MUSEBENCH: BENCHMARKING INTENT-LEVEL AUDIOVISUAL ARTS UNDERSTANDING IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audiovisual arts encompass diverse creative disciplines, including cinema, visual arts, stage performance, and game design, where artistic meaning arises from deliberate combinations of visual, auditory, and narrative elements (e.g., fear amplified through claustrophobic framing, or grief conveyed through silence and lingering close-ups). True artistic understanding extends beyond recognizing what is depicted to reasoning about why it is expressed through particular creative choices. Despite the strong progress of multimodal large language models (MLLMs), this critical aspect of artistic understanding remains underexplored, as existing benchmarks largely measure perceptual recognition while overlooking reasoning about creative intent. To address this gap, we introduce MUSEBENCH, a comprehensive benchmark designed to evaluate MLLMs on nuanced artistic understanding. It comprises 4,016 questions spanning cinematic arts, static visual arts, stage performing arts, and game arts that pair professional commentary with visual demonstration. To capture the open-ended nature of artistic analysis at scale, the benchmark combines single-select and variable-option multi-select questions. All questions are generated and refined through a four-phase iterative pipeline combining shortcut filtering, adversarial distractors, and expert validation. Comprehensive zero-shot evaluation of 28 state-of-the-art MLLMs reveals that even the best-performing model achieves only 48.28% accuracy, substantially below human expert performance of 87.18%, exposing a significant gap in current models' creative domain expertise. Further analysis points to a consistent failure pattern in which models lag sharply on game arts, recover partial interpretations but struggle to identify complete correct option sets on multi-select items, and gain little from adaptive key frame selection, suggesting the bottleneck lies in stylistic vocabulary and cultural priors rather than temporal localization.

1 INTRODUCTION

What does it mean to understand art? It is not merely recognizing what is shown, but interpreting why it is expressed in a particular way. The audiovisual arts (Krupskyy et al., 2021; Lokki et al., 1998; Kot et al., 2021), spanning cinema, visual arts, stage performance, and interactive media, provide a uniquely demanding setting for exposing this distinction. An artistic work is a deliberately designed expressive system in which creators orchestrate camera movement, composition, editing pace, lighting, blocking, and visual style to convey emotion, theme, and aesthetic intent (Carvalho & Lund, 2015; Sławek-Czochra & Sosnowska, 2023). Understanding such artistic expression requires reasoning about why a technique was chosen, how a visual arrangement serves creative intention, and what deeper artistic meaning emerges from the interplay of form and content (Barber, 2012; Naphade & Huang, 2002). For example, as illustrated in Fig. 1, asking why a director pairs symmetric framing with warm lighting, or punctuates a scene with prolonged silence, requires linking visual form to emotional intent rather than naming on-screen objects. This demands a level of comprehension that goes well beyond factual recognition or surface-level description: models must grasp not only what appears on screen but also the creator's underlying intent and the cultural conventions that inform it.

Although multimodal large language models (MLLMs) (Song et al., 2023; Zohar et al., 2025; Yang et al., 2024; He et al., 2026; Wang et al., 2025d; Tao et al., 2025; Song et al., 2025b) have rapidly approached human-level performance on standard perception and reasoning tasks, it remains underexplored whether they can capture such deeper artistic understanding. As illustrated in Fig. 1, existing

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

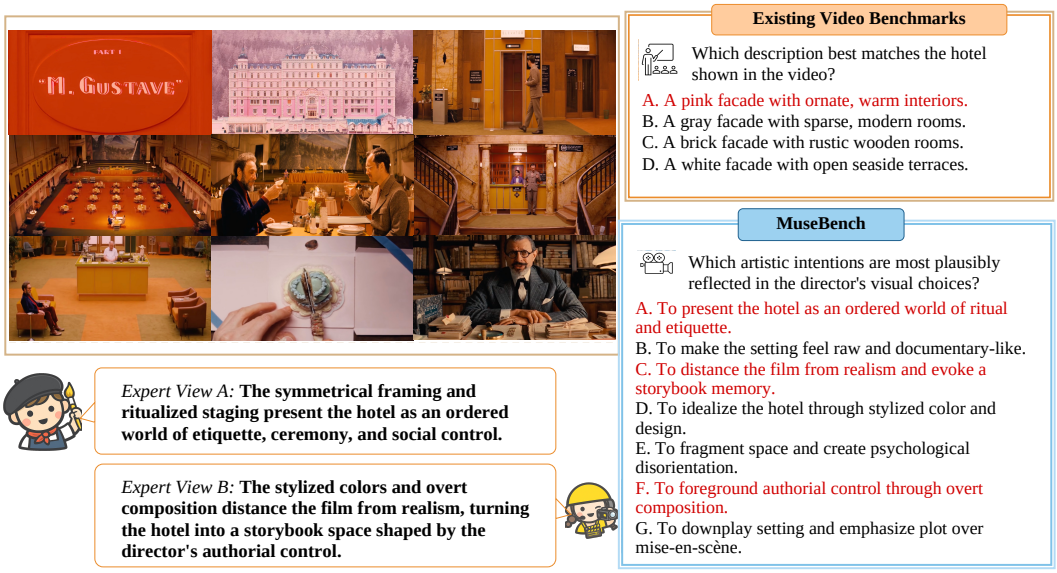


Figure 1: **Overview of MUSEBENCH.** Cinematic frames on the left ground two contrasting questions on the right. Existing video benchmarks (orange) test surface-level recall with a single correct answer, while MUSEBENCH (blue) probes the artistic intent behind directorial choices and allows multiple defensible answers, highlighted in red. The bottom-left dialogue shows how viewers can reach different yet well-supported interpretations of the same scene. This plurality in artistic understanding motivates our multi-select design, in which multiple options may be simultaneously correct.

video understanding benchmarks (Xu et al., 2017; Yu et al., 2019; Tan et al., 2025; Song et al., 2025a) primarily evaluate what is happening in a scene with a single correct option, rather than whether models can infer the intent behind creative decisions, such as why a director relies on symmetric composition, warm palettes, and ritualized blocking, or interpret their artistic significance. However, constructing a rigorous benchmark for intent-level artistic understanding is challenging on three intertwined fronts. **(i) Expert-knowledge scarcity:** Professional artistic analysis is inherently sparse, and authoring intent-level questions at benchmark scale is prohibitively expensive, well beyond what crowdsourcing can reliably supply. **(ii) Multiple valid interpretations:** Many analytical questions are constrained but non-unique and admit several defensible perspectives, so the dominant fixed four-option single-choice format collapses this plurality onto a single answer and reduces to pattern matching. **(iii) Reliable assessment of interpretation:** Even with high-quality questions, evaluation itself is a measurement problem. Naive accuracy on fixed-option items is not comparable across questions with different option counts, conflates successful guessing with genuine interpretation, and may fail to capture partial credit for the set-valued analytical judgments that artistic reasoning naturally produces. Addressing these challenges requires rethinking data sourcing, question format, and evaluation protocol in concert.

We tackle these challenges through three coordinated design choices, each directly aligned with the corresponding issues outlined above. **(i) Constructing expert-supervised data from video essays:** We leverage video essays (Bresland, 2010; Lavik, 2012) (sourced primarily from YouTube, Bilibili, and TikTok), analytical videos in which critics pair professional commentary with on-screen demonstrations, as an ideal source of grounded artistic analysis, since narration explicitly references the displayed visual content and thus yields natural temporal alignment between expertise and visual evidence. We develop a four-phase construction pipeline under iterative in-context updating (Sec. 3.3) that transforms over 10,000 video essays into benchmark questions requiring genuine visual understanding rather than transcript-based shortcuts. Within this pipeline, every distractor is crafted by an adversarial step that combines four complementary strategies, technical misread, over-simplification, factual error, and conceptual confusion, so that all options appear equally plausible to a reader without access to the clip and shortcut-driven guessing is suppressed. **(ii) Representing plurality through mixed formats.** We move beyond the fixed four-option paradigm (Sec. 3.2) and interleave single-select questions, which probe whether a model can identify the most precise interpretation, with multi-select questions that probe whether it can enumerate the full set of valid analytical dimensions, while the per-question option count varies between four and eight so that the

answer space reflects the open-ended structure of artistic reasoning rather than a uniform template. **(iii) Principled evaluation protocol.** For reliable assessment, we introduce a new scoring protocol designed for this heterogeneous setting (Sec. 3.5). We report overall accuracy (ACC) as the fraction of questions answered fully correctly. Our primary format-specific metrics are Chance-Adjusted Accuracy (CAA) for single-select questions, with expected score 0 under uniform guessing, and exact match (EM) for multi-select questions, requiring the complete correct option set. Precision, recall, and F1 diagnose incorrect selections, missed answers, and partial agreement. Empirically (Sec. 4), models often recover part of the correct set but struggle with exact recovery: both omitted answers and excess selections limit multi-select performance (Tab. 10).

