics experiments (optical refraction, electromagnetic induction, fluid dynamics, free fall), astronomy observation, electronic games, board games, sports games. **(C7) Advertising** (19 sub-categories): sneaker, smartphone, automobile, perfume, food and beverage, skincare, movie trailer, game trailer, sports event, tourism destination, app UI demo, e-commerce listing. **(C8) Nature & extreme weather** ($\sim$9 sub-categories): aurora, volcanic eruption, deep-sea bioluminescence, forest fire, super-cell thunderstorm, sand storm, glacier collapse, polar night, monsoon rainfall.

B.2.2 LLM Prompt Templates

Stage 2 of data construction relies on two GPT-5.4 system templates: an *initial-prompt* template that turns a sampled $(theme, subject, scene, style)$ quadruple plus a target shot count into a structured multi-shot script with all evaluation metadata, and a *Prompt-Enhancement (PE)* template that rewrites the initial script into the cinematic global-to-shot format consumed by downstream generators. We show the two templates below.

Initial-Prompt System Template (GPT-5.4)

```
You are an expert prompt designer for multi-shot audio-video generation.  Given the
dimension constraints below and a target shot count, write one high-quality multi-shot
prompt and emit the structured metadata in JSON.

[Constraints]
(1) Anchor category (one of 8):  the narrative core must revolve around it.
(2) Global attributes (constant across the whole video, choose one value each):  visual
style ∈ {photo-realistic, anime, watercolour, pixel, cyberpunk, retro film}; subject;
colour tone (warm / cold / neutral); audio content (speech / singing / instrument /
human-made env.  / natural ambient / mixed).
(3) Per-shot attributes (may vary, but with continuity):  scene (transitions
narratively coherent, no abrupt jumps; do not freeze on one scene); lighting (1–2
dominant variants); audio emotion (1–2 dominant tones); spoken language (single
language, no code-switching); shot scale / angle / camera motion / transition (must
vary for cinematic richness).
(4) Cross-call diversity.  Vary style / subject / scene across calls.  Single- or
multi-character setups allowed.  Never use real personal names; use descriptive
references ("a young woman").
(5) Continuity.  Shots must follow a clear causal / temporal order, each a natural
```

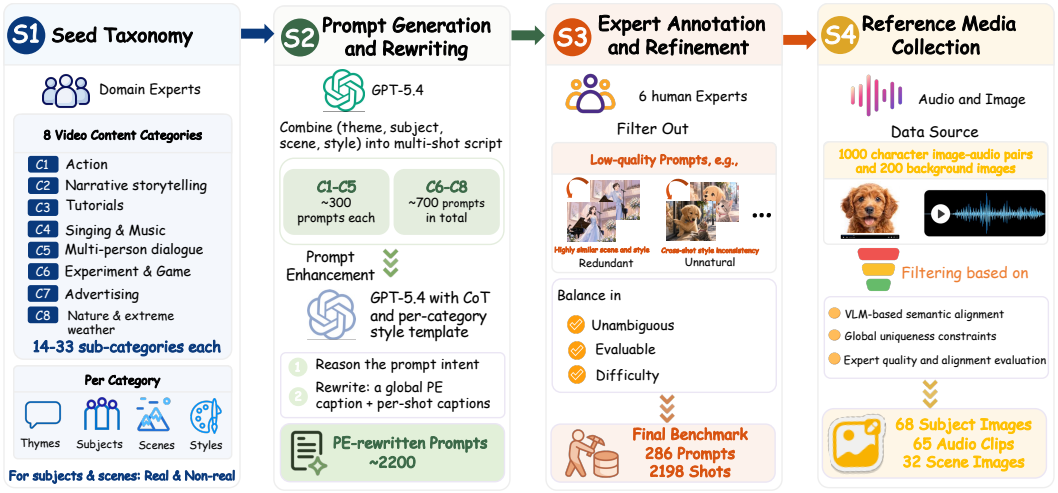


Figure 5: **The data construction pipeline of MSAVBench.** (1) Domain experts define an eight-category seed taxonomy with fine-grained sub-categories, with diverse types of subject, scene, and visual style. (2) GPT-5.4 first samples $(theme, subject, scene, style)$ quadruples and synthesises an initial multi-shot script with structured per-shot metadata; a Prompt-Enhancement (PE) model then rewrites it into the global-to-shot format with explicit cinematic language. (3) Six domain experts review every PE-rewritten script, filter out low-quality / hallucinated cases, and refine ambiguous descriptions. (4) Reference media are sampled from public benchmarks, automatically tagged by Gemini 3.1 Pro, and finally curated by experts to obtain a clean reference-conditioned subset.

continuation of the previous one.
(6) Per-category guidance. Apply the 8-category guidance (Action / Narrative / Tutorial / Singing & Music / Multi-person Dialogue / Science & Game / Advertising / Nature).

