CAST-BENCH: TRACING EVERY SOUND TO ITS SOURCE IN MULTI-SHOT AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

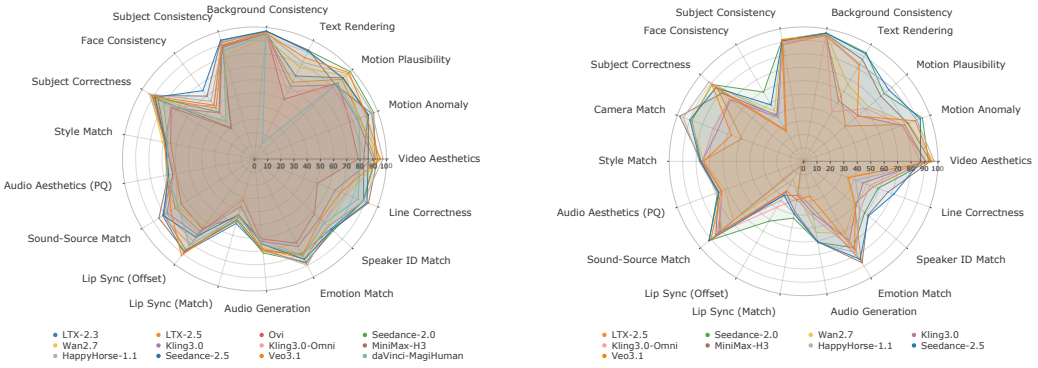
Existing benchmarks for joint text-to-audio-video (T2AV) generation evaluate perceptual quality and audiovisual alignment, but balanced sound-source sampling and separate reporting by shot structure remain limited. This makes it difficult to determine whether a specific sound is missing, attributed to the wrong source, or lost across shot transitions. To address this, we introduce CAST-BENCH, a benchmark designed to trace every generated sound back to its intended source in multi-shot audio-video generation. The core idea is to formulate T2AV correctness as an *audiovisual contract* in which each requested sound must be emitted by the right source, synchronized with the corresponding visual action, and carry the right semantic content. Concretely, the benchmark organizes 900 prompts into 9 sound-source categories, with 100 prompts per category, spanning speech, animals, instruments, ambient sounds, and other emitter types. The prompts include 439 single-shot and 461 multi-shot narratives, each derived from a structured specification that links every audio event to its emitter, subject, and temporal slot. The benchmark provides an 18-metric evaluation suite covering three dimensions: Visual Performance, Content Consistency, and Audio & Semantic Alignment. We further benchmark 12 representative generators and find that visual quality is nearly saturated across leading systems, yet audio semantic scores still differ markedly. Across the nine shared models, introducing shot transitions lowers the mean by 10.71 points and changes model rankings. Multi-person speech, consistent speaker identity across shots, and cross-shot event retention remain the main unresolved challenges. All prompts and evaluation code are available at https://anonymous.4open.science/r/cast-bench-9DE5/.

1 INTRODUCTION

Text-to-audio-video systems are shifting from silent video synthesis (OpenAI, 2024) to native joint video-audio generation (OpenAI, 2025; Gallegos & Iljic, 2025; Kuaishou Technology, 2026; Low et al., 2025; SII-GAIR et al., 2026; Tu et al., 2026). A prompt can now specify who speaks, the language and emotion they use, whether a dog barks off-screen, or whether a sound persists across a shot. Thus, the output must satisfy an audiovisual contract. The requested event is present, emitted by the intended source, synchronized with the visual action, and semantically consistent with the evolving story. However, this contract becomes much harder to satisfy in multi-shot narratives, where sounds must not only match their visual sources within each shot but also maintain identity and coherence across shots. A dog's bark may appear in the soundtrack yet be bound to the wrong animal, a line of dialogue may be spoken by the wrong character, or a sound event may drift into a shot where it was never requested. These are not corner cases but systematic breakdowns of the source-event-time binding that the contract demands, and no single audio or video quality score can reveal which binding failed or why.

Existing evaluation has not kept pace with this transition. We analyze three gaps in current evaluation suites: (1) Sound-source coverage is seldom controlled as a balanced sampling axis. VBench and its extensions organize prompts around visual attributes such as appearance, motion, and intrinsic faithfulness (Huang et al., 2024; 2025; Zheng et al., 2025), while recent T2AV suites categorize data by application, genre, or broad audio type (Zhou et al., 2026b; Wei et al., 2026). As a result, which sound sources are represented, and how frequently, is largely determined by incidental scene content.

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.



(a) Single-shot subset (12 models).

(b) Multi-shot subset (9 models).

Figure 1: **Performance radar for video generation models on CAST-BENCH.** The radar charts compare 12 single-shot and 9 multi-shot video generation models.

This makes the coverage of animals, instruments, ambient sounds, and off-screen sources difficult to assess systematically. (2) Aggregate audio scores can obscure distinct failures. An audio-quality or audio-text alignment score alone cannot distinguish whether an event is generated, whether it comes from the specified sound source, whether its timing is correct, and whether the utterance is produced by the correct speaker. Thus, a low aggregate score provides little insight into which requirement has failed. (3) Single-shot and multi-shot results share one leaderboard, so one-shot scores can dilute the cross-shot failures where sound-source binding is hardest.

These gaps motivate an evaluation protocol centered on sound sources. Each requested sound should have an explicit emitter, event, and temporal relation that can be traced back to the prompt. The protocol should test both common and challenging combinations of audio and visual conditions, separate single-shot quality from cross-shot continuity, and independently inspect audio evidence when visual cues could provide shortcuts. The goal is to identify which part of the audiovisual contract fails, rather than collapse all failures into a single preference score. Figure 1 shows that leading systems cluster relatively closely on visual dimensions, whereas audio and alignment dimensions provide clearer separation. On multi-shot prompts, the largest performance drops occur in speech, synchrony, and source binding. The radar provides a capability map of dimensions that a video-centric score cannot resolve.

We introduce CAST-BENCH, a sound-source-centered benchmark for diagnosing joint video-audio generation. To give evaluation an explicit reference, each prompt is derived from a structured specification of the requested subjects, sounds, and temporal relations. Complementary metrics distinguish visual quality, content correctness, and audiovisual correspondence, while separate reporting for single-shot and multi-shot prompts exposes differences that a pooled score can obscure. Implementation details are provided in Appendix D.

Our contributions are threefold. (1) **Sound-category coverage:** 900 bilingual prompt–specification pairs cover nine categories and both shot settings. (2) **Diagnostic evaluation:** 18 metrics combine specialist estimates with rubric-based judgments of visual and audio-related requirements. (3) **Empirical analysis:** experiments compare 12 generators, analyze speaker-count and shot-count groups, and report agreement with human judgments.

2 Related Work

Joint video–audio generation. Large-scale diffusion models trained with flow-based objectives (Ho et al., 2022; Blattmann et al., 2023; Peebles & Xie, 2023; Lipman et al., 2023; Kong et al., 2024) underpin many text-to-video models. Early systems (OpenAI, 2024; Wan et al., 2025; HaCohen et al., 2024) synthesized silent frames, requiring a separate stage to add Foley, speech, or music. Current text-to-audio-video (T2AV) systems (Team Seedance et al., 2026; Gallegos & Iljic, 2025; Alibaba Cloud, 2026a; HaCohen et al., 2026; Low et al., 2025; SII-GAIR et al., 2026) jointly generate frames and a synchronized waveform within a shared generation process. A complementary line of research focuses on how audio is generated and integrated. Video-to-audio models, such

as MMAudio, Kling-Foley, and HunyuanVideo-Foley, generate synchronized sound from a fixed visual track (Cheng et al., 2025; Wang et al., 2025c; Shan et al., 2025). Joint audio-video systems, including AV-DiT, UniVerse-1, UniForm, MOVA, and JavisDiT, couple the two modalities during generation (Wang et al., 2024; 2025a; Zhao et al., 2025; SII-OpenMOSS Team et al., 2026; Liu et al., 2025). Audio-driven models such as Wan-S2V and StableAvatar instead condition visual generation on supplied audio (Gao et al., 2025; Tu et al., 2025). As these models move from single clips toward cinema-like storytelling, maintaining structure and continuity across subjects, scenes, and sounds becomes increasingly important. This makes the evaluation of multi-shot audio-video generation, especially the consistency between what is seen and heard, both increasingly important and challenging.

T2V and T2AV benchmarks. VBench, EvalCrafter, VBench++, VBench-2.0, and T2V-CompBench evaluate visual quality, motion, and text-video alignment, including compositionality and physical plausibility (Huang et al., 2024; Liu et al., 2024; Huang et al., 2025; Zheng et al., 2025; Sun et al., 2024). AudioCaps and Clotho provide audio-captioning data (Kim et al., 2019; Drossos et al., 2020); AudioTime adds temporal audio-text annotations, and TTA-Bench evaluates text-to-audio systems (Xie et al., 2024; Wang et al., 2025b). These resources do not test correspondence to generated visual sources. Recent joint audio-video benchmarks evaluate multi-shot narratives, dialogue, physics, and audiovisual alignment (Wei et al., 2026; Zhou et al., 2026b;a; Xie et al., 2025; Cao et al., 2025; Hua et al., 2025). Table 1 situates our benchmark among these suites. Our focus is balanced, auditable sampling of emitting sound sources, with separate evaluation of single- and multi-shot outputs.

Benchmark	Task	#M	#A	#P	MS
VBench (Huang et al., 2024)	T2V	16	–	946	✗
EvalCrafter (Liu et al., 2024)	T2V	17	–	700	✗
Video-Bench (Han et al., 2025)	T2V	9	–	419	✗
TTA (Wang et al., 2025b)	T2A	13	5	2,999	✗
Harmony (Hu et al., 2025)	TI2AV	15	3	150	✗
UniAVGen (Zhang et al., 2025)	TI2AV	9	1	100	✗
AVGen (Zhou et al., 2026b)	T2AV	10	3	235	✓
MSAV (Wei et al., 2026)	T2AV	20	3	286	✓
Ours	**T2AV**	**18**	**9**	**900**	✓

Table 1: Compact comparison with representative benchmarks. **#M**: listed metric entries; **#A**: audio categories/subsets; **#P**: prompt entries; **MS**: includes multi-shot prompts. Counts follow each suite's own grouping.

3 CAST-BENCH

3.1 Design principles and task formulation

A generated video may contain a requested sound without placing it in the intended shot or associating it with the intended subject. Assessing these requirements therefore needs an explicit description of the expected events and their context. Each CAST-BENCH item pairs a natural-language prompt with a structured annotation of the requested subjects, sounds, and shot requirements. The generator receives the prompt, while the evaluator also uses the annotation to identify the requirements being assessed.

To give different sound categories comparable representation, we allocate 100 prompts to each of nine categories. We also record shot count, visual style, sounding condition, source complexity, video type, and camera movement. These attributes describe the requirements of each prompt and support analyses across different task conditions.

For multi-shot prompts, the annotation specifies which subjects and audio events are expected in each shot, including identities that should persist across cuts. This distinguishes a requirement to produce an event somewhere in the video from a requirement to produce it in a particular context. We report single-shot and multi-shot results separately, together with individual metric scores, so that differences between task settings and failure types remain visible alongside the overall score.

3.2 Taxonomy and prompt space

Free-form prompt writing can leave sound events implicit or describe the same subject inconsistently across shots. We therefore construct the task requirements before expressing them as natural-language prompts. Figure 2 summarizes the construction process: sound categories guide specification design, and the resulting specifications are converted into prompts and checked for consistency.