Zero-shot evaluation of 28 MLLMs on MUSEBENCH shows that even the best model reaches only 48.28% accuracy against 87.18% human expert accuracy, exposing a gap that existing benchmarks obscure. Our analysis further shows that models struggle consistently with game art across all tiers, recover partial interpretations but struggle to identify complete correct option sets on multi-select questions, and benefit little from adaptive key-frame selection. This suggests that the main bottleneck is not temporal localization, but limited stylistic vocabulary and cultural priors. These findings point to richer artistic supervision and more multi-faceted evaluation, rather than further scaling of generic video understanding.

In summary, this paper makes three key contributions:

- **Benchmark for Audiovisual Arts Understanding.** We introduce MUSEBENCH, a new benchmark for audiovisual arts expertise, covering four art categories and 11 sub-domains, with questions over a variable number of options to capture interpretive plurality.
- **Scalable Expert-Knowledge Pipeline.** We develop a four-phase construction pipeline under iterative in-context updating that leverages video essays and vision-language models to generate visually grounded, intent-level questions at scale, addressing the fundamental challenge of acquiring expert knowledge for creative domains.
- **Principled Evaluation Protocol and Analysis.** We evaluate 28 MLLMs zero-shot using overall ACC, single-select CAA, and multi-select EM, with P/R/F1 diagnosing partial agreement. The best model reaches only 48.28% overall ACC, far below human experts at 87.18%. Multi-select analysis reveals that partial agreement can coexist with poor recovery of the complete correct option set.

2 RELATED WORK

Multimodal Large Language Models for Video Understanding. Multimodal Large Language Models (MLLMs) have advanced rapidly in video understanding, with efficient processing of many frames as a central challenge. One line of work pursues efficient encoding via sparse token memory (Song et al., 2023), visual summarization tokens (Shu et al., 2024), or native sparse attention for long contexts (Song et al., 2025b). A parallel line casts video understanding as agentic retrieval, using tree search (Yang et al., 2024), interleaved reasoning with temporal grounding (Yang et al., 2025b), or multi-agent coordination (Chen et al., 2025). Complementary training-side advances include process rewards for temporal alignment (Tao et al., 2025) and empirical studies of sampling and scaling (Zohar et al., 2025). Despite this progress, existing MLLMs are developed and evaluated almost exclusively on everyday activities, open-domain QA, or academic lectures, with no prior work probing the domain-specific expertise and interpretive reasoning demanded by the audiovisual arts, including cinematographic technique, compositional principles, and performance craft.

Benchmarks for Video Understanding. Video understanding benchmarks have progressed from short-clip QA (Xu et al., 2017; Yu et al., 2019) to story-level and temporal-reasoning frameworks (Huang et al., 2020; Li et al., 2024b). Recent work expands along multi-modal breadth (Fu et al., 2025), long-video scale (Wang et al., 2025c; Tan et al., 2025), and domain knowledge on expert lectures and STEM reasoning (Hu et al., 2025; Song et al., 2025a). Audio-visual perception has been explored in parallel, with AV-Odyssey Bench (Gong et al., 2024) probing fine-grained contrasts such as pitch and loudness, and our work extends this inquiry from low-level perception toward the interpretation of artistic intent. Despite this growing breadth, existing benchmarks largely center on general activities, factual comprehension, or academic STEM knowledge, leaving expertise in cinematographic technique, compositional principles, and performance craft underexplored.

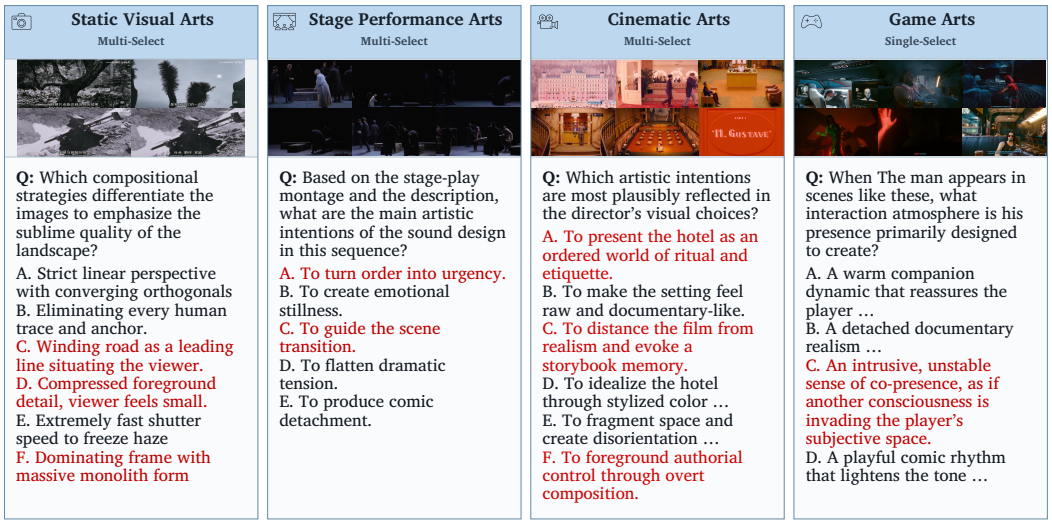


Figure 2: Representative examples from four MUSEBENCH categories.

3 MUSEBENCH CONSTRUCTION

This section details the design and construction of MUSEBENCH. We first motivate video essays as an expert-narrated knowledge source for probing audiovisual analytical understanding (Sec. 3.1), and then introduce our hierarchical capability taxonomy and two complementary question formats (Sec. 3.2). Building on this design, we describe a four-phase construction pipeline that transforms raw video essays into candidate question-answer pairs (Sec. 3.3), followed by an iterative quality review (Sec. 3.4). Finally, we introduce the specific evaluation metrics in MUSEBENCH (Sec. 3.5).

3.1 VIDEO ESSAYS AS A KNOWLEDGE SOURCE

A *video essay* is an analytical audiovisual format in which critics, educators, or practitioners examine artistic works through temporally aligned expert commentary and supporting visual or auditory evidence. Video essays are particularly well suited to our setting due to three key properties: (i) *expert-narration density*, creators explain not only what a technique is but why it produces a particular effect; (ii) *narration-to-evidence alignment*, spoken analysis directly references on-screen evidence; and (iii) *creative-arts coverage* across domains under-represented in existing video benchmarks (Song et al., 2025a), such as cinematography, fine art, photography, stage performance, and game art. Together, these properties enable us to derive intent-level questions about *why* a creative choice was made, rather than only *what* happens in a scene.

3.2 EVALUATION SCOPE AND QUESTION DESIGN

Inspired by (Järvinen, 2002), we establish a hierarchical capability taxonomy that drives both data collection and reporting to ensure comprehensive coverage across the audiovisual arts. At the top level, we identify four art categories (**Cinematic Arts**, **Static Visual Arts**, **Stage Performing Arts**, **Game Arts**) together with 11 sub-domains (see Sec. C.2), informed by the canonical organization of creative disciplines and the artistic topics most actively discussed by expert video essayists. Fig. 2 shows representative pairs from all four art categories. See Sec. H for more examples.

Within this taxonomy, MUSEBENCH combines two complementary question formats to capture the open-ended nature of artistic analysis. **Single-select** questions present a variable number of options (4 to 8) with exactly one correct answer and probe discrete recognition under a known-answer contract. **Multi-select** questions embed 2 to 4 correct answers among the options and probe set-valued analytical judgment, where the existence of multiple valid perspectives is itself part of the signal. Single-select isolates whether a model can discriminate the right interpretation under certainty. Multi-select tests whether it can enumerate the full set of valid interpretations without over-claiming. In both formats, the evaluation instruction indicates whether the question is single-select or multi-select, but does not reveal the exact number of correct answers.