[Writing rules] produce exactly the requested shot count; every shot must be concrete and visually evocative; output one continuous paragraph (*global framing* → *Shot 1* → *Shot 2* → ...); no bullet points; output language matches the user input.

[Output format – JSON only]
{ "prompt": "<the multi-shot prompt>",
"dimensions": {
"video_content","video_style","video_subject","video_tone","audio_content": "<chosen value>",
"num_persons","num_animals","num_object_subjects": <int>,
"per_shot": [{ "shot_id":1, "scene","lighting","audio_emotion","language",
"shot_scale","shot_angle","camera_motion","transition": "<value>" }, ...] } }
Output JSON only; no extra text.

Prompt-Enhancement (PE) System Template (GPT-5.4)

You are an expert prompt engineer for multi-shot AI video generation. Given a user request, return one JSON object:
{ "is_safe": true|false, "reasoning": "...", "caption": "...",
"category": "Action | Narrative | Tutorial | Singing&Music | Multi-person Dialogue | Science&Game | Advertising | Nature" }

[Safety] is_safe=false only on sexual / graphically violent / politically sensitive content; otherwise true.

[Reasoning rules] Produce a structured, decision-oriented analysis (directional, no visual expansion): (i) pick the top-level category from one of the 8 above; (ii) identify the core technical challenge (e.g. lip-sync, dense text rendering, multi-shot editing logic); (iii) plan the shot count (1–15) and the role of each shot (establishing, close-up emotion, action climax); (iv) plan cinematic language: per-shot scale (long / full / medium / close / extreme close), angle (low / eye-level / high), camera motion (push-in / pull-out / pan / tilt), with explicit cross-shot diversity; (v) plan a unified colour tone and a 1–2-style lighting plan; (vi) plan audio: content type, 1–2 dominant emotional tones, a single spoken language, AV-sync points. Apply per-category guidance from the given list.

[Caption rules] Rewrite the user input into a single multi-shot caption with the following template:

- **Global framing** – open with "*This video contains X shots; it is a ⟨category⟩ in ⟨style⟩, ⟨colour tone⟩, accompanied by ⟨audio content⟩ in ⟨language⟩.*" Lighting and emotional colour are anchored on 1–2 dominant variants.
- **Per-shot detail** – each shot begins with "*Shot N [t_s–t_e s]*" (default total duration 15 s) and describes, in order: transition from previous shot; shot scale and angle (diverse across shots); composition and focus; scene; lighting; subject state and props; subject motion / physical change; audio emotional tone, language, and the precise AV interaction; camera-motion trajectory.
- **Continuity** – each shot is a natural continuation of the previous one; respect temporal and causal order; no abrupt jumps.
- **Naming** – replace any real personal name in the input with a descriptive reference ("a young woman").

[Output format] a single coherent natural-language paragraph (*global framing* → *Shot 1* → *Shot 2* → ...); no bullet points, no extra headings; rich vocabulary; output language matches the user input.

Output JSON only; no extra text.
[User original prompt]:

B.3 Data Analysis Details

We summarise the released 286-prompt / 2,198-shot benchmark below. The high-level distributions are visualised in Figure 2 (in the main text), and the cinematic-language distributions (shot scale, camera angle, transition, tone × saturation) are reported in Figure 6.

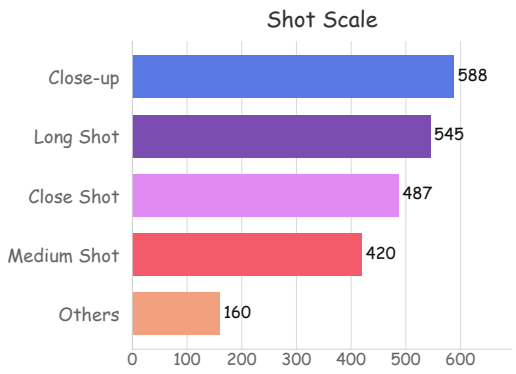
Visual and stylistic diversity. The eight video genres are balanced: Action 16.4%, Tutorial 16.4%, Narrative 15.7%, Singing & Music 16.1%, Multi-person Dialogue 15.7%, Science / Game 8.4%, Advertising 8.4%, Nature 2.8%. Subjects span humans (60.8%), animals (14.7%), inanimate objects (8.0%) and fictional characters (16.4%); scenes are realistic (66.1%) versus non-realistic (33.9%). Six visual styles are represented at the prompt level: photo-realistic 54.2%, anime 13.6%, watercolour 10.8%, pixel art 10.5%, cyberpunk 10.5%, retro film 0.3%.