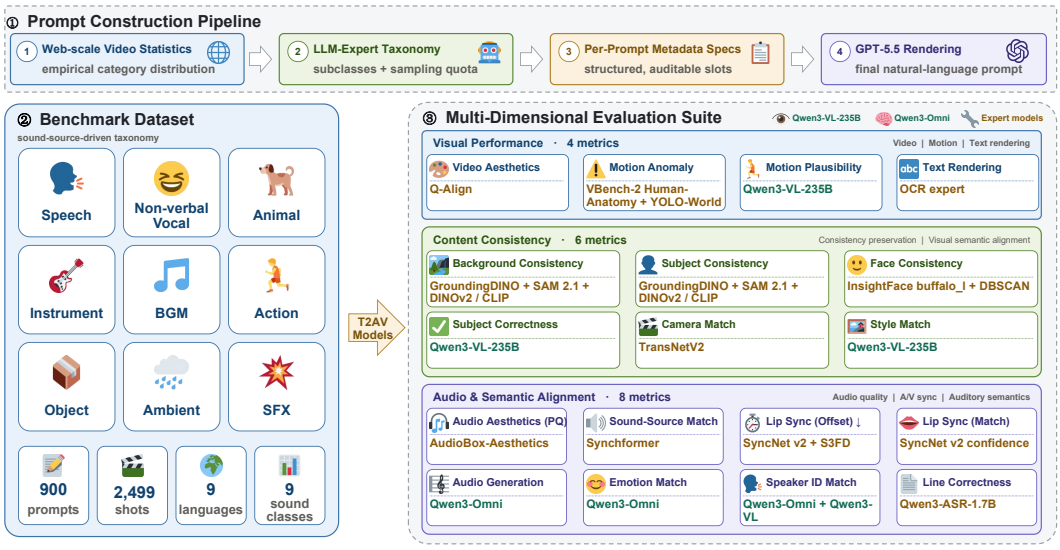


Figure 2: **Overview of the CAST-BENCH evaluation framework.** Panel ① (Prompt Construction Pipeline) shows four stages—Web-scale Video Statistics, LLM-Expert Taxonomy, Per-Prompt Metadata Specs, and GPT-5.5 Rendering—followed by expert validation and coverage repair. Panel ② (Benchmark Dataset) summarizes the sound-source-driven taxonomy and its 900 prompts, 2,499 shots, nine languages, and nine sound classes. Panel ③ (Multi-Dimensional Evaluation Suite) organizes the shot-aware 18-metric harness into Visual Performance, Content Consistency, and Audio & Semantic Alignment, with single- and multi-shot scores reported separately.

This ordering allows coverage and evaluation requirements to be inspected before the prompts are used for generation.

We use observations of publicly available audiovisual examples to inform the sound categories and event descriptions. The purpose is to identify recurring types of audible content that the benchmark should include. The taxonomy groups prompts by sound category and refines each category into event or scenario subclasses, with finer labels where applicable. We combine these labels with the six task attributes when constructing structured specifications. Each specification records the intended subjects and audio events and associates shot-specific requirements with the relevant shots. Specifying these elements before writing the prompt makes the intended coverage explicit and provides a common reference for prompt construction and evaluation. We use GPT-5.5 to express each structured specification as a natural-language prompt. The conversion instructions require the annotated subjects, sound events, dialogue, and shot order to be preserved without introducing additional requirements. Explicit shot markers identify the events assigned to each shot in a multi-shot prompt. Generators receive this natural-language description, while evaluators also have access to its paired annotation. Appendix B.2 provides the conversion template and formatting rules.

A well-formed specification does not guarantee that its natural-language prompt preserves every requirement. We therefore check structural validity separately from prompt–specification consistency. Structural checks cover category labels, entity references, and shot sequences; semantic checks compare the rendered prompt with the specification to identify omitted or added events. We also check for near-duplicate prompts and inspect category coverage. Section 3.3 summarizes the resulting prompt distribution.

3.3 Dataset Statistics and Coverage

Figure 3 summarizes the 900-prompt suite, with 100 prompts in each sound category. It contains 439 single-shot prompts and 461 multi-shot prompts, with the latter requesting between two and nine shots. The requested shot counts sum to 2,499.

The suite includes 288 prompts with one visible sounding subject, 541 with multiple sounding subjects, and 71 with an off-screen emitter. A separate source-complexity attribute identifies 630 multi-

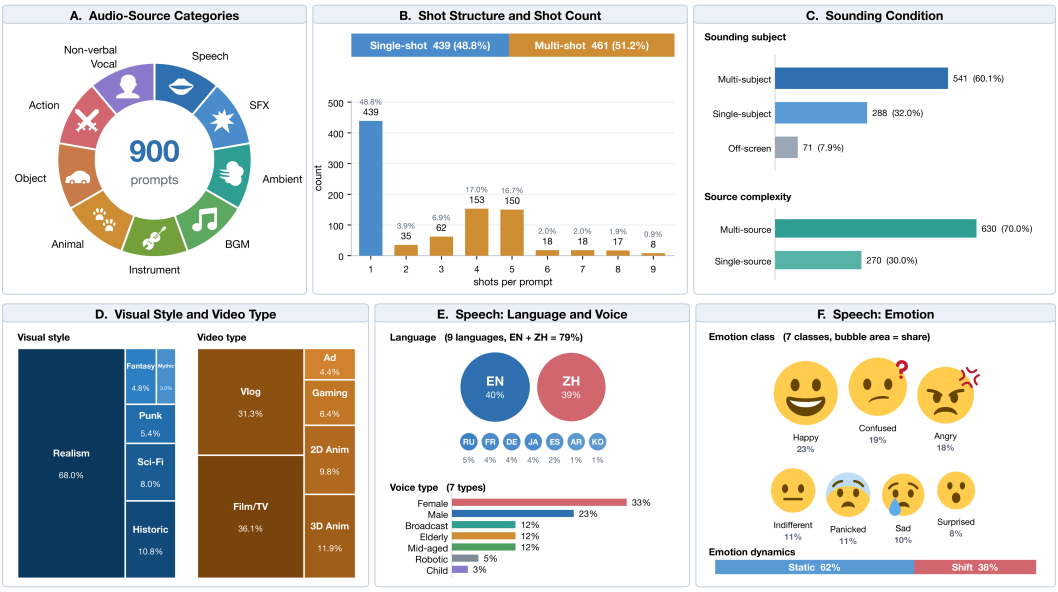


Figure 3: **Composition of the CAST-BENCH prompt suite.** The suite spans nine audio-source categories (A), single- and multi-shot structures with one to nine shots (B), sounding-subject and source-complexity conditions (C), visual styles and video types (D), speech languages and voice types (E), and emotion classes and dynamics (F). Detailed statistics and subclass breakdowns are provided in Appendix B.3.

source and 270 single-source prompts. The first attribute describes the sounding-subject condition, whereas the second describes the number of sound sources specified in the prompt. These conditions provide different tests of audio-content and source-correspondence requirements.

The prompts also vary in visual style, video type, speech language, and emotion requirements. Within the 100-prompt speech category, nine languages are represented, with English and Chinese together accounting for 79 prompts. Thirty-eight prompts in this category specify an emotion change, and 168 prompts across the full suite require on-screen text. These distributions describe the coverage of the benchmark rather than equally represented language or emotion conditions. Appendix B.3 reports the detailed subclass counts and their distribution.

3.4 Evaluation Framework

3.4.1 Hierarchical Evaluation Metrics

Perceptual quality, visual consistency, and audio-related correctness describe different aspects of a generated video. A visually convincing output can still omit a requested sound, render the wrong dialogue, or produce mismatched mouth motion. We therefore report 18 metrics in three groups: Visual Performance, Content Consistency, and Audio & Semantic Alignment. The individual scores help identify which aspects contribute to an overall performance difference.

Visual Performance. These metrics provide a reference for interpreting audio-related performance. *Video Aesthetics* (Wu et al., 2024) measures perceptual visual quality, while *Motion Anomaly* (Zheng et al., 2025) assesses detected abnormalities in human bodies, faces, and hands. *Motion Plausibility* uses five visual questions concerning naturalness, physical plausibility, temporal coherence, action completeness, and subject integrity. *Text Rendering* compares recognized text with the strings requested in the prompt. These measurements address specific aspects of visual output.

Content Consistency. These metrics cover both appearance stability and agreement with visual requirements. *Background Consistency* and *Subject Consistency* measure feature stability within and across detected shots, using separate background and subject regions. *Face Consistency* (Deng et al., 2019) combines agreement with the expected identity count and the stability of detected face

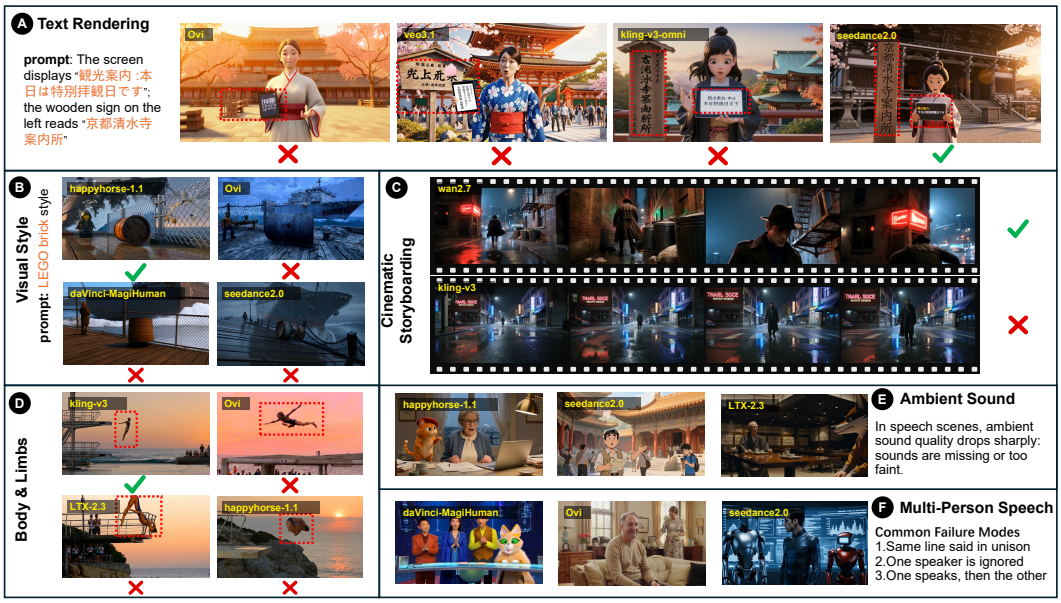


Figure 4: **Representative failure cases on CAST-BENCH.** The examples expose six recurring errors: (A) missing or incorrect in-scene text, (B) style drift, (C) incoherent shot storyboarding, (D) distorted body and limb motion, (E) missing ambient sound during speech, and (F) failure to render concurrent multi-person speech. ✓ denotes a correct result, ✗ a failure, and red boxes mark the evidence inspected by the corresponding metric.

identities. The remaining metrics in this group assess specific visual requirements in the prompt. *Subject Correctness* checks whether the described subjects appear with the specified identity attributes. *Camera Match* evaluates whether the generated video has the requested shot count. *Style Match* assesses agreement with the requested visual style and production type.

Audio & Semantic Alignment. Audio quality and audio-content correctness require different evidence. A soundtrack can be perceptually clear while omitting an event requested by the prompt. We use *Audio Aesthetics (PQ)*, the Production Quality score from AudioBox Aesthetics (Tjandra et al., 2025), to assess audio quality, and *Audio Generation* to assess requested content through event-presence questions and an overall audio–prompt alignment judgment. Reporting both scores distinguishes perceptual quality from whether the expected sounds are actually produced. Temporal correspondence is evaluated separately from audio-content presence. *Sound-Source Match* (Iashin et al., 2024) estimates the temporal offset between audio and corresponding visual activity; a smaller offset indicates better synchrony but does not verify the source assignment of every sound. For visible speech, *Lip Sync (Offset)* (Chung & Zisserman, 2016) measures only the temporal delay between mouth motion and speech, whereas *Lip Sync (Match)* measures whether the observed mouth shapes and speech content exhibit the expected phonetic correspondence. These metrics assess temporal correspondence under their respective applicability conditions and do not replace dialogue-content or speaker-assignment evaluation. For speech, we distinguish emotional expression, speaker assignment, and dialogue accuracy. *Emotion Match* evaluates the specified vocal or facial emotion requirements. *Speaker ID Match* compares a speaker-aware transcript with the requested speakers and utterances, while *Line Correctness* (Shi et al., 2026) measures transcript agreement with the requested dialogue using word error rate. Reporting these scores separately helps identify different speech-related failures.