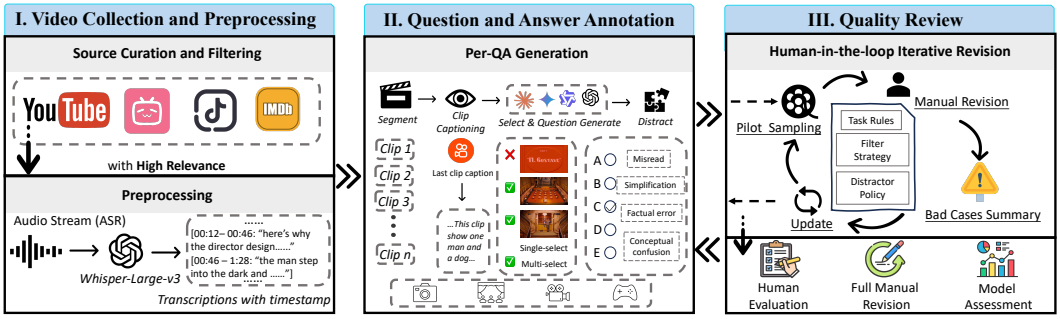


Figure 3: **Construction pipeline of MUSEBENCH.** Panel I curates video essays primarily from YouTube, Bilibili, and TikTok, applies relevance filtering against the audiovisual-arts taxonomy, and separates each retained video into two synchronized outputs with distinct roles, narrator transcripts for question construction and narrator-removed 10-second audiovisual clips for model evaluation. Panel II generates candidate questions through four per-video phases, *Segment*, *Clip Captioning*, *Select & Question Generate*, and *Distract*. Panel III closes a human-in-the-loop revision cycle (*Pilot Generation*, *Manual Revision*, *Bad Cases Summary*, *Update*) that feeds a *Full Regeneration* arrow back to Panel II; each category exits once the stopping criterion is met.

3.3 DATA COLLECTION AND QUESTION-ANSWER ANNOTATION

Video collection. Guided by the taxonomy above, we collect video essays primarily from YouTube, Bilibili, and TikTok that cover a broad range of expert commentary on the four art categories. We use GPT-5.4-mini (OpenAI, 2026) to generate category-aware search keywords and assess candidate relevance, retaining only videos with substantial audiovisual-arts analysis. Each retained video is then transcribed with Whisper-Large-v3 (Radford et al., 2023) to produce timestamped expert commentary aligned with the source video, supporting question generation and revision.

Question-Answer Annotation. As illustrated in Panel II of Figure 3, candidate QA pairs are generated through four successive phases shown in Figure 3.
❶ **Segment** (Figure 3, Panel II top): following (Wang et al., 2025b), each video is partitioned into 10-second intervals to establish a uniform temporal granularity for subsequent analysis.

❷ **Clip Captioning** (Figure 3, Panel II second row): Keye-VL-1.5 (Yang et al., 2025a) samples each 10-second segment at 1 fps and produces a single fine-grained caption per segment, conditioned on the temporally aligned narrator transcript. The captions cover visual attributes such as color, composition, motion and scene context. They are used solely as construction resources for downstream question generation and review, and are never exposed to models under evaluation.

❸ **Select & Question Generate** (Figure 3, Panel II third row): the clip-level captions and full narrator transcripts are provided as inputs for generating 3 to 5 candidate questions per video in single-select and multi-select formats. For each candidate item, relevant evidence clips are first identified, after which the question prompt and correct answer are generated conditioned on those clips. The process follows two constraints: (i) the question must remain answerable solely from the narrator-removed evidence clips, and (ii) the correct answer is formulated prior to any distractor to mitigate stylistic or lexical saliency bias toward the correct option.

❹ **Distract** (Figure 3, Panel II bottom): we generate plausible distractors using four core strategies, *technical misread* (valid domain terminology applied to a wrong analysis), *over-simplification* (partially correct but missing the core insight), *factual error* (contradicts visual or auditory evidence), and *conceptual confusion* (mixes related but distinct concepts), later extended to seven in the final prompt to absorb additional failure modes (see Sec. D.6). Each item draws from multiple strategies, and every option is required to appear equally plausible to a reader who has not seen the clip. Single-select pairs receive 3–7 distractors (4–8 options total); multi-select pairs mix 2–4 correct options with distractors. We further forbid proper nouns, prohibit near-identical phrasing across distractors, and randomly shuffle option positions. Full details in Secs. D.3 to D.6.

Copyright and release. MUSEBENCH is intended for noncommercial research. Data collection followed applicable platform policies and copyright requirements. We excluded works whose terms prohibit redistribution and do not treat public availability as permission to redistribute. We release

QA pairs, annotations, evaluation code, and source pointers, but do not redistribute third-party media unless redistribution is explicitly permitted by the applicable license or rights holder. Each item records provenance, source information, evidence timestamps, and available licensing or attribution metadata, and is reviewed before release. The package contains no third-party media (videos, clips, frames, audio, transcripts, thumbnails, or API responses). Users obtain source media through the original host or other authorized channels and remain subject to the corresponding terms. We maintain a takedown procedure for creators and rightsholders and version the benchmark when items become unavailable or are removed. Full release and rights-review procedures are provided in Appendix A.

3.4 Quality Review

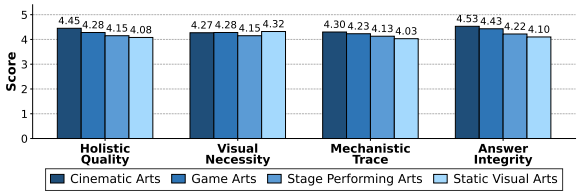


Figure 4: Expert quality ratings on 90 sampled items across four dimensions (0–5 Likert scale). Four domain experts participate, with two matched experts independently rating each category. Per-category averages range from 4.03 to 4.53 (STD: 0.41).

To ensure quality, we run the construction phases through an iterative review loop, where the in-context prompt is updated each round with new exclusion rules and domain-specific constraints. The loop is applied independently to each of the four art categories, with each round proceeding in four steps. ❶ **Pilot Generation**: a batch of candidate QA pairs is generated under the current prompt. ❷ **Manual Revision**: domain-expert reviewers assign binary pass/fail tags to each sampled QA pair under a shared failure taxonomy covering narrator-dependent answerability, ambiguous stems, weak or factually incorrect distractors, and misaligned clip references. Narrator-dependent cases are those answerable only from the expert transcript, not from the narrator-removed evaluation clips. ❸ **Bad Cases Summary**: we consolidate failure tags into a list of newly observed failure types for the round. ❹ **Update**: we rewrite the prompt with additional exclusion rules targeting new failures and then trigger a full regeneration.

During review of an early draft containing $5{,}000$ QA pairs, around 9% were flagged for severe errors in the CRITICAL or HIGH tiers. Across all severity levels, the audit identified eight tagged failure modes grouped into four tiers. The most severe tier covers hard schema violations such as labels pointing to nonexistent options, inline option lists in the stem disagreeing with the canonical options array, and multi-select pairs with an empty answer set. The lower tiers cover option-quality issues such as duplicated option texts, near-identical option prefixes, multi-select pairs that degenerate to single-select, and `correct_answer` fields that paraphrase rather than reproduce the option string. We retired the more severe tiers by replacement from the QA pool, and the lower tiers by a combination of strengthened generation and distractor prompts and programmatic post-hoc alignment. We additionally retired seven systemic issues that resist item-swap remediation, including low option discriminability, overly academic register, imprecise distractors lacking distinct error strategies, and inconsistent enforcement of the visual-evidence requirement, by full prompt-level rewriting. Full details in Secs. D.7 and D.8.

After generation, we manually verify every retained item before model assessment. To further validate the final benchmark, four domain experts are invited to independently rate a set of 90 samples along four quality dimensions on a 0–5 Likert scale (two assigned to Static Visual Arts and Game Arts, the other two to Stage Performing Arts and Cinematic Arts); the resulting per-category averages exceed 4.0 across every dimension with an average Inter-Annotator Agreement of Gwet AC2 (James, 2026; Popplewell et al., 2019) $= 0.855$, indicating near-perfect consistency across raters, as summarized in Fig. 4. Rater backgrounds, assignments, and the scoring rubric are detailed in Sec. D.9.