Acoustic and linguistic diversity. Audio content per prompt is dominated by speech (28.7%), human-made environmental sounds (20.3%), nature ambient (14.3%), instrument / machine (10.1%), singing (10.1%), and human activity sounds (16.4%). Per-shot emotional colour spans the seven categories: joy (42.5%), fear / suspense (18.8%), anger / tension (11.1%), neutral (11.1%), surprise (9.6%), sad (5.5%), and others (1.4%). Spoken content is distributed across six languages – Chinese 165 prompts, English 64, Japanese 15, Korean 15, Spanish 14, French 13 – enabling explicit multilingual evaluation.

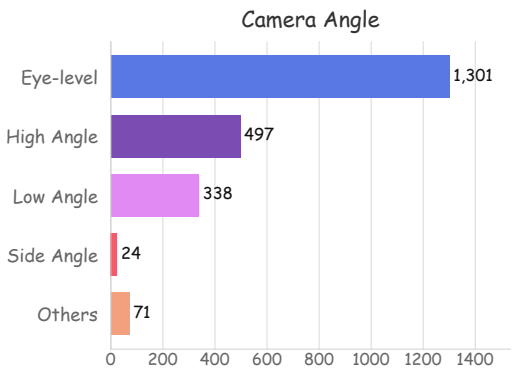
Cinematic language. MSAVBench reports 5 major shot scales (close-up 26.8%, long 24.8%, extreme close 22.2%, mid 19.1%, mid-close+mid-long 5.2%, plus a 1.9% tail) and 5 major shot angles (eye-level 59.2%, top-down 22.6%, low 15.4%, side 1.0%, others 1.8%). Camera motion is reported as 4 major types (push-pull 44.6%, pan-tilt 26.5%, tracking-and-orbit 5.8%, hand-held / shake 23.1%); transitions span 4 major types (hard cut 66.9%, dissolve 18.7%, none 13.0%, match cut / fade 1.4%); and lighting is reported with 5 major types (natural, side, soft, neon, low-key). The distributions are plotted in Figure 6.

Reference assets. The released reference subset contains 68 subject reference images, 65 paired reference audio clips and 32 scene reference images, all assigned across 96 prompts. Subjects span both realistic and anime domains; reference audio clips cover five age buckets (0–19, 19–30, 31–45, 46–60, 60+), multiple ethnicities and six languages; scene reference images cover both indoor environments (restaurants, bedrooms, offices) and outdoor environments (snack streets, playgrounds, parks, coastal areas, courtyards, grasslands).

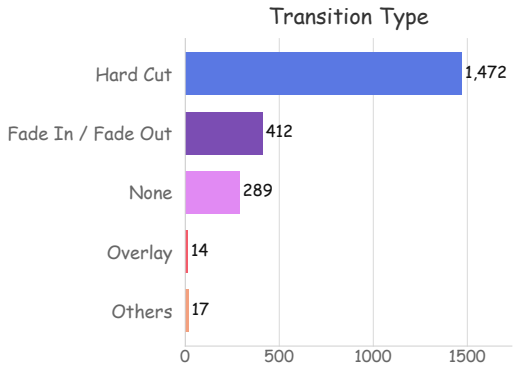
Multi-level task complexity. Shot count per prompt ranges from 2 to 15 (mean 7.7): 7% have 2–3 shots, 19% have 4–5, 23% have 6–7, 23% have 8–9, 18% have 10–11, 7% have 12–13 and 3% have 14–15. 32.2% of prompts require multi-subject composition, with over 10% demanding ≥ 5 simultaneous subjects. Cross-combining reality classes yields four difficulty buckets: realistic-subject × realistic-scene 49.3%, realistic-subject × non-realistic-scene 26.2%, non-realistic-subject × realistic-scene 16.8%, non-realistic-subject × non-realistic-scene 7.7%.



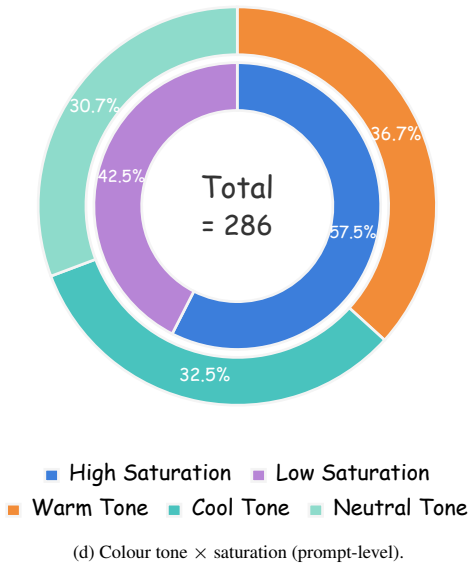
(a) Shot scale distribution (top-4 + tail).



(b) Camera angle distribution (top-4 + tail).



(c) Transition-type distribution (top-4 + tail).



(d) Colour tone × saturation (prompt-level).

Figure 6: **Long-tail cinematic-language and tonal distributions of MSAVBench.** Shot scale, camera angle, transition type and tone×saturation distributions on the released 286-prompt suite.