3.4.2 Comprehensive Evaluation Pipeline

Prompt compliance involves both measurable signals and semantic requirements. We use specialist estimators for signals such as perceptual quality and temporal offset, and rubric-based judges for requirements such as event presence and speaker assignment. For multi-shot outputs, TransNetV2 (Souček & Lokoč, 2020) detects shot boundaries so that applicable metrics can operate on individual segments. The specialist branch runs dedicated estimators for visual and au-

Method	G1 Visual				G2 Consistency						G3 Audio & Alignment								Ovr ↑
	V-Aes ↑	Anat. ↑	Mot.P ↑	T-OCR ↑	BgC ↑	SbC ↑	FcC ↑	SbCor ↑	CamM ↑	StyM ↑	A-PQ ↑	SSM ↓	LipO ↓	LipM ↑	AuG ↑	Emo ↑	SpID ↑	ASR ↑	
Single-shot subset (12 models)																			
Seedance-2.5 (ByteDance Seed Team, 2026)	0.862	0.926	0.909	0.913	**0.974**	**0.936**	**64.887**	0.888	–	0.688	6.954	0.187	2.593	4.394	0.648	0.851	7.960	0.915	78.11
Seedance-2.0 (Team Seedance et al., 2026)	0.914	**0.970**	**0.983**	0.918	0.972	0.917	44.415	0.898	–	0.691	6.973	0.193	1.370	4.878	**0.715**	0.872	**8.080**	0.884	78.03
MiniMax-H3 (MiniMax, 2026)	0.932	0.954	0.833	**0.921**	0.971	0.933	59.699	0.915	–	0.674	6.984	**0.148**	1.240	4.742	0.595	0.883	7.560	**0.927**	77.97
HappyHorse-1.1 (Alibaba Cloud, 2026b)	**0.974**	0.958	0.957	0.820	0.961	0.899	54.709	**0.938**	–	0.687	**7.117**	0.212	2.241	3.553	0.696	**0.899**	7.600	0.886	77.41
Veo3.1 (Gallegos & Iljic, 2025)	0.951	0.910	0.970	0.858	0.957	0.899	51.154	0.906	–	0.695	7.035	0.255	**0.826**	4.407	0.697	0.813	6.800	0.700	74.52
LTX-2.3 (Lightricks, 2026a)	0.828	0.923	0.911	0.703	0.956	0.880	29.841	0.859	–	0.683	6.666	0.183	1.113	**5.089**	0.698	0.804	7.860	0.887	73.66
LTX-2.5 (Lightricks, 2026b)	0.914	0.859	0.896	0.650	0.956	0.887	48.321	0.878	–	0.684	6.987	0.294	3.154	3.769	0.682	0.845	7.560	0.887	73.64
Wan2.7 (Alibaba Cloud, 2026a)	0.959	0.852	0.937	0.777	0.924	0.816	33.798	0.922	–	**0.710**	6.466	0.312	3.612	2.866	0.708	0.823	7.520	0.749	70.43
Kling3.0-Omni (Kling Team et al., 2025)	0.848	0.794	0.807	0.564	0.960	0.923	50.831	0.752	–	0.674	6.583	0.180	1.846	3.099	0.616	0.819	7.120	0.738	69.81
Kling3.0 (Kuaishou Technology, 2026)	0.899	0.837	0.833	0.615	0.962	0.896	43.642	0.742	–	0.689	6.610	0.222	2.815	3.209	0.627	0.741	6.680	0.756	68.74
daVinci-MagiHuman (SII-GAIR et al., 2026)	0.802	0.849	0.851	0.137	0.944	0.853	36.012	0.809	–	0.668	6.269	0.321	1.800	4.553	0.645	0.823	7.200	0.843	67.75
Ovi (Low et al., 2025)	0.787	0.775	0.856	0.506	0.967	0.891	28.720	0.741	–	0.646	6.457	0.348	3.410	4.555	0.609	0.713	6.160	0.512	62.34
Multi-shot subset (9 models)																			
Seedance-2.5 (ByteDance Seed Team, 2026)	0.872	0.925	0.828	0.925	0.969	0.913	**62.487**	0.852	0.900	0.770	**7.704**	**0.080**	7.092	3.994	0.609	0.842	**6.320**	**0.715**	73.28
Seedance-2.0 (Team Seedance et al., 2026)	0.930	**0.942**	0.783	**0.933**	**0.971**	0.920	59.547	0.890	0.885	0.764	7.636	0.082	**4.851**	**4.297**	0.613	0.743	5.940	0.589	72.15
MiniMax-H3 (MiniMax, 2026)	0.945	0.827	0.753	0.876	0.961	0.910	39.547	0.790	**0.981**	0.761	7.609	0.146	7.349	3.780	0.608	**0.872**	6.250	0.669	69.45
HappyHorse-1.1 (Alibaba Cloud, 2026b)	**0.974**	0.896	**0.857**	0.837	0.961	0.903	38.413	**0.919**	0.953	**0.776**	7.169	0.204	6.648	3.960	**0.617**	0.786	5.460	0.413	66.41
Kling3.0-Omni (Kling Team et al., 2025)	0.870	0.788	0.611	0.426	0.960	0.891	40.388	0.719	0.717	0.753	6.793	0.163	6.059	2.867	0.425	0.771	5.520	0.527	60.22
Wan2.7 (Alibaba Cloud, 2026a)	0.969	0.832	0.632	0.659	0.965	**0.922**	43.254	0.893	0.894	0.743	6.596	0.317	8.432	2.522	0.537	0.623	4.980	0.343	59.80
Kling3.0 (Kuaishou Technology, 2026)	0.909	0.780	0.530	0.521	0.961	0.913	37.751	0.714	0.680	0.755	6.819	0.177	7.058	2.574	0.394	0.683	5.120	0.471	57.94
Veo3.1 (Gallegos & Iljic, 2025)	0.951	0.884	0.523	0.829	0.951	0.919	27.085	0.890	0.571	0.746	6.881	0.101	7.458	2.933	0.264	0.760	5.040	0.356	56.91
LTX-2.5 (Lightricks, 2026b)	0.908	0.801	0.408	0.559	0.953	0.882	25.645	0.834	0.490	0.754	7.143	0.198	9.605	2.427	0.344	0.861	4.960	0.552	56.13

Table 2: **Leaderboard on CAST-BENCH. Visual Performance:** V-Aes: Video Aesthetics, Anat.: Motion Anomaly, Mot.P: Motion Plausibility, T-OCR: Text Rendering. **Content Consistency:** BgC/SbC/FcC: Background/Subject/Face Consistency, SbCor: Subject Correctness, CamM: Camera Match, StyM: Style Match. **Audio & Semantic Alignment:** A-PQ: Audio Aesthetics (PQ), SSM: Sound-Source Match, LipO/LipM: Lip Sync (Offset)/(Match), AuG: Audio Generation, Emo: Emotion Match, SpID: Speaker ID Match, ASR: Line Correctness. Per column, **deep** / medium / light orange highlight the top-1/2/3 scores; light-blue model name = open-source.

dio aesthetics, human anatomy, OCR, segmentation (Ravi et al., 2025) and feature tracking, face identity, lip synchrony, source/onset displacement, and ASR. The LLM branch contains six rubric-based modules: Motion Plausibility, Subject Correctness, Style Match, Audio Generation, Emotion Match, and Speaker ID Match. GPT-5.5 first generates a saturated pool of questions from the prompt/specification and the metric-specific evidence schema. We filter this pool for entailment, observability, non-redundancy, and coverage of all required fields to obtain the final standard question bank. Qwen3-Omni serves as the audio-understanding judge (Xu et al., 2025) and Qwen3-VL as the video-understanding judge (Bai et al., 2025); they answer binary questions or fixed metric-specific scales (0–5 for Style Match; 1–5 for audio alignment and speaker assignment). Speaker ID Match is a staged exception: Omni first produces turn-level transcripts and speaker identities through multi-turn audio inspection, and VL then compares those identities and lines with the cast table and dialogue in the prompt. Detailed results retain the reporting scale specified for each metric, while overall scores use a common higher-is-better scale from 0 to 100. The two offset metrics are reversed during this conversion. Appendix D provides the metric-level scoring rules.

4 Experiments

4.1 Experimental setup

We evaluate twelve representative models: seven closed-source generators and five open-source models (MiniMax-H3, LTX-2.3, LTX-2.5, daVinci-MagiHuman, and Ovi). All models are evaluated on the original English prompts, except the Kling models, which use Chinese prompts because of their input-length limits. Outputs are 720p and nominally 10 s (8 s for Veo3.1), with three independent seeds or repeated API calls. We take a weighted average of the valid metrics after mapping them to a common $[0, 100]$ higher-is-better scale, reversing the two native lower-is-better measures. Because Ovi, LTX-2.3, and daVinci-MagiHuman do not support multi-shot generation, they are excluded from the multi-shot evaluation; the remaining nine models are evaluated on both single- and multi-shot prompts.

4.2 Performance across shot settings and evaluation dimensions

Table 2 reports the single-shot and multi-shot results. The three highest single-shot scores are close: 78.11 for Seedance-2.5, 78.03 for Seedance-2.0, and 77.97 for MiniMax-H3. The same systems lead the multi-shot subset with scores of 73.28, 72.15, and 69.45, respectively. Elsewhere, the ordering

differs between subsets: Veo3.1 ranks fifth among the nine shared models on single-shot prompts and eighth on multi-shot prompts.

Across the nine models, the mean overall score is 74.30 on single-shot prompts and 63.59 on multi-shot prompts, a difference of 10.71 points. All nine models obtain a lower overall score on the multi-shot subset. For Seedance-2.0, Motion Plausibility decreases from 0.983 to 0.783, Audio Generation from 0.715 to 0.613, and Line Correctness from 0.884 to 0.589. These results identify weaker performance under the multi-shot requirements in this benchmark.

Performance differences also depend on the evaluation dimension. On the single-shot subset, the mean of the eight normalized Audio & Semantic Alignment metrics is 60.3 for Ovi, 76.3 for Seedance-2.0, and 75.2 for LTX-2.3. LTX-2.3 therefore obtains a similar audio-related group score to Seedance-2.0 despite a lower visual-group mean, 84.1 versus 94.6. This comparison illustrates that visual and audio-related scores capture different aspects of performance; high visual scores alone do not establish audiovisual correctness.

Qualitative inspection identifies six recurring failures:

1. Insufficient text rendering: fails under multilingual and artistic font conditions.
2. Inaccurate video style: styles are difficult to match correctly across shots in multi-style videos.
3. Failures in high-dynamic scenes: cinematic shot sequencing becomes chaotic or shot breakdown fails, with large-scale limb deformation.
4. Concurrent dialogue is rarely realized: turns are serialized, one line is dropped, or both faces mouth the same words, breaking speaker and transcript alignment.
5. Insufficient speech-rate adaptation: long lines are clipped rather than accommodated by rate adaptation, while short lines are padded with unrequested hallucinated speech.
6. Low-quality ambient sound generation: in scenes with speech, background ambient sound is under-rendered, drowned out by speech, or largely missing.

Figure 4 shows representative cases and the metric evidence behind these diagnoses.

4.3 Detailed analysis

We compare performance across speaker-count and shot-count groups.

4.3.1 Speaker multiplicity and shot setting

Multiple speakers introduce additional requirements for dialogue and speaker assignment, while multi-shot prompts require these relations to be maintained across cuts. To examine performance under these two conditions, we compare the G3 scores of one-speaker and multi-speaker prompts separately within each shot setting (Table 3). Averaged equally across the nine models, the single-shot score is 68.72 for one-speaker prompts and 66.04 for multi-speaker prompts, a difference of 2.69 points. The corresponding multi-shot scores are 51.04 and 48.47, a difference of 2.57 points. The gap between shot settings is larger: 17.68 points for one-speaker prompts and 17.57 points for multi-speaker prompts.