3.5 Evaluation Metrics

Single-select questions contain $K_i \in \{4, ..., 8\}$ options, so uniform-random accuracy varies as $1/K_i$. Raw accuracy therefore mixes model performance with item-specific guessing probability. We report chance-adjusted accuracy (CAA), which subtracts the per-item chance baseline and rescales the score:

$$\mathrm{CAA}_i = \frac{a_i - 1/K_i}{1 - 1/K_i}, \qquad \mathrm{CAA} = \frac{1}{N_{\mathrm{single}}} \sum_{i=1}^{N_{\mathrm{single}}} \mathrm{CAA}_i, \tag{1}$$

where $a_i \in \{0, 1\}$ indicates correctness and N_{single} is the number of single-select questions. CAA is normalized to 0 under uniform guessing and 1 for a correct response; negative aggregate scores indicate worse-than-chance performance. This enables fair comparison across subdomains with different option-count distributions. For multi-select questions, exact-match accuracy (EM) requires the predicted option set to match the ground truth exactly, measuring complete coverage of valid answers without selecting distractors. CAA and EM are the primary format-specific metrics. Overall ACC summarizes fully correct answers across both formats, with equal weight per question and no chance adjustment. Precision, recall, and F1 diagnose incorrect selections, missed answers, and partial agreement, respectively. Complete definitions and supplementary metrics are in Sec. E.

3.6 BENCHMARK STATISTICS

MUSEBENCH contains 4,016 expert-validated questions across four art categories and 11 sub-domains, drawn from over 10,000 candidate video essays. Each retained video contributes 3–5 questions with 4–8 options each (Table 4). Distributions by sub-domain, source platform, language, and cultural or regional provenance are provided in Sec. C.4.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate 28 MLLMs zero-shot in three tiers: proprietary (Singh et al., 2025; Anthropic, 2026; Team et al., 2023; xAI, 2025; Bai et al., 2025; Hong et al., 2025; Team et al., 2026), open source general-purpose (Bai et al., 2025; Xu et al., 2025; Chen et al., 2024b; Li et al., 2024a; Yu et al., 2026; Team et al., 2024), and open source video-specific (Cheng et al., 2024; Zhang et al., 2025a; Feng et al., 2025; Shen et al., 2024; Wang et al., 2025a; Li et al., 2025; 2024b; Shu et al., 2024), the last including 5 models with dynamic key frame selection (Tang et al., 2025; Zhang et al., 2025b; Yang et al., 2025b; Fei et al., 2024; Ren et al., 2024). Frames are sampled from narrator-removed evidence clips at 1 fps, capped at each model's maximum frame count. Audio-capable models receive the full narrator-removed audiovisual clip; other models receive a transcript of the same audio. All prompts specify single-select or multi-select.

4.2 MAIN RESULTS

Table 1 reports zero-shot results; Sec. F provides full per-category breakdowns.

Finding 1. Audiovisual-arts reasoning remains far from saturated, suggesting gaps in current training corpora. Frontier proprietary systems lead the leaderboard yet fall well short of expert-level performance, while video-specialized models cluster in a narrow band below them, offering no decisive advantage: even smaller general-purpose MLLMs match or surpass them. Together with the limited gains from dynamic key frame selection discussed below, this pattern suggests that stronger video processing alone is insufficient. Current pretraining and instruction-tuning recipes appear to cover only part of the stylistic vocabulary, cultural priors, and grounded inference required for expert-level artistic understanding. This motivates treating audiovisual arts as more than specialized video understanding and integrating richer artistic and cultural supervision into future training pipelines.

Finding 2. Game arts are a shared weakness while other categories surface divergent competence profiles. Performance decomposition along MUSEBENCH's four top-level categories reveals sharply different competence profiles (Fig. 5, left and middle), with no model dominating all axes. Models competitive on cinematic, static visual, and stage performing arts drop markedly on game footage across all tiers and formats, ruling out a metric artifact—likely due to the limited representation of interactive visuals, real-time camera control, and stylized rendering in web-scale corpora. Outside game arts, frontier proprietary systems maintain broad coverage, whereas open source models exhibit pronounced category specialization, pairing strong cinematic or static-visual subscores with severe weakness elsewhere. Aggregate accuracy thus obscures per-category reliability, motivating per-axis reporting for model selection.

Table 1: Zero-shot results on MUSEBENCH (%). Single-select uses chance-adjusted accuracy (CAA); multi-select uses exact match (EM). Overall ACC is the percentage of questions answered exactly. CAA bootstrap standard error is about 1.3 points. Best results are **bold**; second-best are underlined.

Model	Modality	Overall	Overall		Cinematic Arts		Static Visual Arts		Stage Perf. Arts		Game Arts	
		ACC	Single CAA	Multi EM	Single CAA	Multi EM	Single CAA	Multi EM	Single CAA	Multi EM	Single CAA	Multi EM
Random	–	13.55	0.02	6.04	0.04	5.85	-0.04	4.47	0.02	6.34	0.06	6.37
Human Expert	–	87.18	90.98	78.00	98.74	86.42	90.13	76.18	89.42	70.55	86.15	78.83
Proprietary MLLMs												
GPT-5.4 (Singh et al., 2025)	V+A+T	44.57	50.28	25.50	56.50	31.58	54.24	27.15	56.43	30.79	32.00	14.53
Claude-4.6-Opus (Anthropic, 2026)	V+T	**48.28**	55.13	**28.91**	63.26	31.58	58.51	**30.38**	**62.65**	**38.42**	34.07	**18.36**
Gemini-3.1-pro-preview (Team et al., 2023)	V+A+T	36.90	43.77	14.88	43.16	15.98	42.70	16.40	49.50	19.74	38.72	9.18
Grok-4.1 (xAI, 2025)	V+A+T	20.54	13.71	8.00	14.19	6.63	19.70	10.48	15.90	14.21	3.20	3.06
Qwen-3.5-Plus (Bai et al., 2025)	V+T	47.26	**58.52**	23.21	**68.88**	29.04	64.36	18.82	60.69	34.47	**38.80**	12.43
Doubao-Seed-1.8-Pro	V+A+T	46.12	55.00	24.22	62.10	29.04	**65.63**	25.54	56.84	27.37	32.86	16.25
GLM-4.5v (Hong et al., 2025)	V+T	17.13	5.43	8.61	16.17	14.62	-2.97	9.14	13.60	10.26	-4.34	1.15
Kimi-K2.5 (Team et al., 2026)	V+T	19.92	18.33	2.07	23.06	2.73	24.05	2.69	23.62	3.42	0.35	0.00
Open Source General-Purpose MLLMs												
Qwen3.5-397B-A17B (Bai et al., 2025)	V+T	44.77	53.42	22.71	62.65	**33.33**	57.45	23.12	56.60	24.47	35.79	10.71
Qwen2.5-Omni-7B (Xu et al., 2025)	V+A+T	32.69	30.71	18.18	36.47	21.44	40.76	17.47	34.43	24.21	8.31	11.09
InternVL3-8B (Chen et al., 2024b)	V+T	33.07	29.49	20.30	43.41	28.07	36.23	27.69	30.48	17.11	6.66	9.75
InternVL3-78B (Chen et al., 2024b)	V+T	37.80	47.03	13.53	47.98	14.23	48.79	18.01	57.25	17.37	31.59	6.88
LLaVA-OneVision-7B (Li et al., 2024a)	V+T	20.42	21.24	0.50	22.01	0.00	25.77	1.88	25.09	0.00	10.24	0.38
MiniCPM-o (Yu et al., 2026)	V+A+T	31.35	27.05	18.90	35.72	23.20	32.92	19.62	31.49	22.11	6.14	11.85
Gemma-4-E4B (Team et al., 2024)	V+A+T	27.61	28.67	9.06	39.40	11.89	32.55	13.17	31.44	8.68	10.30	3.63
Open Source Video-Specific MLLMs												
VideoLLaMA2 (Cheng et al., 2024)	V+A+T	20.34	20.07	1.17	31.35	1.95	18.07	1.61	25.34	1.32	5.40	0.00
VideoLLaMA3 (Zhang et al., 2025a)	V+A+T	27.19	26.82	9.90	34.37	10.33	24.55	11.56	33.76	9.47	14.02	8.60
Video-R1 (Feng et al., 2025)	V+T	26.72	28.41	7.21	30.50	5.07	28.70	9.68	38.49	10.26	13.87	5.35
LongVU (Shen et al., 2024)	V+T	14.87	8.21	1.01	14.40	0.19	6.50	1.34	5.46	0.79	7.75	1.72
VideoRFT (Wang et al., 2025a)	V+T	26.12	26.17	8.17	26.50	7.99	30.14	10.75	36.04	8.68	9.01	6.12
VideoChat-R1 (Li et al., 2025)	V+T	26.07	26.49	7.77	35.87	5.26	29.05	12.10	33.04	6.05	6.46	8.41
VideoChat2 (Li et al., 2024b)	V+T	17.78	15.27	0.34	17.20	0.19	14.35	0.00	18.67	0.79	10.45	0.38
Video-XL-2 (Shu et al., 2024)	V+T	24.18	29.91	0.11	28.63	0.00	29.29	0.54	40.42	0.00	19.17	0.00
AKS (Tang et al., 2025)	V+T	19.30	18.99	0.00	17.61	0.00	21.46	0.00	28.01	0.00	6.34	0.00
Q-Frame (Zhang et al., 2025b)	V+T	18.75	9.65	8.05	13.81	9.55	13.83	9.68	3.30	9.21	8.19	4.59
LongVT (Yang et al., 2025b)	V+T	20.52	17.14	4.64	17.99	2.73	21.61	9.14	21.92	6.32	5.02	2.10
Video-CCAM (Fei et al., 2024)	V+T	17.53	15.10	0.00	23.38	0.00	18.31	0.00	16.92	0.00	1.05	0.00
TimeChat (Ren et al., 2024)	V+T	14.42	7.79	0.34	12.27	0.58	9.66	0.00	4.61	0.26	5.04	0.38