C More Evaluation Suite Details on MSAVBench

C.1 Metric Definitions, Tools and Score Mapping

Our evaluation framework contains 20 metrics organized into four levels: *Story*, *Cross-Shot*, *Intra-Shot*, and *Reference*. For each metric, we briefly specify *(i)* what it measures, *(ii)* the tool or judge used, *(iii)* how it is computed, and *(iv)* how the raw output is mapped to a score in $[0, 1]$.

C.1.1 Story-Level Metrics

(1) Narrative coherence. Measures whether the video forms a coherent story or valid procedural sequence across shots. It is evaluated by a rubric-based VLM judge (Qwen 3.5 (Team, 2026)) over uniformly sampled frames from the full video. The judge answers predefined binary questions about event ordering, causal validity, and completeness. The final score is the proportion of positive answers.

(2) Visual quality. Measures whether prompt-specified visual attributes are correctly realized. It is evaluated by a rubric VLM judge using prompt-instantiated multiple-choice questions. Each prompt slot is converted into an MCQ and scored by answer accuracy. The final score is the average MCQ accuracy.

(3) Audio-visual synchronization. Measures temporal synchronization between visual events and sound at the whole-video level. It is evaluated by DeSync metric, which is predicted by the Synchformer model (Iashin et al., 2024). The tool predicts the global audio-video offset. The raw offset Δt is first mapped to $[0, 1]$ by $\max(0, 1 - |\Delta t|/2.0\,\mathrm{s})$.

(4) Lip-speech synchronization. Measures lip-sync quality for dialogue-bearing shots. It is evaluated using active-speaker localization (Liao et al., 2025), speaker diarization (Park et al., 2024), and StableSyncNet (Li et al., 2024). Matched speaking segments are scored and averaged across the video. The raw sync confidence is directly used as the score.

(5) Sound attribution. Measures whether speech is temporally aligned with the correct visible speaker. It is evaluated using visual active-speaker detection (Liao et al., 2025) and audio diarization (Park et al., 2024). Speakers are matched across modalities and their temporal overlap is computed. The final score is the mean overlap ratio.

C.1.2 Cross-Shot-Level Metrics

(6) Cross-shot layout consistency. Measures spatial coherence of the main subject across shots, including position, orientation, scale, and prompt-specified hand relations. It is evaluated by a tool-grounded agentic judge with grounding (Liu et al., 2024a) and pose (Bazarevsky et al., 2020) tools. The score is computed from adjacent-shot consistency checks. The final score is the average pass rate.

(7) Subject consistency. Measures identity and appearance consistency of the main subject across shots. It is evaluated using a VLM localizer (Team, 2026), DINOv2(Oquab et al., 2023), and ArcFace (Deng et al., 2019). Subject crops are extracted and encoded, and pairwise similarities are computed across shots. The final score is $\max(0, \cos)$ averaged over pairs.

(8) Background consistency. Measures background stability across shots after removing the foreground subject. It is evaluated using foreground erasure and CLIP (Radford et al., 2021) image embeddings. Background embeddings are compared pairwise across shots. The final score is the mean clipped cosine similarity.

(9) Style consistency. Measures whether the visual style remains consistent across shots. It is evaluated using CSD-ViT-L (Somepalli et al., 2024) style embeddings. Pairwise cosine similarities are computed across shots. The final score is the mean clipped cosine similarity.

(10) Illumination consistency. Measures stability of lighting, shadow, and brightness across shots. It is evaluated by a rubric-based VLM judge over sampled frames. Adjacent shot pairs are checked for lighting consistency. The final score is the average pass rate.

(11) Colour consistency. Measures consistency of tone, saturation, and contrast across shots. It is evaluated by a rubric-based VLM judge. Adjacent shot pairs are compared for color consistency. The final score is the average pass rate.

(12) Music consistency. Measures continuity of background music across shots. It is evaluated using Demucs (Défossez et al., 2019), MuQ (Zhu et al., 2025), and MIR-AIDJ All-in-one (Kim & Nam, 2023). The score combines music embedding similarity, BPM agreement, and beat alignment. The final score is a weighted sum of these components in $[0, 1]$.

(13) Voice timbre consistency. Measures speaker timbre consistency across dialogue-bearing shots. Demucs (Défossez et al., 2019) isolates the vocal track, and SortFormer (Park et al., 2024) assigns global speaker identities across the concatenated audio. For each speaker–shot unit, we extract a w2v-BERT-2.0 (Chung et al., 2021) embedding and compare embeddings assigned to the same speaker across shots. The final score is a speech-duration-weighted mean of cosine similarities. Clipping to $[0, 1]$ is used for score aggregation.