Model	G3 Audio & Semantic Alignment			
	Single-shot		Multi-shot	
	1 spk.	>1 spk.	1 spk.	>1 spk.
Seedance-2.0	**75.41**	**74.39**	**59.12**	**57.25**
MiniMax-H3	74.32	73.22	56.30	53.62
Veo3.1	71.05	67.76	44.70	41.11
HappyHorse-1.1	69.76	66.22	54.24	51.58
Seedance-2.5	68.94	65.46	55.97	54.51
LTX-2.5	67.89	62.90	46.00	41.59
Kling3.0-Omni	64.85	63.25	53.39	50.66
Wan2.7	64.07	60.28	41.27	39.59
Kling3.0	62.22	60.85	48.38	46.30

Table 3: G3 Audio & Semantic Alignment score (0–100) by speaker count and shot setting.

The direction of these differences is consistent across all nine models. Every model scores lower on multi-speaker prompts than on one-speaker prompts in both shot settings, and lower on multi-shot prompts than on single-shot prompts within each speaker-count group. This consistency indicates that multi-speaker dialogue and cross-shot continuity are systematic bottlenecks rather than weaknesses of a particular architecture. Progress on either dimension alone is therefore unlikely to close the overall gap; models must jointly maintain speaker assignment and shot-level coherence.

4.3.2 Performance across requested shot counts

The single-shot/multi-shot comparison does not show whether performance declines progressively as more shots are requested. To examine this pattern, Figure 5 groups prompts by requested shot count and includes the single-shot subset as a reference. Among the three displayed systems, Seedance-2.0 has the highest score in every group, followed by Wan2.7 and Kling3.0. From one to nine shots, their scores decrease from 78.03 to 69.88, from 70.43 to 56.71, and from 68.74 to 53.30, respectively. The corresponding gaps of 8.15, 13.72, and 15.44 points show that the magnitude of the difference across shot-count groups varies substantially between models.

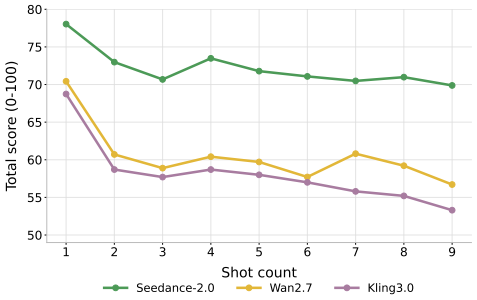


Figure 5: Overall score by required shot count for three representative models.

4.4 Sampling sensitivity and agreement with human judgments

An evaluation should not depend excessively on the particular prompt subset used to compare models. Figure 6 examines this sensitivity at sampling ratios of 25%, 50%, 75%, and 100% for Seedance-2.0, Wan2.7, and Kling3.0, separately for single-shot and multi-shot prompts. Across the displayed subsets, every curve varies by less than three points, and the ordering of the three systems remains unchanged within each shot setting. For each model, the multi-shot score is also lower than the single-shot score at every sampling ratio. These observations indicate good generalizability and a relatively uniform distribution of the prompt set, and that the overall evaluation system is relatively stable.

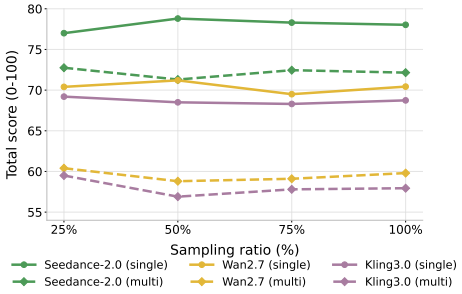


Figure 6: Overall score for three representative models as the evaluated prompt subset grows from 25% to the full suite.

The human evaluation yields Spearman correlations of 0.833 and 0.784 for the single-shot and multi-shot subsets, respectively, and 0.872, 0.767, and 0.602 for Speaker ID Match, Sound-Source Match, and Motion Plausibility. Three annotators independently score each video; Appendix E describes the scoring procedure, aggregation rules, and inter-rater reliability analysis.

5 Conclusion

We present CAST-BENCH, a 900-prompt benchmark organized by sound category with structured task annotations and 18 evaluation metrics, covering nine sound-source categories. Experiments on 12 representative generators show that strong visual quality does not guarantee reliable audio semantics or speaker correspondence, while the mean overall score decreases by 10.71 points on multi-shot prompts; further analyses reveal additional degradation under more demanding shot and speaker configurations. CAST-BENCH provides an auditable framework for measuring and comparing progress in complex audio-video generation.

References

Alibaba Cloud. Wan 2.7. Alibaba Cloud Model Studio documentation, 2026a. URL https://www.alibabacloud.com/help/en/model-studio/text-to-video-api-reference.

Alibaba Cloud. HappyHorse 1.1. Alibaba Cloud Model Studio model documentation, 2026b. URL https://www.alibabacloud.com/help/en/model-studio/happyhorse-1-1-t2v.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv,

Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. URL https://arxiv.org/abs/2304.08818.

ByteDance Seed Team. Seedance 2.5. Official model release, 2026. URL https://seed.bytedance.com/en/seedance2_5.

Zhe Cao, Tao Wang, Jiaming Wang, Yanghai Wang, Yuanxing Zhang, Jiahao Wang, Jialu Chen, Miao Deng, Yubin Guo, Chenxi Liao, Yize Zhang, Zhaoxiang Zhang, and Jiaheng Liu. T2AV-Compass: Towards unified evaluation for text-to-audio-video generation. *arXiv preprint arXiv:2512.21094*, 2025. URL https://arxiv.org/abs/2512.21094.

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 *CVPR*, 2025. URL https://arxiv.org/abs/2412.15322.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *ACCV Workshops*, 2016. URL https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/chung16a.pdf.

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 (CVPR)*, pp. 4690–4699, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP*, 2020. URL https://arxiv.org/abs/1910.09387.

Jess Gallegos and Thomas Iljic. Introducing Veo 3.1 and advanced capabilities in Flow. Google official announcement, 2025. URL https://blog.google/innovation-and-ai/products/veo-updates-flow/.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-S2V: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025. URL https://arxiv.org/abs/2508.18621.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. URL https://arxiv.org/abs/2501.00103.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233, 2026. URL https://arxiv.org/abs/2601.03233.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, and Yongxin Ni. Video-Bench: Human-aligned video generation benchmark. In *CVPR*, 2025. URL https://arxiv.org/abs/2504.04907.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *NeurIPS*, 2022. URL https://arxiv.org/abs/2204.03458.

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

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. *arXiv preprint arXiv:2512.09299*, 2025. URL https://arxiv.org/abs/2512.09299.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. URL https://arxiv.org/abs/2311.17982.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3633890. URL https://arxiv.org/abs/2411.13503.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP*, 2024. URL https://arxiv.org/abs/2401.16423.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *NAACL*, 2019. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, Xiao Hu, Xiaohua Hu, Boyuan Jiang, Fangyuan Kong, Hang Li, Jie Li, Qingyu Li, Shen Li, Xiaohan Li, Yan Li, Jiajun Liang, Borui Liao, Yiqiao Liao, Weihong Lin, Quande Liu, Xiaokun Liu, Yilun Liu, Yuliang Liu, Shun Lu, Hangyu Mao, Yunyao Mao, Haodong Ouyang, Wenyu Qin, Wanqi Shi, Xiaoyu Shi, Lianghao Su, Haozhi Sun, Peiqin Sun, Pengfei Wan, Chao Wang, Chenyu Wang, Meng Wang, Qiulin Wang, Runqi Wang, Xintao Wang, Xuebo Wang, Zekun Wang, Min Wei, Tiancheng Wen, Guohao Wu, Xiaoshi Wu, Zhenhua Wu, Da Xie, Yingtong Xiong, Yulong Xu, Sile Yang, Zikang Yang, Weicai Ye, Ziyang Yuan, Shenglong Zhang, Shuaiyu Zhang, Yuanxing Zhang, Yufan Zhang, Wenzheng Zhao, Ruiliang Zhou, Yan Zhou, Guosheng Zhu, and Yongjie Zhu. Kling-Omni Technical Report. arXiv preprint arXiv:2512.16776, 2025. URL https://arxiv.org/abs/2512.16776.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. URL https://arxiv.org/abs/2412.03603.

Klaus Krippendorff. Computing Krippendorff's alpha-reliability. Annenberg School for Communication, University of Pennsylvania, 2011. URL https://www.asc.upenn.edu/krippendorffs-alpha-reliability. Literature updated September 13, 2013.

Kuaishou Technology. Kling AI launches 3.0 model, ushering in an era where everyone can be a director. Official model announcement, 2026. URL https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-30-model-ushering-era-where-everyone-can-be.

Lightricks. LTX-2.3. Official model release, 2026a. URL https://huggingface.co/Lightricks/LTX-2.3.

Lightricks. LTX-2.5. Official model release; technical details in the LTX-2 report, 2026b. URL https://huggingface.co/Lightricks/LTX-2.5.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023. URL https://arxiv.org/abs/2210.02747.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, and Tat-Seng Chua. JavisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025. URL https://arxiv.org/abs/2503.23377.

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 (CVPR)*, 2024.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. arXiv preprint arXiv:2510.01284, 2025. URL https://arxiv.org/abs/2510.01284.

MiniMax. MiniMax-H3. Official technical release, 2026. URL https://minimax.cn/blog/minimax-h3.

OpenAI. Video generation models as world simulators. OpenAI technical report, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. URL https://arxiv.org/abs/2212.09748.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations (ICLR)*, 2025.

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. URL https://arxiv.org/abs/2508.16930.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-ASR Technical Report. *arXiv preprint arXiv:2601.21337*, 2026. URL https://arxiv.org/abs/2601.21337.

SII-GAIR, Sand.ai, Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, Lingzhi Li, Lyumanshan Ye, Min Hu, Qiangang Wang, Quanwei Qi, Steffi Chern, Tao Bu, Taoran Wang, Teren Xu, Tianning Zhang, Tiantian Mi, Weixian Xu, Wenqiang Zhang, Wentai Zhang, Xianping Yi, Xiaojie Cai, Xiaoyang Kang, Yan Ma, Yixiu Liu, Yunbo Zhang, Yunpeng Huang, Yutong Lin, Zewei Tao, Zhaoliang Liu, Zheng Zhang, Zhiyao Cen, Zhixuan Yu, Zhongshu Wang, Zhulin Hu, Zijin Zhou, Zinan Guo, Yue Cao, and Pengfei Liu. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. arXiv preprint arXiv:2603.21986, 2026. URL https://arxiv.org/abs/2603.21986.

SII-OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, Wenming Tu, Xiangyu Peng, Yang Gao, Yanru Huo, Ying Zhu, Yinze Luo, Yiyang Zhang, Yuerong Song, Zhe Xu, Zhiyu Zhang, Chenchen Yang, Cheng Chang, Chushu Zhou, Hanfu Chen, Hongnan Ma, Jiaxi Li, Jingqi Tong, Junxi Liu, Ke Chen, Shimin Li, Shiqi Jiang, Songlin Wang, Wei Jiang, Zhaoye Fei, Zhiyuan Ning, Chunguo Li, Chenhui Li, Ziwei He, Zengfeng Huang, Xie Chen, and Xipeng Qiu. MOVA: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026. URL https://arxiv.org/abs/2602.08794.

Tomáš Souček and Jakub Lokoč. TransNet V2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. URL https://arxiv.org/abs/2008.04838.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024. URL https://arxiv.org/abs/2407.14505.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu

Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing video generation for world complexity. arXiv preprint arXiv:2604.14148, 2026. URL https://arxiv.org/abs/2604.14148.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025. URL https://arxiv.org/abs/2502.05139.

Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. StableAvatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025. URL https://arxiv.org/abs/2508.08248.

Shuyuan Tu, Qi Tian, Zihan Yang, Yue Wu, Xintong Han, Weijie Kong, Jiangfeng Xiong, Jian-Wei Zhang, Zhao Zhong, Liefeng Bo, Zuxuan Wu, and Yu-Gang Jiang. Baton: Explicit semantic blueprints for joint video-audio generation. *arXiv preprint arXiv:2605.25195*, 2026. URL https://arxiv.org/abs/2605.25195.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

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

Hui Wang, Cheng Liu, Junyang Chen, Haoze Liu, Yuhang Jia, Shiwan Zhao, Jiaming Zhou, Haoqin Sun, Hui Bu, and Yong Qin. TTA-Bench: A comprehensive benchmark for evaluating text-to-audio models. *arXiv preprint arXiv:2509.02398*, 2025b. URL https://arxiv.org/abs/2509.02398.

Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, Zihan Li, Yuzhe Liang, Xiaopeng Wang, Haorui Zheng, Ming Wen, Kang Yin, Yiran Wang, Nan Li, Feng Deng, Liang Dong, Chen Zhang, Di Zhang, and Kun Gai. Kling-Foley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*, 2025c. URL https://arxiv.org/abs/2506.19774.

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. URL https://arxiv.org/abs/2406.07686.

Yujie Wei, Yujin Han, Zhekai Chen, Yongming Li, Kaixun Jiang, Zhihang Liu, Quanhao Li, Zhiwu Qing, Xiang Wang, Zhen Xing, Ruihang Chu, Lingyi Hong, Yefei He, Junjie Zhou, Junqiu Yu, Yang Shi, Difan Zou, Kai Zhu, Shiwei Zhang, Yingya Zhang, Yu Liu, Xihui Liu, and Hongming Shan. MSAVBench: Towards comprehensive and reliable evaluation of multi-shot audio-video generation. *arXiv preprint arXiv:2605.20183*, 2026. URL https://arxiv.org/abs/2605.20183.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 54015–54029, 2024.

Tianxin Xie, Wentao Lei, Kai Jiang, Guanjie Huang, Pengfei Zhang, Chunhui Zhang, Fengji Ma, Haoyu He, Han Zhang, Jiangshan He, Jinting Wang, Linghan Fang, Lufei Gao, Orkesh Ablet, Peihua Zhang, Ruolin Hu, Shengyu Li, Weilin Lin, Xiaoyang Feng, Xinyue Yang, Yan Rong, Yanyun Wang, Zihang Shao, Zelin Zhao, Chenxing Li, Shan Yang, Wenfu Wang, Meng Yu, Dong Yu, and Li Liu. PhyAVBench: A challenging audio physics-sensitivity benchmark for physically grounded text-to-audio-video generation. *arXiv preprint arXiv:2512.23994*, 2025. URL https://arxiv.org/abs/2512.23994.

Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. AudioTime: A temporally-aligned audio-text benchmark dataset. *arXiv preprint arXiv:2407.02857*, 2024. URL https://arxiv.org/abs/2407.02857.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report. arXiv preprint arXiv:2509.17765, 2025. URL https://arxiv.org/abs/2509.17765.

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

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. URL https://arxiv.org/abs/2502.03897.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. URL https://arxiv.org/abs/2503.21755.

Yang-Hao Zhou, Haitian Li, Rexar Lin, Heyan Huang, Jinxing Zhou, Changsen Yuan, Tian Lan, Ziqin Zhou, Yudong Li, Jiajun Xu, Jingyun Liao, Yi-Ming Cheng, Xuefeng Chen, Xian-Ling Mao, and Yousheng Feng. MTAVG-Bench: A diagnostic benchmark for multi-talker dialogue-centric audio-video generation. *arXiv preprint arXiv:2602.00607*, 2026a. URL https://arxiv.org/abs/2602.00607.

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. URL https://arxiv.org/abs/2604.08540.

A USE OF AI TOOLS

In this work, we used AI tools for the following tasks: reviewing and correcting the English grammar of the manuscript; surveying the literature on audio-video and benchmark evaluation to support the related-work analysis; typesetting and formatting the paper; checking reference correspondence; and drafting the Introduction and human-evaluation protocol. All AI-assisted content has been reviewed and verified by the authors prior to submission. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of AI tools.

B PROMPT-CONSTRUCTION PIPELINE AND TEMPLATES

CAST-BENCH's prompt suite is produced by a two-stage LLM pipeline. Figure 7 shows the overall data flow; the two system prompts driving Stages S1 and S2 are listed verbatim in sections B.1 and B.2. Boilerplate rules that are not discussed in the main paper are elided as "[. . .]".

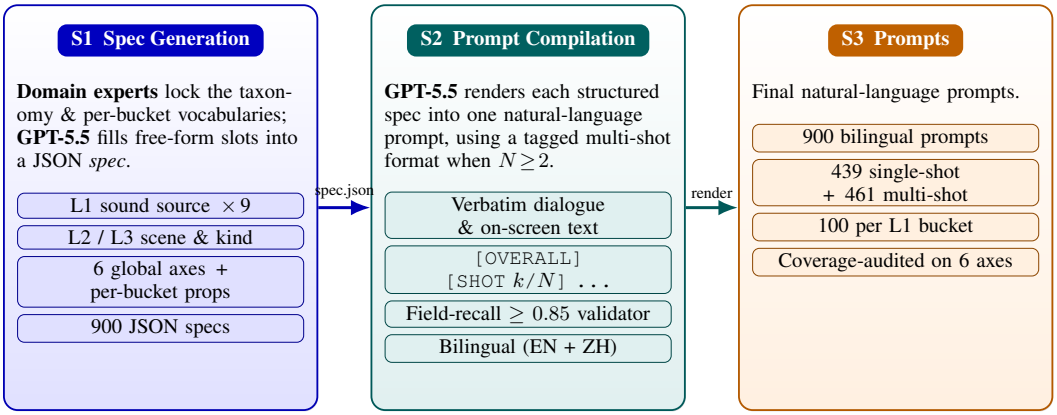


Figure 7: **Reverse-flow prompt-construction pipeline.** Domain experts lock the acoustic taxonomy and the per-bucket vocabularies (S1, blue); GPT-5.5 fills the free-form slots to produce 900 structured JSON *specs*; GPT-5.5 then renders each spec into a single natural-language prompt with a tagged multi-shot format (S2, teal); the final bilingual prompt set (S3, orange) will be released publicly.

B.1 S1 — SPEC-GENERATION SYSTEM PROMPT (GPT-5.5)

Given the L1 sound source (fixed by the caller) and the per-bucket vocabularies, GPT-5.5 emits one structured JSON *spec* per call. Vocabulary-controlled slots (`visual_style`, `video_type`, `camera_movement`, `audio_situation`, `sound_complexity`, `shot_count`, `speech_attributes.*`, ...) are enumerated in the system message; free-form slots (`scene_setting`, `subjects[*].descriptor`, `physics_phenomena`, camera details) are left to the model.

S1: Spec-Generation System Prompt (GPT-5.5)

```
You are a T2AV benchmark spec generator. Produce ONE JSON object
describing a single video--audio sample. Use ONLY values from the
provided vocabularies; do NOT invent enumeration values.

[Constraints]
 (1) `category` and its L1 bucket are fixed by the caller.
 (2) `shot_count` in {single_shot, multi_shot}. If multi_shot, `
shots[]` MUST contain EXACTLY N items, where N is provided by the
caller and 2 <= N <= 9.
```

```
 (3) The 6 global axes (`visual_style`, `video_type`, `
camera_movement`, `audio_situation`, `sound_complexity`, `shot_count
`) MUST equal the pre-allocated values.
 (4) When `audio_situation` == "off_screen", the sound emitter MUST
NOT be visible in the frame.
 (5) NARRATIVE CONTINUITY (HARD, multi-shot): all shots MUST tell
ONE coherent story; each shot picks up from the previous one.
 (6) Every `subjects[*].descriptor` MUST be >= 8 words with concrete
 visual attributes (age, clothing, material, lighting).
 (7) For speech buckets, `subjects[*].speech` or `shots[*].dialogue
[*].text` MUST be written in the script/language specified by `
speech_attributes.language`.
 [ ... boilerplate about JSON escaping / vocab lists ... ]

[Vocabulary]
 [[VOCAB_ENUMERATION]]        # injected at runtime

[Output] JSON only. No markdown, no commentary, no ```json fences.

Now produce the spec for:
 category = [[CATEGORY]]
 shot_count = [[SHOT_COUNT]]     (N = [[N_SHOTS]] if multi)
 pre-allocated axes = [[AXES]]
```

B.2 S2 — PROMPT-COMPILER SYSTEM PROMPT (GPT-5.5)

Each JSON spec is then handed to GPT-5.5, which *compiles* it into one natural-language prompt. Multi-shot specs use the tagged format `[OVERALL] ... [SHOT 1/`N`] ... [SHOT` N/N`]` so downstream generators can detect shot boundaries with a simple regex.

S2: Prompt-Compiler System Prompt (GPT-5.5)

```
You are a T2AV prompt compiler. Convert a structured spec JSON into
a single natural-language video-generation prompt.

Rules:
 (1) Embed every concrete element from the spec -- subjects,
audio_events, camera, scene_setting, and all top-level global axes (
visual_style, video_type, ...).
 (2) NEVER invent details not in the spec. ONLY describe what the
spec contains; do NOT add atmospheric flourish, similes, metaphors,
emotional commentary, or extra sound effects that the spec did not
ask for. Stay strictly factual.
 (3) Output ONLY the prompt text; no preamble, no JSON.
 (4) sound_complexity == single_source: the prompt MUST NOT mention
any sound source other than the primary audio subject.
 (5) audio_situation == off_screen: do NOT describe the visual
appearance of the off-screen emitter; describe only its audible
qualities.
 (6) DIALOGUE INTEGRITY (HARD): quote every line from `subjects[*].
speech` / `shots[*].dialogue` VERBATIM, in the original script;
never paraphrase, translate, abridge, or write "...". Every quoted
line MUST end with terminal punctuation.
 (7) ON-SCREEN TEXT: entries in `text_on_screen` MUST be rendered as
 visible text in the scene, never as spoken dialogue. Dialogue and
on-screen text are two disjoint sets.
 (8) EMOTION ARC (HARD, speech): when `speech_attributes.
emotion_change == true`, the prompt MUST describe a single emotional
 turn from `emotion_from` to `emotion_to`; the OVERALL header MUST
mention this arc in plain words.
```

```
 (9) Multi-shot format:
       [OVERALL] <one short sentence embedding all global axes>
       [SHOT 1/N, t1s, <camera>] <shot description> ...
       [SHOT N/N, tN s, <camera>] <shot description> Output one
continuous paragraph -- no bullet points.
 [ ... boilerplate about length limits / language ... ]

Now compile the following spec:
 [[SPEC_JSON]]
```

B.3 PROMPT-SUBCLASS STATISTICS

This section gives the fine-grained subclass statistics omitted from the main text. Percentages are rounded to one decimal place. For L2, the denominator is the number of valid L2 event occurrences in the corresponding L1 bucket, so a prompt containing several audio events contributes to several L2 counts. For L3, we count only the first valid event of the corresponding L1 bucket (the designated main event), giving a denominator of 100 prompts per bucket. Categories without an L3 vocabulary use their L2 subclasses as the finest event-level partition.

Global-axis marginals. The suite mixes 439 single-shot prompts (48.8%) with 461 multi-shot prompts (51.2%), the latter spanning two to nine ordered shots; four-shot (153, 17.0%) and five-shot (150, 16.7%) narratives are the largest groups, followed by three-shot (62, 6.9%), two-shot (35, 3.9%), six- and seven-shot (18 each, 2.0%), eight-shot (17, 1.9%), and nine-shot (8, 0.9%) cases. On source binding, 541 prompts (60.1%) contain multiple sounding subjects, 288 (32.0%) contain one visible sounding subject, and 71 (7.9%) use an off-screen emitter; independently, 630 prompts (70.0%) contain multiple sound sources and 270 (30.0%) a single source. Visual style is realistic in 612 prompts (68.0%), historical 97 (10.8%), sci-fi 72 (8.0%), punk 49 (5.4%), fantasy 43 (4.8%), and xuanhuan 27 (3.0%). Video type is film/TV 325 (36.1%), vlog 282 (31.3%), 3D animation 107 (11.9%), 2D animation 88 (9.8%), games 58 (6.4%), and advertisements 40 (4.4%). Speech spans nine languages—English 40 (40.0%), Chinese 39 (39.0%), Russian 5 (5.0%), French, German, and Japanese 4 each (4.0%), Spanish 2 (2.0%), and Arabic and Korean 1 each (1.0%). Emotion classes are joyful 23 (23.0%), confused 19 (19.0%), angry 18 (18.0%), neutral and panicked 11 each (11.0%), sad 10 (10.0%), and surprised 8 (8.0%); 38 prompts (38.0%) additionally specify an emotion change, and 168 prompts (18.7%) require on-screen text.