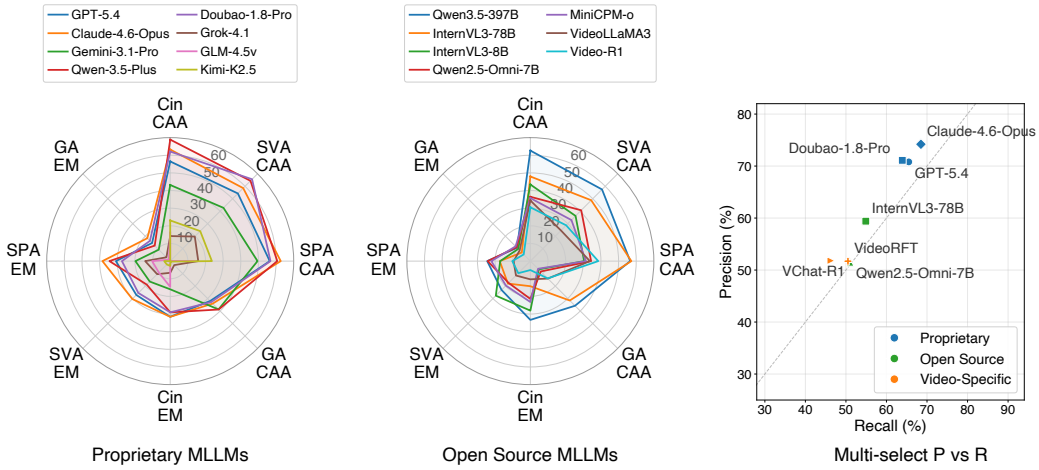


Figure 5: Per-category Single CAA (upper axes) and Multi EM (lower axes) for proprietary (left) and open-source (middle) models; multi-select precision versus recall (right). Cin/SVA/SPA/GA denote Cinematic/Static Visual/Stage Performing/Game Arts.

4.3 IN-DEPTH ANALYSIS

Beyond the leaderboard, we report four complementary analyses on key frame selection, multi-select behavior, modality contributions, and option position bias.

Finding 3. Key frames provide limited gains. We evaluate five MLLMs equipped with dynamic key frame selection (AKS (Tang et al., 2025), Q-Frame (Zhang et al., 2025b), LongVT (Yang et al., 2025b), Video-CCAM (Fei et al., 2024), and TimeChat (Ren et al., 2024)), which adaptively pick informative frames at inference time rather than ingesting a fixed uniform sample. Despite this added flexibility, all five cluster between 14.42 and 20.52 ACC, at or below the lower end of the video-

specialized tier and trailing the strongest video-specific models by 7 to 13 points. Content-aware frame selection therefore does not translate into a leaderboard advantage for these systems. This is consistent with a bottleneck in artistic vocabulary and cultural priors: locating salient frames alone does not supply the knowledge needed to interpret their artistic significance.

Finding 4. Partial agreement does not ensure a complete interpretation. On multi-select questions, precision exceeds recall for most evaluated models (Fig. 5, right), indicating that they often omit valid readings. The reverse error also matters. On game arts, LongVU reaches 91.65% recall and 67.64% F1 but only 1.72% EM; Claude-4.6-Opus has lower recall (55.78%) and F1 (59.59%) yet substantially higher EM (18.36%) (Tabs. 1 and 10). Broad coverage can therefore yield strong partial agreement while still producing an incorrect option set. Precision, recall, and EM together distinguish missed interpretations from excess selections.

The distinction also changes model comparisons within a family. InternVL3-78B surpasses InternVL3-8B in overall ACC (37.80 versus 33.07) and single-select CAA (47.03 versus 29.49), but falls behind in multi-select EM (13.53 versus 20.30). The larger model's advantage on choosing one answer does not extend to recovering all valid answers. Selecting one plausible interpretation is therefore an insufficient proxy for recovering all defensible readings; model choice should account for the answer format, even within the same family.

Finding 5. Modality gain. On all 4,016 questions, adding audio to video-and-text input improves Overall ACC by 7.52 points for MiniCPM-o (Yu et al., 2026) and 6.42 points for Gemma-4-E4B (Team et al., 2024) (Table 2). Single CAA rises by 7.77 and 8.11 points, respectively, while Multi EM rises by 7.18 and 2.79 points. This matched comparison isolates the contribution of audio alongside video and text, showing that it supplies task-relevant evidence across both question formats.

Table 2: Audio ablation on MUSEBENCH (%): video+text (V+T) versus video+audio+text (V+A+T).

Model	Input	ACC	CAA	EM
MiniCPM-o	V+T	23.83	19.28	11.72
MiniCPM-o	V+A+T	**31.35**	**27.05**	**18.90**
Gemma-4-E4B	V+T	21.19	20.56	6.27
Gemma-4-E4B	V+A+T	**27.61**	**28.67**	**9.06**

Finding 6. Open-source MLLMs exhibit a pronounced first-position bias. On the 1,407 single-select questions carrying five or more answer choices (Fig. 6), the gold labels are approximately uniformly distributed between positions A through E at 15.9–18.2% each, yet the model predictions focus heavily on the early positions. Position A alone accounts for 30.9% of all model predictions, roughly double its 15.9% gold share, while positions E through H collectively receive only 22.2% of predictions against a 33.0% gold share. This front-loading is markedly stronger among open source MLLMs, whose 34.9% A-prediction rate exceeds the gold share by 19.0 points; proprietary models, by contrast, predict A at 22.4%, only 6.5 points above the gold baseline. Because MUSEBENCH randomizes answer placement across all items, the asymmetry reflects a model-intrinsic positional prior rather than a label-ordering artifact, and it points to a default-to-first fallback that open source MLLMs invoke when the visual and cultural evidence is insufficient to discriminate among options.

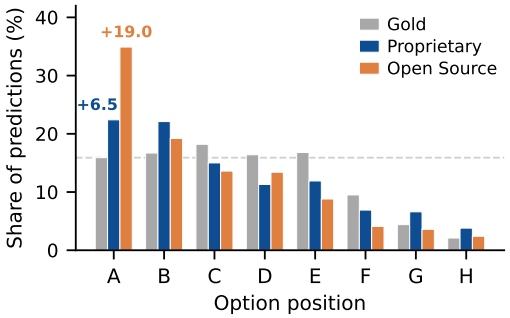


Figure 6: Option position bias (n=1,407; ≥5 choices).

5 CONCLUSION

We introduce MUSEBENCH, a benchmark for audiovisual arts understanding with 4,016 expert-validated questions across four art categories, constructed from video essays through a four-phase pipeline with iterative human-in-the-loop quality review. Zero-shot evaluation of 28 MLLMs yields a best-model accuracy of 48.28%, compared with 87.18% for human experts. Models lag sharply on game arts, recover partial interpretations but struggle to identify complete correct option sets on multi-select questions, and gain little from adaptive key frame selection, suggesting a bottleneck in stylistic vocabulary and cultural priors rather than temporal localization.

AI USE STATEMENT

AI models were used both in benchmark construction and to assist manuscript preparation. As described in Sec. 3.3, GPT-5.4-mini assisted with search-keyword generation and source-relevance assessment, Whisper-Large-v3 transcribed the source videos, and Keye-VL-1.5 generated clip captions. Language models generated candidate questions, answers, and distractors from the construction materials (Secs. D.5 and D.6). The construction pipeline incorporated iterative human review and prompt revision, and every retained item was manually verified before model assessment (Sec. 3.4). LLMs also assisted with writing and revising code.