C.1.3 Intra-Shot-Level Metrics

(14) Intra-shot layout-text alignment. Measures whether the spatial arrangement and hand actions within a shot match the shot caption. It is evaluated by a tool-grounded agentic VLM judge with grounding and pose tools. The judge answers predefined sub-questions for each shot. The final score is the average pass rate.

(15) Camera parameter adherence. Measures adherence to prompt-specified shot scale, angle, motion, and framing. It is evaluated by a rubric VLM judge over sampled shot frames. Each specified camera attribute is checked independently. The final score is computed as the average pass rate.

(16) Audio quality. Measures acoustic and production quality of generated audio. It is evaluated using Audiobox-Aesthetic (Tjandra et al., 2025). We use its production-quality sub-score for each shot. The raw score is mapped to $[0, 1]$ by $(\mathrm{PQ} - 1)/9$.

(17) Text rendering accuracy. Measures character-level fidelity of rendered on-screen text. It is evaluated using PP-OCRv5 (Cui et al., 2025b;a; 2026) on advertising-style prompts. Recognized text is compared against the target text using character error rate. The final score is $1 - \mathrm{CER}$, clipped to $[0, 1]$.

(18) ASR transcription (WER). Measures speech transcription accuracy against the prompt-specified script. It is evaluated using FireRedASR2-LLM (Xu et al., 2026) or Whisper-large-v3 (Radford et al., 2023), depending on language. The transcription is compared against the target using word error rate. The final score is $1 - \min(\mathrm{WER}, 1)$.

C.1.4 Reference-Level Metrics

(19) Subject fidelity. Measures whether the generated subject matches the reference image in identity and appearance. It is evaluated using the same subject-embedding pipeline as cross-shot subject consistency. Generated subject embeddings are compared with the reference image embedding. The final score is the mean clipped cosine similarity.

(20) Voice fidelity. Measures whether the generated speaker matches the reference voice in timbre. It is evaluated using the same speaker-embedding pipeline as cross-shot voice consistency. Generated speech embeddings are compared with the reference voice embedding. The final score is the mean clipped cosine similarity.

C.1.5 Overall Score Aggregation

Some atomic metrics reflect fine-grained sub-dimensions of the same underlying capability and therefore partially overlap in evaluation scope. Treating them as separate dimensions would overweight that capability in the final aggregation. We thus merge five visual consistency metrics into *Visual Quality* and four dialogue-related audio metrics into *Multi-Speaker Dialogue Audio*, yielding 11 final dimensions. Specifically, subject, background, style, illumination, and color consistencies are combined into *Visual Quality*, while voice timbre consistency, lip-sync, sound attribution, and ASR transcription are combined into *Multi-Speaker Dialogue Audio*. The remaining dimensions are kept separate.

All dimensions are mapped to $[0, 1]$ using metric-specific deterministic rules as described above, and then averaged. To account for structural failures in multi-shot generation, we further multiply the average by a shot-completion penalty coefficient, defined as the proportion of valid generated shots relative to the prompt-specified shot count. As reported in Table 6, this design shows strong alignment with human judgments.

C.2 Stratified Scoring Paradigms

All 20 metrics are implemented with one of three scoring paradigms.

(1) Specialized expert models (10 **metrics).** These metrics are computed by task-specific expert models or deterministic signal-processing pipelines without VLM-based reasoning. The corresponding metrics are: audio-visual sync., lip-speech sync., sound attribution, style consistency, music consistency, voice timbre consistency, audio quality, text rendering accuracy, ASR transcription (WER), and voice fidelity.

(2) Instance-wise rubric-based scoring (5 **metrics).** These metrics are computed by a single-pass VLM judge using fixed rubrics, with the final score given by the pass rate over applicable sub-questions. The corresponding metrics are: narrative coherence, visual quality, illumination consistency, colour consistency, and camera parameter adherence.

(3) Tool-grounded agentic scoring (5 metrics). These metrics rely on tool-grounded evaluation, where localized evidence from perception tools is used to support scoring. The corresponding metrics are: cross-shot layout consistency, subject consistency, background consistency, intra-shot layout-text alignment, and subject fidelity.

D Additional Experimental Details

D.1 Implementation

All perception tools are deployed as independent FastAPI micro-services on $8\times$A100 hosts. GPT-5.4 (OpenAI, 2026) is used for initial prompt generation and prompt enhancement. For VLM-based evaluation, Gemini 3.1 Pro (Gemini Team, 2026) is used for audio-related judgments, while Qwen3.5 (Team, 2026) is used for visual-related judgments. Tool outputs are cached at the case level and reused across metrics whenever possible. We use GPT-5.4 (2026-03-05) for prompt construction and enhancement, Gemini-3.1-Pro-preview for reference tagging, and Qwen3.5-397B-A17B as the primary visual judge with temperature 0. The smaller Qwen2.5-VL-32B-Instruct configuration is evaluated in Appendix F.4.