Speech. L2 occurrences (184): daily dialogue 93 (50.5%), broadcast 43 (23.4%), commentary 29 (15.8%), voiceover 18 (9.8%), and singing 1 (0.5%). The main-event L3 distribution (100): film commentary 8 (8.0%), press conference 8 (8.0%), radio broadcast 8 (8.0%), tour guide 8 (8.0%), casual conversation 7 (7.0%), documentary narration 7 (7.0%), film voiceover 7 (7.0%), interview 7 (7.0%), lecture 7 (7.0%), meeting 7 (7.0%), platform livestream 7 (7.0%), quarrel 7 (7.0%), news broadcast 6 (6.0%), and sports commentary 6 (6.0%). Voice timbre is female 33 (33.0%), male 23 (23.0%), broadcast 12 (12.0%), elderly 12 (12.0%), middle-aged 12 (12.0%), robotic 5 (5.0%), and child 3 (3.0%). Languages are English 40 (40.0%), Chinese 39 (39.0%), Russian 5 (5.0%), French 4 (4.0%), German 4 (4.0%), Japanese 4 (4.0%), Spanish 2 (2.0%), Arabic 1 (1.0%), and Korean 1 (1.0%). Emotions are joyful 23 (23.0%), confused 19 (19.0%), angry 18 (18.0%), neutral 11 (11.0%), panicked 11 (11.0%), sad 10 (10.0%), and surprised 8 (8.0%); 62 prompts (62.0%) have no marked emotion change and 38 (38.0%) contain a change.

Non-verbal vocalization. L2 occurrences (227): physiological 103 (45.4%), emotional 96 (42.3%), and mimicry 28 (12.3%). The main-event L3 distribution (100) is cheer, cough, cry, fart, hiccup, laugh, moan, scream, smirk-laugh, sneeze, sniffle, snore, tongue click, whistle, and yawn at 6 each (6.0%), followed by beatbox and sigh at 5 each (5.0%).

Animal. L2 occurrences (203): bird call 52 (25.6%), insect sound 45 (22.2%), wild beast roar 34 (16.7%), livestock 29 (14.3%), pet vocalization 22 (10.8%), and amphibian 21 (10.3%). The main-event L3 distribution (100) is bear, chicken, dog, elephant, mosquito, pig, and snake at 6 each

(6.0%); cat, cricket, crow, cuckoo, duck, eagle, frog, horse, tiger, and wolf at 5 each (5.0%); and bee and cicada at 4 each (4.0%).

Object. L2 occurrences (215): interaction 104 (48.4%), transportation 67 (31.2%), and weapon 44 (20.5%). The main-event L3 distribution (100) is car, collision, dart, motorcycle, and tear at 6 each (6.0%); airplane, arrow, bicycle, bomb, friction, knife, liquid, missile, pistol, police car, rifle, shatter, ship, and train at 5 each (5.0%).

Action. L2 occurrences (176): daily activity 108 (61.4%) and sports 68 (38.6%). The main-event L3 distribution (100) is fancy diving and martial-arts fighting at 11 each (11.0%), basketball layup contest at 9 (9.0%), sword fighting at 8 (8.0%), wrestling at 7 (7.0%), unarmed fighting at 5 (5.0%), boxing, clapping, knocking, stick fighting, and swimming race at 4 each (4.0%), basketball, gymnastics, running, soccer, table tennis, and walking at 3 each (3.0%), billiards, chopping vegetables, eating, soccer dribbling, and typing at 2 each (2.0%), and parkour at 1 (1.0%).

Ambient sound. There is no separate L3 layer. L2 occurrences (289) are weather 81 (28.0%), urban street 74 (25.6%), water body 45 (15.6%), crowd/applause 34 (11.8%), forest/field 31 (10.7%), and indoor ambience 24 (8.3%).

Instrument. There is no separate L3 layer. L2 occurrences (157) are piano 71 (45.2%), string instrument 56 (35.7%), percussion 16 (10.2%), synthesizer 7 (4.5%), wind instrument 7 (4.5%), and ethnic instrument 0 (0.0%). Among the 98 prompts with an annotated playing mode, 53 are solo (54.1%) and 45 are ensemble (45.9%); two prompts have no mode annotation.

Background music. There is no separate L3 layer. L2 occurrences (144) are film score 91 (63.2%), vlog BGM 31 (21.5%), and game music 22 (15.3%). The music-kind attributes are instrumental 43 (43.0%), song 30 (30.0%), and rhythmic beat 27 (27.0%).

Sound effect. There is no separate L3 layer. L2 occurrences (341) are exaggerated SFX 86 (25.2%), UI notification 81 (23.8%), transition SFX 67 (19.6%), sci-fi SFX 58 (17.0%), magic-skill SFX 36 (10.6%), and ambience SFX 13 (3.8%).

C General-Model Evaluation Details

For every VLM-graded metric we materialise spec-grounded questions or score prompts at benchmark-build time; at evaluation time these are handed to the judge model together with the generated clip. Binary metrics use yes/no answers, whereas style, overall audio alignment, and speaker identity use explicit integer scales. This appendix lists, for each metric, one representative spec together with (i) the exact questions we ask the judge, and (ii) the corresponding grader-rubric prompt that defines the judge's evidence and scoring rules. Boilerplate is elided as "[. . .]".

C.1 Audio Generation

Audio Generation asks whether every audible sound described in the prompt is actually rendered in the generated video, one yes/no question per sound source, plus one overall 1–5 alignment question. Questions are written per shot and must never invent sounds that the prompt does not describe.

Prompt & Questions: Audio Generation

```
------ Prompt ------ A gritty urban basketball court at dusk, chain-
link fences and graffiti walls surrounding cracked asphalt with
faded lines. A male player in a red jersey and black shorts sweats
as he drives for a layup, sneakers screeching on pavement, ball
bouncing rhythmically before release. A female defender in a blue
jersey leaps to block, body tensing mid-air. A male referee in a
striped shirt watches intently from the sideline, whistle dangling.
The digital scoreboard above the hoop reads HOME 78 - GUEST 76 with
```

```
12.0 seconds left; a banner at the bottom displays GAME ON THE LINE.
 Distant crowd murmurs, a car engine idles faintly, and a siren
wails far away. Body
movements produce natural sounds: dribble thumps, sneaker friction,
jump impact, net swish. No camera movement -- fixed frame captures
the physics of momentum, gravity, collision, and friction in real-
time.

------ Questions ------
Q1.  Is there a sound of sneakers screeching on pavement in the
video?
Q2.  Is there a sound of a basketball bouncing rhythmically in the
video?
Q3.  Is there a sound of a dribble thump in the video?
Q4.  Is there a sound of sneaker friction in the video?
Q5.  Is there a sound of a jump impact in the video?
Q6.  Is there a sound of a net swish in the video?
Q7.  Is there a sound of distant crowd murmurs in the video?
Q8.  Is there a sound of a car engine idling faintly in the video?
Q9.  Is there a sound of a siren wailing far away in the video?
Q10. How well does the audio in the video match the prompt?     [1-5
 rubric]
```

Grader Rubric: Audio Generation

```
You are a strict audio judge for one short, single-shot clip. The
original
generation prompt is the ground truth. Use the visuals only to
identify a
sound source; judge the audible soundtrack itself.

There are two question types. For a specific sound or speaker
question,
answer yes only when the requested sound is clearly audible and
consistent
with the prompt; otherwise answer no. For a speaker question, also
require
the same gender, age, role, language, content, and visible mouth
alignment,
and do not merge two prompt-described speakers. Silent or non-
speaking clips
are no for speaker questions. Output one object per numbered
question:
{"reason":"brief audible evidence","answer":"yes"|"no"}.

The exact overall question is: "How well does the audio in the video
 match
the prompt?" Score it from 1 to 5 using the following rubric:
  5  Every required sound is present, correctly timed and balanced,
with no
     extraneous or distorted audio.
  4  All major sounds are present, with only a minor audible flaw.
  3  Some required sounds are present but coverage or timing is
incomplete.
  2  The soundtrack is mostly inconsistent with the prompt.
  1  Audio is wrong, silent, unintelligible, or contradicts the
prompt.
Use 5 only for flawless audio; any faint, late, missing, distorted,
or extra
sound forces a score of at most 4, and a silent or contradictory
track scores
```

```
1. For this question output {"reason":"heard audio and score
justification",
"score":1--5}. Reasons must describe what is heard rather than
restating the
prompt or relying on visual evidence. Return only the required JSON
object or
array, with no extra keys or prose.
```

C.2 EMOTION MATCH

Emotion Match probes whether the vocal emotion in a speech clip matches the spec, and—when the spec declares an emotion *change*—whether the on-screen facial expression follows the same arc. Questions are emitted at the spec level (not per shot) since emotion labels are global.

Prompt & Questions: Emotion Match

```
------ Prompt ------
[OVERALL] Realistic vlog-style multi-shot sequence capturing a
corporate meeting's emotional arc from angry to confused, featuring
multi-subject dialogue with broadcast-tone English speech, fixed-
frame shots, and a cartoon coffee mug confessing a system error.

[SHOT 1/3, medium_shot] A middle-aged male executive in a dark suit,
 stern expression, slams his fist while pointing at a laptop screen
showing red financial charts. He leans forward, jaw tight, eyes
blazing as he delivers, "This quarter's numbers are unacceptable --
we're hemorrhaging revenue and no one has an explanation! I want
answers by noon or heads will roll." A young female analyst fidgets
with her pen; a male engineer leans against the wall. A cartoon-
style animated coffee mug blinks its eyes slowly.
Audio: only the executive's voice, continuous and forceful.

[... SHOT 2/3 and SHOT 3/3 elided;
     emotion arc completes: angry -> confused ...]

------ Questions ------
Q1.  Does the speaking voice in the video start with an angry
emotion?
Q2.  Does the speaking voice in the video transition from an angry
emotion to a confused emotion?
Q3.  Does the on-screen character's facial expression shift in a way
 that matches the vocal emotional transition?
```

Grader Rubric: Emotion Match

```
You are a strict audio-visual emotion judge. For every question,
return only
{"reason":"concrete evidence in at most 20 words","answer":"yes"|"no
"}.

Global override: if no human speech is detectable (silence, music,
ambience,
sound effects, non-verbal vocalizations, or speech too faint or
unintelligible),
answer no for every question.

Question types are modality-specific:
```

```
  A. "Is the speaking voice in the video expressing a <emotion>
emotion?"
     Judge the audible speech only. If several people speak, one
clearly
     matching speaker is sufficient.
  B. "Do the on-screen characters visually display a <emotion>
emotion?"
     Judge frames only; at least one visible person must show the
requested
     expression. No person, an unclear face, or the wrong emotion is
 no.
  C. "Are there no people visible in the video?" Judge frames only;
answer
     yes only when no human or humanoid appears anywhere in the clip
.

Do not transfer evidence between modalities. Broken or undecodable
frames make
the visual questions no. Return one object or an array matching the
numbered
questions, with no extra fields or prose.
```

C.3 SPEAKER ID MATCH

Speaker ID Match evaluates whether the generated clip preserves the identity of every speaking subject across turns and shots, and whether each utterance is assigned to the correct speaker. The implementation uses two stages: Qwen3-Omni first produces an ordered speaker-aware transcript, then a rubric-based judge compares that transcript with the ground truth prompt and subject table. The two stages below are faithful English renderings of the prompts used by the evaluation code; the executable files will be provided with the benchmark release.