For manuscript preparation, LLMs assisted with drafting, revising, and polishing the text, including improving clarity and organization. The authors retain responsibility for the research design, experimental results, interpretation, citations, and final content of the manuscript.

ETHICS STATEMENT

MUSEBENCH uses third-party video essays and artistic works, making source rights, attribution, and responsible interpretation central to its release. The planned public package contains QA pairs, annotations, source pointers, timestamps, and provenance records rather than copies of third-party media; Sec. A describes the rights review and correction and takedown procedures. Human quality review involved four domain experts participating as unpaid collaborators, with a separate two-expert performance baseline, as detailed in Sec. D.9. The benchmark's answers represent source-supported interpretations rather than exhaustive judgments of artistic meaning. Its language and regional coverage limits the traditions represented (Secs. C.4 and G); scores should be interpreted within this scope rather than as measures of the value of artworks, creators, or cultures.

REPRODUCIBILITY STATEMENT

Secs. 3 and D document source curation, transcription, evidence selection, question and distractor generation, and human review, including the construction prompts. Secs. 4.1 and F describe the evaluation inputs, model access, and computational resources, while Secs. 3.5 and E define the scoring rules. Dataset composition and human evaluation procedures are provided in Secs. C.4 and D.9. We plan to release the benchmark annotations, source pointers and timestamps, provenance records, and evaluation code under the protocol in Sec. A.

REFERENCES

Anthropic. Introducing Claude Opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, 2026. February 5, 2026.

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

Sian Barber. Understanding online audio-visual content: a european initiative, media literacy and the user. *Medijske studije*, 3(06):28–41, 2012.

John Bresland. On the origin of the video essay. *Blackbird: an online journal of literature and the arts*, 9(1), 2010.

Ana Carvalho and Cornelia Lund. *The audiovisual breakthrough*. Collin & Maierski Print GbR, 2015.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20237–20246, 2025.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, pp. 179–195. Springer, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

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

Muhammad Iqbal Hasan Chowdhury, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Hierarchical relational attention for video question answering. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 599–603. IEEE, 2018.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

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

Yangfan He, Changgyu Boo, and Jaehong Yoon. Are video reasoning models ready to go outside? *arXiv preprint arXiv:2603.10652*, 2026.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Kairui Hu, Penghao Wu, Fanyi Pu, Wangding Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *ArXiv*, abs/2501.13826, 2025. URL https://api.semanticscholar.org/CorpusID:275820371.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European conference on computer vision*, pp. 709–727. Springer, 2020.

Joseph James. Counting on consensus: Selecting the right inter-annotator agreement metric for nlp annotation and evaluation. *arXiv preprint arXiv:2603.06865*, 2026.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2758–2766, 2017.

Aki Järvinen. Gran stylissimo: The audiovisual elements and styles in computer and video games. In *Computer games and digital cultures conference proceedings*, 2002.

Halyna M Kot, Olena G Levchenko, Tetiana O Kravchenko, and Oksana S Musiienko& Kostiantyn V Hrubych. Problems of intertextuality in audio-visual arts. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 13(1), 2021.

IV Krupskyy, NI Zykun, AP Ovchynnikova, SI Gorevalov, and OA Mitchuk. Determinants and modern genres of audio-visual art. *Journal of the Balkan Tribological Association*, 27(4), 2021.

Erlend Lavik. The video essay: The future of academic film and television criticism? *Frames Cinema Journal*, 1(1):19, 2012.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

Tapio Lokki, Jarmo Hiipakka, Rami Hänninen, Tommi Ilmonen, Lauri Savioja, and Tapio Takala. Realtime audiovisual rendering and contemporary audiovisual art. *Organised Sound*, 3(3):219–233, 1998.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Milind R Naphade and Thomas S Huang. Extracting semantics from audio-visual content: the final frontier in multimedia retrieval. *IEEE Transactions on Neural Networks*, 13(4):793–810, 2002.

OpenAI. GPT-5.4 Mini Model. https://developers.openai.com/api/docs/models/gpt-5.4-mini, 2026. Accessed September 26, 2026.

Michael Popplewell, John Reizes, and Chris Zaslawski. Appropriate statistics for determining chance-removed interpractitioner agreement. *The Journal of Alternative and Complementary Medicine: Paradigm, Practice, and Policy Advancing Integrative Health*, 25(11):1115–1120, 2019.

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

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26160–26169, 2024. URL https://api.semanticscholar.org/CorpusID:272827076.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Małgorzata Sławek-Czochra and Joanna Sosnowska. Perspective of the audiovisual arts: On ways and tools of studying emotions in the current visuals. *Roczniki Kulturoznawcze*, 14(1):153–167, 2023.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tianbo Ye, Yang Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232, 2023. URL https://api.semanticscholar.org/CorpusID:260333927.

Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. Video-mmlu: A massive multi-discipline lecture understanding benchmark. *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 6158–6172, 2025a. URL https://api.semanticscholar.org/CorpusID:277955206.

Enxin Song, Wenhao Chai, Shusheng Yang, Ethan Armand, Xiaojun Shan, Haiyang Xu, Jianwen Xie, and Zhuowen Tu. Videonsa: Native sparse attention scales video understanding. *arXiv preprint arXiv:2510.02295*, 2025b.

Xichen Tan, Yuanjing Luo, Yunfan Ye, Fang Liu, and Zhiping Cai. Allvb: All-in-one long video understanding benchmark. In *AAAI Conference on Artificial Intelligence*, 2025. URL https://api.semanticscholar.org/CorpusID:276928535.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29118–29128, 2025.

Sicheng Tao, Jungang Li, Yibo Yan, Junyan Zhang, Yubo Gao, Hanqian Li, ShuHang Xun, Yuxuan Fan, Hong Chen, Jianxiang He, et al. Moss-chatv: Reinforcement learning with process reasoning reward for video temporal reasoning. *arXiv preprint arXiv:2509.21113*, 2025.

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

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025a.

Shihao Wang, Guo Chen, De-an Huang, Zhiqi Li, Minghan Li, Guilin Li, Jose M Alvarez, Lei Zhang, and Zhiding Yu. Videoitg: Multimodal video understanding with instructed temporal grounding. *arXiv preprint arXiv:2507.13353*, 2025b.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025c.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Conference on Empirical Methods in Natural Language Processing*, 2025d. URL https://api.semanticscholar.org/CorpusID:280149603.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024.

xAI. Grok 4. *https://x.ai/news/grok-4*, 2025.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025a.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. Vca: Video curious agent for long video understanding. *ArXiv*, abs/2412.10471, 2024. URL https://api.semanticscholar.org/CorpusID:274776498.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. *arXiv preprint arXiv:2511.20785*, 2025b.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

Zhou Yu, D. Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. *ArXiv*, abs/1906.02467, 2019. URL https://api.semanticscholar.org/CorpusID:69645185.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025b.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18891–18901, 2025.

Appendix Contents

A Pointer-First Release Plan, Provenance, and Rights **16**

B Limitation and Future Work **16**

C Benchmark Details **17**

C.1 Comparison with Existing Benchmarks 17

C.2 Category and Sub-domain Definitions 17

C.3 Dataset Statistics 18

C.4 Composition by Sub-domain and Source 18

C.5 Benchmark Vocabulary Overview 18

D Construction Details **20**

D.1 Source Curation 20

D.2 Transcription 22

D.3 Phase A: Clip Segmentation 22

D.4 Phase B: Clip Captioning 22

D.5 Phase C: Evidence Selection and Question Generation 25

D.6 Phase D: Distractor Generation 25

D.7 Quality Review Loop 25

D.8 Failure Taxonomy 26

D.9 Human Evaluation Details 30

E Detailed Evaluation Metrics **32**

E.1 Single-Select Evaluation: Chance-Adjusted Accuracy 32

E.2 Multi-Select Evaluation: Exact Match, Precision, Recall, and F1 32

E.3 Overall Accuracy 33

F Complete Experimental Results **33**

F.1 Language and Option-Set Shortcut Controls 34

G Broader Impact **34**

H Additional Examples of MuseBench **35**

A Pointer-First Release Plan, Provenance, and Rights

Pointer-first public release. Upon acceptance, we plan to release MuseBench and the associated evaluation code on Hugging Face under a *pointer-first* protocol for noncommercial research use. The package will contain question–answer pairs, benchmark annotations, source pointers (source URLs or platform video IDs), evidence timestamps, and item-level provenance and attribution records. Third-party content, including videos, clips, frames, audio, transcripts, thumbnails, platform API responses, or cached copies, will be included only when redistribution is explicitly permitted by the applicable license or rightsholder; otherwise, the corresponding item will remain pointer-only. The intermediate pool of more than 10,000 candidate videos will not be released. Any samples shared privately for review are separate from the public package.