D.2 Cost-Efficient Evaluation

Evaluating multi-shot audio-video generation is inherently challenging, requiring a careful balance between evaluation accuracy and computational cost. Our framework is designed to remain efficient in both tool usage and VLM calls. *First*, not all metrics rely on VLM judges: many metrics are handled by specialized expert models or deterministic pipelines, which substantially reduces evaluation cost. *Second*, intermediate results are reused across metrics whenever possible; for example, shared outputs from subject localization, embedding extraction, foreground removal, OCR, and ASR are computed once and consumed by multiple metrics. *Third*, our framework is robust to different VLM backbones, and the smaller Qwen-based judge still achieves competitive alignment with human judgments, as shown in Sec. 4.4.

E Human Expert Annotation

E.1 Experts for Benchmark Construction

The Stage 1 taxonomy design and Stage 3 prompt curation in data construction pipeline are carried out by six domain experts, all of whom are full-time researchers in AIGC and audio-video generation. Each expert holds a graduate degree in computer vision, multimedia, or audio signal processing. During Stage 3, each PE-rewritten prompt is reviewed by at least two experts; disagreements on filtering or refinement are escalated to a third senior expert and resolved by majority vote. After this process, 286 of the original 2,200 PE-rewritten prompts (13.0%) are retained in the released benchmark, highlighting the strictness of the curation process.

E.2 Evaluation Experts and Pairwise Annotation Protocol

For the human-alignment study in Sec. 4.4, we recruit two groups of annotators, all of whom are full-time AIGC researchers and aesthetic-quality annotators with prior experience in aesthetic-quality annotation. The first group consists of 30 experts for system-level evaluation, comparing 16 video-generation models in terms of overall quality. Each annotator labels 40 video pairs, yielding a total of 1,200 pairwise judgments. The second group consists of 10 experts for fine-grained evaluation on three metrics: narrative coherence, cross-shot layout consistency, and intra-shot layout-text alignment. For each metric, annotators compare 10 candidate methods and label 36 pairs each, resulting in 360 judgments per metric. To ensure broad coverage, video pairs are uniformly sampled across genres, including realistic and stylized content, single- and multi-subject scenes, and videos with varying numbers of shots.

To reduce annotation bias, all videos are anonymized and presented in random order, and annotators follow a unified rubric for each evaluation metric. They are allowed to select one of three outcomes for each pair: “A wins,” “B wins,” or “both good / both bad.” Ties are counted as 0.5 for each method

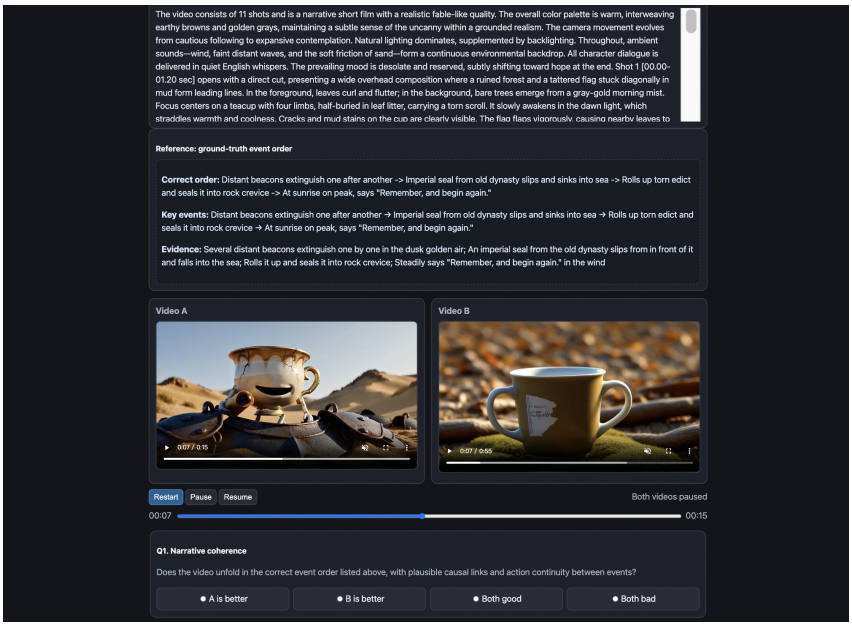


Figure 7: Screenshot of the annotation interface used for pairwise expert evaluation.

when computing win rates. The resulting method rankings are then compared with automatic metrics using Spearman's ρ.

E.3 Annotation Interface

Human evaluation is conducted via a custom web interface for fully anonymized pairwise comparison. Annotators are presented with two candidate videos together with the corresponding prompt and relevant metadata, and select the preferred result under the specified evaluation criterion. The resulting pairwise preferences are aggregated into system-level rankings. Figure 7 illustrates the interface.