Qwen3-Omni System Prompt (Stage 1)

```
You will see a video containing people speaking, including both
visual frames and audio. Transcribe ALL audible spoken dialogue and
narration in temporal order. Annotate every segment with:
 - speaker: a speaker ID (S1, S2, ...)
 - subject: a short description of the speaking subject (appearance
or ``off-screen voice'')
 - text: the spoken content of that segment

Speaker-ID rules:
 1. The first subject to speak is S1. Off-screen narration,
voiceover, and inner monologue count as speakers; if one speaks
first, it is S1.
 2. Assign S2, S3, ... to newly appearing speakers in order of first
 speech.
 3. When a speaker speaks again, reuse exactly the ID assigned at
first appearance.
 4. Off-screen narration is an independent speaker and must not be
merged with an on-screen character.
 5. Retain meaningful singing lyrics and meaningful vocalizations (
for example, ``ah'', ``um'', or ``hey''). Do not transcribe
instrumental background music or pure environmental sound.

Subject-description rules:
 1. For off-screen narration, voiceover, or inner monologue, use one
 fixed ``off-screen voice'' label.
```

```
 2. For visible people, use a brief description containing gender,
age range, and salient appearance or clothing; use at most 15
Chinese characters in the executable prompt.
 3. The subject string must remain identical for repeated segments
from the same speaker.
 4. The executable prompt uses Chinese for the subject description
even when the spoken language is English, French, Japanese, or
another language.
 5. If appearance cannot be determined, use a consistent fallback
such as ``unknown male voice'', ``unknown female voice'', or ``
unknown speaker''.

Output format (strict): output exactly one JSON object, with no
Markdown fences, explanation, timestamps, or extra fields:
{"segments":[{"speaker":"S1","subject":"off-screen voice","text
":"..."},{"speaker":"S2","subject":"middle-aged man in a white shirt
","text":"..."}]}

Additional requirements:
 - speaker must be a string of the form S1, S2, S3, ... .
 - subject must be a non-empty string following the rules above (
Chinese in the executable prompt).
 - text must contain only the spoken dialogue; do not emit [Music],
[Noise], or [Unintelligible] tags.
 - segments must be ordered by the actual time of occurrence in the
video.
 - If no one speaks, output {"segments":[]}.
```

Grader Rubric: Speaker ID Match

```
You are a strict judge for speaker-ID consistency. You will receive
two JSON objects:

GT contains video_id, shot_count, prompt_text, and subjects.
prompt_text is the original English generation prompt and is the
highest-priority ground truth for speaker identity and quoted lines.
 subjects is an ordered table of {id, descriptor, gender}; off-
screen narrators and anthropomorphic characters with their own lines
 are independent subjects. HYP contains segments [{speaker, subject,
 text}] transcribed from the generated video in temporal order.

HYP assigns S1 to the first subject that actually speaks, then
assigns new IDs in order of first speech; repeated speakers must
retain their original IDs. Before scoring, internally extract the
ordered (speaker descriptor, expected line) list from prompt_text,
map it to GT subject IDs, and compare it with HYP. A single subject
may have multiple lines; merging or splitting those lines is
acceptable when attribution and wording remain essentially correct.

Score ONLY these two dimensions independently, each from 1 to 5 (5 =
 fully consistent, 1 = completely incorrect):

General strict-scoring rules:
 - Start from 3; increase only with explicit evidence and decrease
for identifiable errors.
 - Any key-attribute error (gender, off-screen versus on-screen
status, speaker count, line attribution, a missing complete line, or
 translation into another language) caps the score at 3.
 - Use 5 only when errors are essentially unnoticeable; use 4 only
for one minor wording imperfection. When uncertain between 3 and 4,
use 3.
```

```
 - Do not raise scores because the model appears to have made a good
 effort. The two dimensions are independent.

A. subject_consistency -- compare every HYP speaker ID with the same
-numbered GT subject.
  5: ID count is exactly correct and every subject description
identifies the same entity with all key attributes correct.
  4: ID count and all key attributes are correct, with only one
minor wording imprecision.
  3: one key attribute is wrong, one speaker is missed or added, two
 or more descriptions are materially imprecise, or a cartoon/animal
speaker is not treated as independent.
  2: at least two IDs are mismatched, IDs are globally swapped, or
at least two speakers are missed or hallucinated.
  1: ID count and subject descriptions are almost entirely unrelated
 to GT.

B. line_assignment -- compare quoted lines in GT prompt_text (and
dialogues, if present) with HYP segments.
  Tolerate only punctuation, case, spacing, or one-to-two-character
transcription slips. Penalize added or removed content exceeding one
 third of a line, changed predicates or key nouns, missing complete
lines, hallucinated lines, translation into another language, or
incorrect order / speaker attribution.
  5: all lines are attributed correctly and are nearly verbatim,
with no omissions, hallucinations, translations, or order errors.
  4: attribution is correct and there is exactly one noticeable but
limited wording deviation, or one harmless merge/split.
  3: roughly one third of lines are misattributed, one complete line
 is missing or hallucinated, one line has a severe wording change,
or part of the order is wrong.
  2: more than half of lines are misattributed, at least two
complete lines are missing or hallucinated, the whole transcript is
translated, or wording is extensively corrupted.
  1: almost no line is both correctly attributed and textually
matched to GT.

Output exactly one JSON object and no additional text:
{
  "subject_consistency": <int 1-5>,
  "line_assignment": <int 1-5>,
  "errors": ["short error description", ...],
  "comment": "one-sentence overall assessment"
}
The errors list must contain concise phrases (at most 30 Chinese
characters in the executable prompt). If HYP is missing, empty, or
unparseable, assign 1 to both dimensions and write ``HYP missing or
unparseable'' in errors.
```

C.4 STYLE MATCH

Style Match pairs a `visual_style` question with a `video_type` question, both of them video-level and vocabulary-bound. `visual_style` takes one of {sci_fi, fantasy, realistic, punk, historical, xuanhuan}; `video_type` takes one of {2d_animation, 3d_animation, game, film_tv, vlog, advertisement}.

Prompt & Questions: Style Match

```
------ Prompt ------ A realistic 2D animation American cartoon style
 of a dimly lit jazz club with wooden floors and velvet curtains,
```

```
featuring a polished black wood grand piano with ivory keys and open
 lid center stage, reflecting warm spotlights. Behind it, a middle-
aged male bass player sways while plucking strings of his double
bass, and a middle-aged female saxophonist leans into the mic, eyes
closed, blowing into her saxophone. The camera follows in a medium
shot from a slightly low angle, capturing the ensemble's
synchronized playing motion and rhythmic tempo cues. A neon sign
above the bar reads JAZZ NIGHT, and a banner at the bottom displays
Live at 8PM. Audio blends continuous piano notes, repeated plucked
bass strings, and sustained saxophone tones, layered with distant
clinking glasses and murmured chatter. No spoken dialogue.

------ Questions ------
Q1. Is the visual style of this video realistic?
Q2. Is this video a 2D animation in American cartoon style?
    [each question is scored on a 0--5 scale]
```

Grader Rubric: Style Match

```
You are a strict visual-style judge. The clip is evaluated without
using audio.
Each question is scored from 0 to 5 and returns only
{"reason":"concrete visual evidence in at most 25 words","score
":0--5}.

Questions beginning "Is the visual style of this video ...?"
evaluate the
world and subject matter (realistic, sci-fi, fantasy, xuanhuan,
historical, or
punk), not rendering technique. Questions beginning "Is this video
...?"
evaluate production or rendering type (2D animation, 3D animation,
video
game, film/TV, vlog, or advertisement). A vlog must show
characteristic
irregular handheld movement; do not use evidence from one question
to answer
the other.

Use this strict scale for each question independently:
  5 world-class reference-quality match with no meaningful weakness;
  4 clear match with only minor flaws;
  3 mostly matches but has visible problems;
  2 weak or partial match;
  1 barely hints at the requested style or type;
  0 completely wrong, unwatchable, or without decodable frames.
Return one object or an array matching the numbered questions, with
no extra
keys or prose. If no frame can be decoded, every score is 0.
```

C.5 Subject Correctness

Subject Correctness asks, for every VISUAL subject positively described in the prompt, one yes/no question of the fixed template `Does this video contain {subject description}?`. The description is restricted to static visual attributes only.

Prompt & Questions: Subject Correctness

```
------ Prompt ------
[OVERALL] Realistic vlog-style multi-shot sequence tracking a
cricket's dusk journey from garden to living room, with layered
animal vocalizations (cricket chirping, cat hiss, crow caw, mouse
squeak) and handheld tracking camera movements across three distinct
 environments.

[SHOT 1/3, close-up + tracking] Handheld macro tracking shot focuses
 on a small brown cricket perched on a leaf, its chitinous
exoskeleton catching fading dusk light, long antennae twitching as
it chirps continuously. A gray tabby cat, striped fur low to the
ground, stalks silently nearby, tail flicking with predatory tension
. The cricket's open mouth vibrates with each chirp. A large black
crow perched on a lamppost emits repeated harsh caws, beak wide open
, glossy feathers ruffled. From under the porch step, a small brown
mouse squeaks once -- tiny ears twitching, furry tail darting.

[... SHOT 2/3 and SHOT 3/3 elided ...]

------ Questions ------
Q1. Does this video contain a small brown cricket with a chitinous
exoskeleton and long antennae?
Q2. Does this video contain a gray tabby cat with striped fur and a
flicking tail?
Q3. Does this video contain a large black crow with glossy feathers
and an open beak?
Q4. Does this video contain a small brown mouse with twitching ears
and a furry tail?
```

Grader Rubric: Subject Correctness

```
You are a strict visual-subject judge for one or more questions of
the form
"Does this video contain <SUBJECT_DESCRIPTOR>?". Return only
{"reason":"concrete visual evidence in at most 25 words","answer":"
yes"|"no"}.

Answer yes only when one visible subject matches every load-bearing
attribute:
category or species, gender or age, role, primary clothing type and
color,
face or body markers, key prop, stylistic qualifier, and any
explicitly stated
posture or action. Do not silently substitute a similar subject. A
matching
subject may be anywhere in the frame; background statues,
photographs,
reflections, tiny distant figures, or non-visible subjects do not
count. When
multiple subjects are described, one fully matching subject is
sufficient for
that question.

Use only visual evidence and answer no when frames are black, static
,
corrupted, or undecodable. Doubt is no. Return one object or an
array matching
the numbered questions, with no extra fields or prose.
```

C.6 MOTION PLAUSIBILITY

Motion Plausibility uses five visual yes/no questions per shot, covering naturalness, physical plausibility, temporal coherence, action completeness, and subject integrity. The final reported metric is computed from this binary shot-level evaluation; an auxiliary full-video humanoid score is retained for diagnosis only.

Prompt & Questions: Motion Plausibility

```
------ Prompt ------ A handheld vlog-style medium shot at eye level,
 fixed frame, captures a young male basketball player in a white t-
shirt and blue shorts energetically dribbling and shooting a
standard orange rubber basketball on a sunlit outdoor court. The
asphalt surface, metal hoop with white net, and chain-link fence are
 visible. Each dribble produces a sharp rhythmic bounce sound; the
ball's impact with the ground and net generates kinetic energy
transfer and air resistance effects. Audio includes the repeated
basketball bounce (associated with the player), continuous ambient
city sounds and distant car horns, and a single burst of sound as
the ball hits the hoop. No dialogue. Visual style is realistic. The
scene emphasizes body movement and its natural audio by-product --
no camera movement, only subject motion.

------ Questions ------
Q1. Is the motion in the video natural and not stiff?
Q2. Is the motion in the video physically plausible, without any
abnormal or impossible movements?
Q3. Is the motion in the video temporally coherent and smooth,
without jitter, sudden jumps, or flicker?
Q4. Is the action process in the video complete, with a clear
beginning and end, and not abruptly cut off?
Q5. Is the moving subject in the video structurally intact
throughout the motion, without unreasonable splitting, merging, or
severe deformation?
```

Grader Rubric: Motion Plausibility

```
You are a strict visual-motion judge. Answer each of the five
questions with
one JSON item {"reason":"concrete visual evidence","answer":"yes"|"
no"}.
Judge only what is visible in the clip; do not use audio, scenery,
lighting,
clothing, camera motion, or text as evidence. A faithful answer is
yes only
when the relevant motion is clearly natural, physically possible,
temporally
smooth, complete, and structurally intact. If the motion is absent,
ambiguous,
or the frames are missing/corrupted, answer no. Return one object or
 an array
in question order, with no extra keys or prose.
```

D METRIC EVALUATION PROCEDURES

This appendix describes the models, observable evidence, and score construction used by the 18 metrics in the benchmark. We report single-shot and multi-shot results separately whenever a metric is defined for both settings.