Source eligibility and review. During source curation, we excluded sources whose stated terms were incompatible with our collection or release protocol, and we did not treat public availability as permission to redistribute. Each item is reviewed before release, taking into account available license information, creator permissions, and attribution requirements for the source video and, where identifiable, the underlying works it discusses.

Access to source media. The pointer-first protocol limits what we redistribute; it does not grant access to, or rights in, the linked media. Users obtain source media through the original host or other official or authorized channels, such as the platform's own interface or approved APIs, and remain subject to the corresponding terms of service and licensing conditions.

Provenance and attribution. Each released item records its source pointer, platform, creator or channel, evidence timestamps, available license or attribution metadata, and, when identifiable, the underlying film, game, performance, artwork, or other work. Rights in the linked media and underlying works remain with their respective rightsholders, and we do not claim ownership of them.

Copyright and platform compliance. Data collection and benchmark construction were conducted with consideration of applicable platform policies and copyright requirements. The benchmark is intended for noncommercial research. The public benchmark is pointer-first; third-party media are included only when redistribution is explicitly permitted.

Correction and takedown. The release documentation will describe how creators and rightsholders can request correction or removal of an item. When a request is accepted, we will remove the corresponding pointer and annotations, record the change in the release notes, and not provide a cached substitute. The benchmark release does not grant permission to download, repost, or redistribute linked source media.

Versioning and continued availability. Because linked media may be edited, made private, or removed by their hosts, the set of evaluable items may change over time. We will maintain versioned releases with a changelog recording added, corrected, removed, and unavailable items. Items that become unavailable will be flagged in subsequent versions. Items removed at a rightsholder's request will no longer be distributed in any hosted release, while the changelog will preserve the record that the item was removed. We recommend reporting results together with the benchmark version and the number of items evaluated, so that results remain comparable across versions.

B Limitation and Future Work

MuseBench currently focuses on four art categories and relies on video essays as the primary data source. While video essays provide rich expert commentary, they may not fully capture the diversity of artistic expression in raw creative works, and their availability varies across art forms and languages. Future work will explore expanding to additional art forms (e.g., music, architecture), incorporating more diverse multilingual sources, developing open-ended evaluation beyond multiple-choice, and investigating whether targeted fine-tuning on arts-specific data can close the observed gap.

Benchmarks	#Clips	Avg Len. (s)	#QA Pairs	Anno.	QA Tok.	Sub. Tok.	Open Domain	Aud.	Domain Expert	Visual Dep.
MSRVTT-QA (Xu et al., 2017)	2,990	15.2	72,821	A	8.4	✗	✓	✗	✗	✗
MSVD-QA (Chowdhury et al., 2018)	504	9.8	13,157	A	7.6	✗	✓	✗	✗	✗
TGIF-QA (Jang et al., 2017)	9,575	3.0	8,506	A&M	20.5	✗	✓	✗	✗	✗
ActivityNet-QA (Yu et al., 2019)	800	111.4	8,000	M	10.2	✗	✗	✗	✗	✗
STAR (Wu et al., 2024)	7,098	11.9	7,098	A	19.5	✗	✓	✗	✗	✗
NExT-QA (Xiao et al., 2021)	1,000	39.5	8,564	A	25.3	✗	✓	✗	✗	✗
MVBench (Li et al., 2024b)	3,641	16.0	4,016	A	27.3	✗	✓	✗	✗	✗
Video-Bench (Han et al., 2025)	5,917	56.0	17,036	A&M	21.3	✗	✓	✗	✗	✗
EgoSchema (Mangalam et al., 2023)	5,063	180.0	5,063	A&M	126.8	✗	✗	✗	✗	✗
AutoEval-Video (Chen et al., 2024a)	327	14.6	327	M	11.9	✗	✓	✗	✗	✗
TempCompass (Liu et al., 2024)	500	11.4	7,540	A&M	49.2	✗	✓	✗	✗	✗
Video-MMMU (Hu et al., 2025)	900	588.4	900	M	–	✓	✗	✗	✓	✗
MUSEBENCH (Ours)	4,016	50.26	4,016	A	168	✓	✓	✓	✓	✓

Table 3: **Comparison of MUSEBENCH with existing video understanding benchmarks.** "Domain Expert" indicates whether the benchmark requires domain-specific expertise. "Visual Dep." indicates whether questions are explicitly designed to require visual evidence beyond text transcripts. MUSEBENCH is a comprehensive benchmark that simultaneously targets domain expertise in audio-visual arts and enforces visual dependency.

C BENCHMARK DETAILS

C.1 COMPARISON WITH EXISTING BENCHMARKS

Table 3 situates MUSEBENCH against representative video understanding benchmarks along scope, annotation, and modality axes. Existing benchmarks largely remain confined to short clips over everyday activities or open-domain factual QA, without coupling subtitles and audio, without enforcing domain expertise, and without controlling for visual dependency in question design. Recent expert-oriented efforts such as Video-MMMU (Hu et al., 2025) introduce multi-level difficulty over lecture videos, yet they target STEM-style knowledge transfer rather than artistic interpretation and still do not require joint use of subtitle and audio.

In contrast, MUSEBENCH is the only benchmark in the table that simultaneously satisfies the four capability axes Open Domain, Sub.&Aud., Domain Expert, and Visual Dep. Each axis traces back to a component of our construction pipeline (Sec. 3). The hierarchical taxonomy across four art categories and eleven sub-domains realizes the domain expertise dimension, and the narrator-removed audiovisual clips paired with timestamped expert commentary (Sec. 3.3) jointly enforce visual dependency and audiovisual reasoning. These properties position MUSEBENCH as a comprehensive benchmark targeted at expert-level analytical understanding of the audiovisual arts.

C.2 CATEGORY AND SUB-DOMAIN DEFINITIONS

MUSEBENCH organizes 4,016 questions under a hierarchical capability taxonomy: four top-level art categories, each partitioned into fine-grained sub-domains for a total of 11 sub-domains. The taxonomy is grounded in the canonical organization of creative disciplines and the topics most actively analyzed by expert video essayists.

Cinematic Arts. This category covers the analytical study of moving-image works, including narrative film, television drama, documentary, and animation. Items probe how filmmakers translate dramatic intent into image and sound through four sub-domains.

- **Cinematography** – shot composition, camera movement, framing, lensing, depth of field, lighting, and exposure choices that shape how a scene is seen.
- **Mise-en-scène** – on-screen staging including blocking, props, set design, costume, and spatial composition that organize the world inside the frame.
- **Editing and Pacing** – shot-to-shot construction, including cuts, transitions, montage, parallel action, and the rhythm imposed by edit length.

- **Sound** – diegetic and non-diegetic audio, score, ambient sound, silence, and sound bridges that anchor or extend the visual narrative.

Static Visual Arts. This category covers analyses of still images and tangible artifacts, where temporal structure is replaced by compositional and material reasoning.

- **Fine Art** – painting, drawing, printmaking, and sculpture, focusing on composition, color theory, perspective, brushwork, chiaroscuro, and material technique.
- **Photography** – documentary, fine-art, and commercial photography, focusing on exposure, aperture, lensing, framing, and the photographer's documentary intent.

Stage Performing Arts. This category covers live performance forms in which meaning is constructed through performer presence, staged space, and time-bound audience reception.

- **Performance** – acting, vocal delivery, physical gesture, body language, and the live communicative work of the performer, including drama, dance, and stand-up.
- **Stage Design** – the spatial language of the stage, including set design, scenery, props, platforms, and the blocking that organizes performers within them.
- **Theatrical Lighting** – the artistic use of stage lighting, including spot, follow spot, side and back lighting, color washes, and warm/cool contrasts that direct attention and shape mood.

Game Arts. This category covers the audiovisual craft of video games, where artistic choices co-exist with interactive systems and player agency.

- **CG** – the rendered visual language of games, including art style, real-time and pre-rendered cinematics, shading and lighting models, and post-processing aesthetics.
- **Interactive Visuals** – visual elements bound to player interaction, including level design, environmental storytelling, on-screen guidance and HUD, navigation cues, and gameplay-conditioned camera and animation.