F Additional Validation Results

F.1 Validation of Audio-Visual Metrics

We validate four audio-visual metrics using two complementary protocols: controlled positive/negative examples and human preferences on generated videos. These protocols test sensitivity to specific corruptions and agreement with human judgments, respectively.

Controlled positive/negative examples. Correctly matched real clips serve as positive samples. For lip synchronization and audio-visual synchronization, we use 100 talking clips from TalkVid (Chen et al., 2026); negatives are formed by shifting the audio (e.g., by 0.8 or 1.6 seconds) or replacing it with mismatched speech. For sound attribution, we use 78 multi-speaker clips from *The Big Bang Theory* and construct negatives by placing dialogue over scenes in which nobody on screen is speaking. For speaker timbre consistency, we use 60 single-speaker TalkVid clips and construct negatives by pairing speech segments from different speakers. We measure discrimination between positive and negative samples using ROC-AUC. As shown in Table 7, the four metrics achieve ROC-AUCs of 0.981–0.998, indicating that they distinguish correctly matched samples from the tested corruptions.

Human agreement on generated videos. Ten experts annotate a total of 300 video pairs from eight generation models on the multi-person dialogue subset. For each metric, we compare the resulting human model ranking with the automatic ranking using Spearman's ρ_s. As shown in Table 7, all four metrics correlate positively with human preferences.

Table 7: Validation of audio-visual metrics.

Metric	ROC-AUC ↑	Human Spearman ρ_s ↑
Lip synchronization	0.986	0.64
Audio-visual synchronization (DeSync)	0.981	0.72
Sound attribution	0.998	0.62
Speaker timbre consistency	0.998	0.61

Together, the controlled tests and human-alignment results support the reliability of our AV metrics for evaluating audio-visual generation.

F.2 Shot Self-Correction Ablation

Our evaluation self-correction operates on generated videos and refines their segmentation; it does not regenerate video content. A VLM inspects the current segments, invokes merge/split tools, and uses the resulting segments for the next inspection, with at most two iterations. This feedback loop is distinct from the agentic *generation* pipeline evaluated in Appendix F.5.

We compare corrected segments with raw TransNet V2 boundaries in two experiments. First, using the same human ranking as the main human-alignment study, self-correction improves overall Spearman correlation from 0.893 to 0.915. Second, ten experts annotate a subset of 200 videos. Compared with Raw TransNet V2, correction increases exact shot-count accuracy from 55% to 78% and reduces shot-count MAE from 0.68 to 0.25 (Table 8).

Table 8: Effect of evaluation self-correction.

Configuration	Human ρ_s ↑	Count accuracy ↑	Count MAE ↓
Raw TransNet V2	0.893	0.55	0.68
Agentic self-correction	**0.915**	**0.78**	**0.25**

F.3 Aggregation and Shot-Completion Penalty

To examine sensitivity to score aggregation, we compare our 11-dimension scheme with an unweighted average of all metrics and a three-level hierarchy that averages within the global, cross-shot, and intra-shot levels before averaging the three levels equally. All schemes use the same metric normalization and reference human ranking.

Table 9 shows that our scheme has the highest observed correlation with human rankings. The three aggregation schemes retain the same top-five members with modest reordering, and LTX-2.3 (TI2AV) remains above Sora-2. Removing the completion penalty reduces human correlation from 0.915 to 0.887.

Table 9: Sensitivity to aggregation and completion penalty.

Configuration	Human ρ_s ↑	Top-five ordering
11 dimensions (Ours)	**0.915**	Seedance-2.0 > LTX-2.3 (TI2AV) > Wan2.7-T2V > Kling-V3-T2V > HappyHorse
Mean of 18 metrics	0.906	Seedance-2.0 > Wan2.7-T2V > HappyHorse > Kling-V3-T2V > LTX-2.3 (TI2AV)
Mean of three levels	0.896	Seedance-2.0 > HappyHorse > Wan2.7-T2V > LTX-2.3 (TI2AV) > Kling-V3-T2V
Ours without penalty	0.887	Seedance-2.0 > Kling-V3-T2V > LTX-2.3 (TI2AV) > Wan2.7-T2V > HappyHorse

F.4 Robustness Across VLM Backbones

We replace Qwen3.5 with the smaller Qwen2.5-VL-32B-Instruct (Bai et al., 2025) while retaining the rubric- and tool-grounded evaluation designs. The smaller judge maintains positive human agreement on all three complex metrics and exceeds direct Qwen3.5 scoring (Table 10). This demonstrates that MSAVBench's evaluation framework is robust to the specific VLM choice, further validating the reliability of our metric design.

Table 10: Spearman correlation with human experts across scoring designs and VLM backbones.