D.1 Visual Performance

Video Aesthetics. We use the Q-Align OneAlign video scorer to assign each generated video a continuous perceptual-quality score. The model evaluates visual clarity, composition, detail, naturalness, and overall aesthetic quality directly from the video, without using the generation prompt. The reported score is the arithmetic mean over valid videos in each shot-count split.

Motion Anomaly. This metric follows the VBench2 Human-Anatomy pipeline. YOLO-World detects humans in every frame and then locates faces and hands inside the human boxes; three SimMIM-finetuned ViT anomaly classifiers process the human, face, and hand crops. A person is anomalous if any of the three classifiers fires. For a video with T frames, the score is

$$1 - \frac{1}{T} \sum_{t=1}^{T} \frac{N_t^{\text{abnormal}}}{N_t^{\text{person}}},$$

so larger values indicate more anatomically plausible motion.

Motion Plausibility. Qwen3-VL-235B-A22B-Instruct answers five binary questions for every shot, covering naturalness, physical plausibility, temporal coherence, action completeness, and subject integrity. The judge is instructed to use visible motion only and to ignore audio, scenery, clothing, camera motion, and text. We convert yes/no answers to zero/one, average them within a shot, average shots within a multi-shot spec, and finally average valid specs.

Text Rendering. PaddleOCR (PP-OCR) detects and recognizes text in sampled frames. Detections are merged within a frame and linked across frames by overlap to form text tracks. Each track is compared with the prompt-specified target text after Unicode, case, and punctuation normalization; the best SequenceMatcher ratio r is combined with OCR confidence c as $r(0.5 + 0.5c)$. The video score is the mean best-match score over all required text strings.

D.2 Content Consistency

Background Consistency. SAM2.1 Hiera-Large produces and propagates a subject mask independently within each shot. We retain the background, suppress the subject, and extract frame features with CLIP ViT-B/32. For each frame we average the cosine similarity to the first frame and to the preceding frame; shot and spec scores are frame-weighted means of these values.

Subject Consistency. Using the same SAM2 masks, we retain the subject and suppress the background, then extract features with DINOv2 ViT-B/14. The frame stability score is

$$s_t = \frac{1}{2} \left[\cos(f_t, f_1) + \cos(f_t, f_{t-1}) \right].$$

Higher values indicate smaller appearance drift of the subject. The scores are aggregated over valid frames, shots, and specs.

Face Consistency. InsightFace buffalo_l detects faces and extracts identity embeddings frame by frame. Detections are linked into face tracklets using bounding-box IoU and embedding cosine similarity; tracklet embeddings are weighted by detection confidence and relative face area, and DBSCAN clusters tracklets into identities. The final score is on a 0–100 scale and is reported to three decimal places. It combines identity-count agreement (40 percent) and identity stability (60 percent), including tracklet-level feature stability and identity switching.

Subject Correctness. Qwen3-VL-235B-A22B-Instruct answers one visual yes/no question for each prompt-described subject. A positive answer requires the requested category or species, age or gender, clothing, key appearance markers, props, and explicitly stated pose or action to be simultaneously visible; uncertainty or an undecodable frame is scored as no. We compute the correct-question rate per spec and then average over specs within each shot-count split.

Camera Match. TransNetV2 receives 27 by 48 RGB frames and predicts a hard-cut probability for each frame. Thresholding at 0.5 and grouping successive cut boundaries yields the predicted shot count A, which is compared with the annotated count E. The per-video score is

$$\max\left(0, 1 - \frac{|A - E|}{E}\right),$$

and the metric is reported for multi-shot samples.

Style Match. Qwen3-VL-235B-A22B-Instruct scores two video-level questions on a 0–5 Likert scale. One question evaluates the visual world or subject-matter style (e.g., realistic, fantasy, historical, or sci-fi), while the other evaluates the production or rendering type (e.g., 2D or 3D animation, game, film/TV, or vlog). The two scores are summed to obtain a 0–10 rubric score and divided by 10 to obtain the 0–1 value reported in Table 2. These reported values are then averaged over the applicable samples.

D.3 Audio and Semantic Alignment

Audio Aesthetics. Meta AudioBox Aesthetics predicts four perceptual axes from the extracted audio: Content Enjoyment, Content Usefulness, Production Complexity, and Production Quality. We use Production Quality (PQ), on its 1–10 scale, as the benchmark audio-aesthetics metric and average it over valid videos.

Sound-Source Match. The Synchformer model (Iashin et al., 2024) trained on AudioSet jointly analyzes the video and audio and predicts a continuous temporal offset between a sound and its corresponding visual event. We retain samples with an identifiable sound source and report the mean absolute offset; smaller values indicate better synchrony.

Lip Sync (Offset). SyncNet combines S3FD face detection with audio and mouth-region visual features. Within temporal windows it finds the offset that maximizes audio–visual similarity. We report the mean absolute offset in frames (one frame is approximately 40 ms at 25 fps), with smaller values indicating better alignment.

Lip Sync (Match). The same S3FD and SyncNet inference also produces a maximum audio–visual matching confidence. Unlike the offset metric, this measure rewards a strong correspondence between phonetic audio features and mouth motion; larger values are better. Videos without a reliable face or mouth track do not contribute to the valid mean.

Audio Generation. Qwen3-Omni-30B-A3B-Instruct evaluates whether prompt-described speech, animal sounds, ambience, object sounds, action sounds, instruments, and music are actually audible. Each specified event is evaluated with a yes/no question, together with an overall audio–prompt alignment question scored from 1 to 5. The judge must use the audio track as evidence rather than inferring a sound from the visual action; event correctness and the overall score are normalized and aggregated first per shot and then per spec.

Emotion Match. Qwen3-Omni-30B-A3B-Instruct evaluates audio and visual emotion with separated modalities. Speech-emotion questions receive audio only and test the emotion category, initial emotion, or transition; facial-expression questions receive video only and test visible emotion or the presence of people. Answers are binary, with a within-spec consistency bonus when all questions are correct, followed by normalization and averaging over specs.

Speaker ID Match. This metric uses two multimodal stages. Qwen3-Omni-30B-A3B-Instruct first transcribes dialogue and narration in temporal order and assigns speaker IDs by order of first speech, reusing each ID for later turns. Qwen3-VL-235B-A22B-Instruct then compares the transcript with the original prompt and subject table, assigning independent 1–5 scores for subject identity consistency and line-to-speaker assignment, including speaker count, identity attributes, order, and wording. The two dimension scores are summed for each video, giving a nominal 2–10 score, and these sums are then averaged over samples.

Line Correctness. Qwen3-ASR-1.7B (Shi et al., 2026) transcribes the generated audio. The reference is formed by concatenating annotated dialogue in shot order, excluding on-screen text. After case, punctuation, and language-appropriate token normalization, we compute WER with jiwer and convert it to

$$s = \max\left(0, 1 - \min(\mathrm{WER}, 1)\right).$$

The model score is the mean sample accuracy, so lower WER yields a higher benchmark score.

D.4 Aggregation Convention

Binary metrics use the fraction of positive or correct answers; Likert metrics are aggregated in their native scale before any normalization; and offset metrics retain the lower-is-better direction in per-metric reporting. Scores are first computed at the video, shot, or spec level as appropriate, then averaged separately for single-shot and multi-shot samples. The three metric groups above provide the visual-performance, content-consistency, and audio–semantic-alignment components used in the benchmark summary.

For the overall score, all valid metrics are mapped to a common $[0, 100]$ higher-is-better scale, reversing the two offset metrics. We place additional emphasis on five metrics: Face Consistency (FcC), Audio Aesthetics (A-PQ), Audio Generation (AuG), Speaker ID Match (SpID), and Line Correctness (ASR), each with a weight of 10%. The remaining valid metrics collectively account for the other 50%.

E HUMAN EVALUATION PROTOCOL

Participants and procedure. We sample 50 generated videos per model for human evaluation. Three annotators independently score each generated video against its prompt, structured specification, and applicable evaluation questions. The evaluation covers different sound categories, generators, and both single-shot and multi-shot settings. Before scoring, the annotators review the same metric definitions and example cases to establish a shared understanding of the rubric. Videos are presented in randomized order; annotators may replay them and inspect individual shots. Scores are recorded independently, without consulting other annotators or automatic judgments.

Scoring rules. Human judgments follow the metric definitions in Appendices C and D. Motion Plausibility, Subject Correctness, sound-event presence, and Emotion Match use binary responses: 1 when the requirement is satisfied and 0 otherwise. Style Match uses the existing 0–5 rubric, while overall audio alignment and the two Speaker ID Match components use 1–5 ratings. For perceptual properties assessed by specialist models, annotators rate the corresponding quality, consistency, or synchrony on a five-point scale: 1 indicates a clear failure, 2 major defects, 3 partial fulfillment, 4 minor defects, and 5 full fulfillment without perceptible defects. Each judgment concerns only the specified property. Inapplicable questions are excluded; an explicitly requested but missing event is scored as a failure rather than marked inapplicable.

Evidence requirements. Visual properties are assessed from muted video, and audio content and vocal emotion from the audio track. Audiovisual synchrony and visible speaker attribution use both modalities. Annotators verify audible evidence instead of inferring sounds from visual actions. For multi-shot videos, they inspect the full sequence and relevant shots, checking event timing, dialogue assignment, and identity continuity across cuts. Short evidence notes identify the relevant event or timestamp when a requirement is violated.

Aggregation and agreement. Question and shot scores are aggregated within each video using the corresponding metric's rule, and the three annotators' scores are averaged to obtain the human reference. Spearman correlation compares automatic scores with these human references separately for single-shot and multi-shot videos, using average ranks for ties and reversing the direction of offset metrics. Inter-rater reliability is computed from the original independent ratings using Krippendorff's α, with nominal distance for binary responses and ordinal distance for ordered scales (Krippendorff, 2011). Across all rated items, Krippendorff's $\alpha = 0.82$, indicating good agreement among the three annotators. Any subsequent discussion of disagreements is reserved for qualitative analysis and does not replace the independent ratings.

Figures 8 and 9 illustrate the cross-model review and prompt–video–question views used to inspect the evaluation material. Automatic score summaries are consulted only after independent scoring is complete.



Figure 8: Cross-model review interface for inspecting generated videos and score summaries after independent annotation.

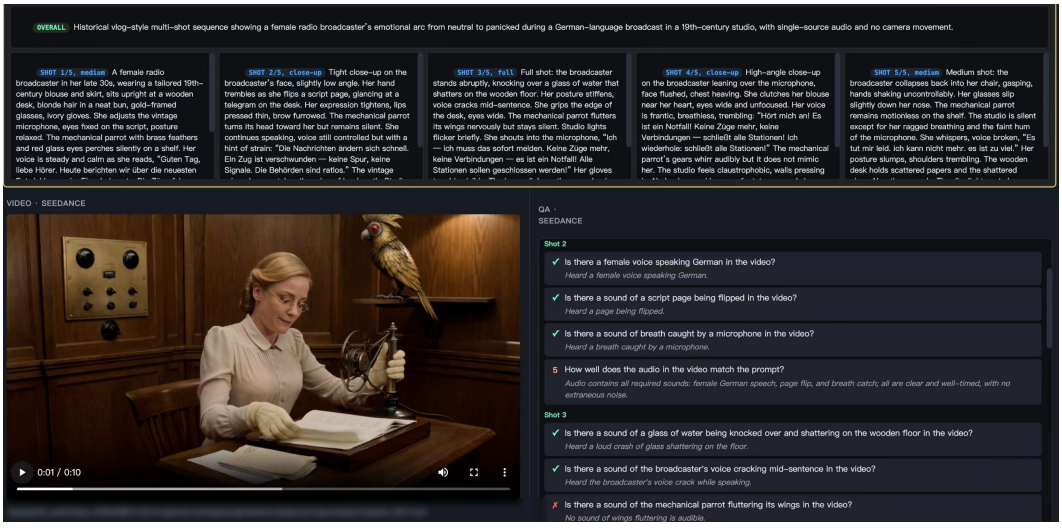


Figure 9: Detailed review interface showing a generated video, its prompt, shot descriptions, and evaluation questions.