Metric	Direct Qwen3.5	Ours, Qwen2.5-VL	Ours, Qwen3.5
Narrative coherence	0.600	0.820	**0.850**
Cross-shot layout consistency	0.429	0.732	**0.767**
Intra-shot layout-text alignment	0.405	0.741	**0.786**

F.5 Results on Agentic Generation

To test agentic generation directly, we evaluate the ViMax framework (Huang et al., 2026a) with LTX-2.3 as its video backbone. ViMax decomposes generation into narrative planning, storyboard and shot design, reference-image creation, and per-shot synthesis. A consistency-checking loop refines references and coordinates character, scene, and camera continuity before final assembly. In our configuration, LTX-2.3 supplies native audio-video generation. We compare this pipeline with the same backbone in T2AV and TI2AV modes.

Table 11 reports the 11 metrics used for aggregation. The agentic pipeline improves the overall score to 0.741, compared with 0.726 for TI2AV and 0.642 for T2AV, with gains in camera adherence and the composite dialogue score. This observation demonstrates the potential of the agentic pipeline for MSAV generation, providing further evidence for our Finding 1.

Table 11: Agentic generation with LTX-2.3. V-Det. is the visual-detail dimension; VQ aggregates visual consistency; Dialog aggregates speaker timbre, lip synchronization, sound attribution, and transcription.

Method	I-Lay.	C-Lay.	Cam.	Mus.	AV-sync	Aud.Q	OCR	Narr.	V-Det.	VQ	Dialog	Overall
ViMax + LTX-2.3	0.665	0.644	0.954	0.454	0.980	0.649	0.630	0.785	0.792	0.846	0.619	**0.741**
LTX-2.3 (TI2AV)	0.765	0.670	0.814	0.422	0.968	0.662	0.687	0.803	0.732	0.837	0.496	0.726
LTX-2.3 (T2AV)	0.348	0.439	0.781	0.423	0.958	0.660	0.586	0.768	0.754	0.690	0.507	0.642

F.6 Results on More Powerful Dubbing Models

To test whether the observed dialogue failures depend solely on HunyuanFoley, we re-dub videos from LongLive and a chained Wan2.2 pipeline using MMAudio (Cheng et al., 2025) and ThinkSound (Liu et al., 2025).

Table 12 shows that replacing HunyuanFoley with MMAudio improves LongLive's overall score from 0.586 to 0.593 and reduces WER from 7.55 to 3.54. However, dialogue accuracy and synchronization remain weak despite some gains in individual metrics. Thus, stronger dubbing yields modest or inconsistent overall improvements and does not fully resolve the observed failures. These results motivate designing a unified audio-video architecture.

F.7 Reproducibility Details

Human-study scope. The original overall ranking uses 30 annotators and 1,200 pairwise judgments. The original three complex-metric studies use ten annotators and 360 judgments per metric over ten candidate methods. The additional AV study uses ten experts and 300 video pairs from eight models, while the shot-count study uses 200 videos annotated by ten experts.

Table 12: Alternative dubbing models.

Pipeline	AV-sync ↑	Lip-sync ↑	Attribution ↑	WER ↓	Audio quality ↑	Overall ↑
LongLive + HunyuanFoley	0.801	0.096	0.360	7.55	0.585	0.586
LongLive + MMAudio	0.835	0.106	0.351	3.54	0.636	0.593
LongLive + ThinkSound	0.683	0.117	0.344	7.06	0.635	0.583
Wan2.2-chained + HunyuanFoley	0.783	0.144	0.583	1.95	0.564	0.634
Wan2.2-chained + MMAudio	0.714	0.119	0.655	2.42	0.599	0.630
Wan2.2-chained + ThinkSound	0.560	0.142	0.459	5.17	0.577	0.614

Dataset structure and access. The benchmark contains 286 prompts and 2,198 shots. Each prompt has story-level metadata and a list of shot-level controls; the 96 reference-conditioned prompts additionally identify their conditional assets. These assets comprise 68 subject images, 65 audio clips, and 32 scene images. Reference media retain the licensing conditions of their original sources.

G Ethics, Privacy, and Licensing

The text prompts in MSAVBench are synthetically generated from expert-designed taxonomies and subsequently reviewed by domain experts. They do not contain personal data, identifiable individuals, sensitive political or geographic content, or real proper names. The reference images and audio clips are drawn from previously published benchmarks with open redistribution terms and are used in accordance with their respective licenses. We further review these assets to exclude content that may raise privacy or cultural-sensitivity concerns.

Accordingly, MSAVBench does not introduce new privacy risks. The generated videos used in our experiments are produced solely for evaluation and are not redistributed. Upon release, we will provide the prompt set, the reference assets that can be legally shared, and the evaluation framework.

H Limitations

We discuss the limitations of our MSAVBench. First, some components of our agentic evaluation pipeline rely on multimodal foundation models as judges, which may introduce additional cost in large-scale evaluations. Nevertheless, as shown in Sec. 4.4, the framework remains well aligned with human judgment even when instantiated with a smaller open-source model, suggesting that our evaluation method is robust to the choice of VLM backbone.