C.3 Dataset Statistics

Table 4 consolidates the statistics that characterize MuseBench, covering global scope, source data, question format.

Table 4: Comprehensive statistics of MuseBench, covering global properties and construction parameters.

Property	Value
Total questions	4,016
Art categories	4 (cinematic, static visual, stage performing, game)
Sub-domains	11
Candidate source video essays	>10,000
Sampling rate during captioning	1 fps
Questions per video	3 to 5
Options per question	4 to 8

C.4 Composition by Sub-domain and Source

The distribution below covers all 4,016 retained questions. We report source language separately from the cultural or regional provenance of the work discussed in a video essay, since narration language does not necessarily identify that work's origin.

C.5 Benchmark Vocabulary Overview

We visualize the dominant terms of MuseBench in Fig. 7 after removing function words and generic analytical fillers, exposing the artistic vocabulary that drives stems, options, and core intents.

Table 5: Question counts by art category and sub-domain.

Category	Sub-domain	Single	Multi	Total	Share (%)
Cinematic Arts	Cinematography	204	219	423	10.53
Cinematic Arts	Mise-en-scène	85	83	168	4.18
Cinematic Arts	Editing and Pacing	93	97	190	4.73
Cinematic Arts	Sound	102	117	219	5.45
Static Visual Arts	Fine Art	363	213	576	14.34
Static Visual Arts	Photography	262	162	424	10.56
Stage Performing Arts	Performance	315	197	512	12.75
Stage Performing Arts	Stage Design	151	97	248	6.18
Stage Performing Arts	Theatrical Lighting	142	86	228	5.68
Game Arts	CG	195	211	406	10.11
Game Arts	Interactive Visuals	310	312	622	15.49
Total	11 sub-domains	2,222	1,794	4,016	100.00

Table 6: Distribution of retained questions by source platform, narration language, cultural or regional provenance, and genre. Each dimension totals 4,016 questions; percentages are rounded.

Dimension	Group	Questions	Share (%)
Platform	YouTube	2,073	51.62
	Bilibili	1,696	42.23
	TikTok	156	3.88
	IMDb	91	2.27
Source language	English	2,519	62.72
	Chinese	1,497	37.28
Culture/region	North America	1,196	29.78
	Global/Mixed	1,028	25.60
	Europe	924	23.01
	East Asia	791	19.70
	Others	77	1.92
Genre	Mixed	885	22.04
	Action/Adventure	344	8.57
	Role-Playing	279	6.95
	Portrait/Figure	271	6.75
	Drama	236	5.88
	Variety/Circus/Magic	218	5.43
	Landscape/Cityscape	200	4.98
	Comedy	187	4.66
	Musical	187	4.66
	Puzzle/Platformer	185	4.61
	Others	1,024	25.50



Figure 7: Vocabulary of MUSEBENCH. Word size reflects token frequency in questions, options, and core intents after removing function words and generic analytical fillers.

D Construction Details

This appendix expands the construction pipeline summarized in Section 3.3 and the human-in-the-loop quality review described in Section 3.4. The subsections that follow track the per-video phases of the main text in order: source curation feeds the corpus; transcription and category inference attach the metadata each video needs; *Phase A* segments the video into 10 s clips; *Phase B* captions every clip; *Phase C* first selects evidence ranges and then generates 3 to 5 candidate questions per video; *Phase D* synthesizes distractors. The two final subsections document the iterative review loop that mutates the in-context prompt across rounds and the final-stage human evaluation that validates the released benchmark.

D.1 Source Curation

This subsection details the source-discovery stage that feeds the construction pipeline. The goal is a corpus of long-form video essays in which a domain expert verbally analyzes an artistic artifact while the visual content is shown on screen. Source discovery proceeds in four LLM-controlled stages. (1) A category-aware keyword generator produces an initial keyword list grounded in the controlled vocabulary that defines each category in the main text. (2) The keywords are issued to public web video search and a candidate pool of `(video_id, title, channel, description, view_count)` records is collected. (3) A relevance judge inspects each candidate's metadata and emits a binary verdict together with a confidence score; only candidates with `is_relevant=true` and confidence at or above 0.55 are admitted. (4) When an active keyword exhausts its top results without yielding new admitted videos, a variant generator is asked to extend the keyword list, and crawling continues until the per-category quota is reached. A final human-vetting pass then removes residual failure modes that the metadata filter cannot detect. The four LLM stages are each governed by an explicit system and user prompt, reproduced in Figs. 8 to 11. A summary of the deterministic fallback keyword lists, with per-block sizes and 3–5 representative entries per block, is reported in Table 7.

Keyword generation prompt. Fig. 8 reproduces the system and user prompt used at stage 1, instantiated for Cinematic Arts. The same envelope is reused for the other three categories with the focus list and example terms swapped to the matching controlled vocabulary. Output is constrained to a single JSON object `{"keywords":[...]}` and the decoding temperature is fixed at 0.2.

Keyword generation prompt (Cinematic Arts)

System.
You are a video search expert. Given a user topic, generate high-quality search queries that target long-form video essays.
Requirements. (1) Focus on film breakdowns, film essays, audiovisual language, and director shot technique. (2) Use cinematic terminology such as montage, shot language, mise-en-scène, long take, color composition. (3) Match the way analytical viewers actually phrase searches. (4) Diversify across film genres and directorial styles. (5) Each query must be short, analytically loaded, and avoid pure title-only searches.
Output. A single JSON object `{"keywords":[...]}` and nothing else.

User.
Topic: `{topic}`. Generate `{num_keywords}` effective search queries that cover film breakdowns, film essays, and cinematic technique. Required focal points (combine freely):
1. Cinematography. Shot size (long shot, extreme close-up), camera angle (Dutch angle, bird's eye), movement (dolly zoom, tracking), lighting (high key, low key), composition (rule of thirds, leading lines).
2. Editing. Montage, long take, jump cut, parallel cut, flashback, transitions (match cut, smash cut), Kuleshov effect.
3. Mise-en-scène. Set and prop symbolism, blocking, costume.
4. Sound design. Sound-to-image relation, diegetic vs. non-diegetic, sound bridge, ambient sound, silence.
Examples may anchor on a director or auteur case.

Figure 8: Keyword generation prompt for Cinematic Arts. The Static Visual, Stage Performing, and Game Arts variants share the same envelope, with the four numbered focal points replaced by the corresponding controlled vocabulary of the target category.

Relevance judgment prompt. For every candidate returned by the search step, the relevance judge sees only public metadata plus the active keyword list and emits a JSON verdict. Of the four categories, Stage Performing is the broadest, since it admits stand-up specials, sketch comedy, dance

theater, and traditional opera forms in addition to musical theater and opera. Its prompt, reproduced in Fig. 9, is therefore the most explicit about positive and negative cases. The Cinematic, Static Visual, and Game variants share the same envelope and additionally hard-exclude DIY tutorials, software walkthroughs, and gameplay-only content respectively.

Relevance judgment prompt (Stage Performing Arts, broad form)

System.
You are a content analyst for stage and live performance video essays. Decide whether the given video belongs to analysis, commentary, or in-depth discussion of any stage or live performance form.
Scope (broad). Theater, opera, musical theater (stage design, lighting, directorial style); stand-up comedy specials; sketch comedy; long-form and short-form improv; ballet, modern dance, dance theater; cabaret and drag; circus and magic; mime, physical comedy, clowning; spoken word; talk-show performance; traditional opera forms such as Peking opera, kabuki, and noh.
Positive. Any of the above forms with explanation, analysis, ranking, or in-depth discussion. Comedy special reviews, comedian technique breakdowns, musical theater reviews.
Negative. Pure performance recording without analytical commentary, music videos, film analysis, gaming or anime content, news segments, or popular-science content.
Output. A single JSON object `{"is_relevant": bool, "reason": str, "confidence": float in [0,1]}`.

User.
Active keywords: `{keywords}`.
Candidate metadata.
Title: `{title}`
Description: `{description}`
Channel: `{channel}`
View count: `{view_count}`
Decide whether this candidate focuses on analysis, commentary, or discussion of stage or live performance, and provide a short reason.

Figure 9: Relevance judgment prompt for Stage Performing Arts. The Cinematic, Static Visual, and Game variants share the envelope and additionally hard-exclude DIY tutorials, software walkthroughs, and gameplay-only content respectively.

Variant expansion prompt. When a keyword exhausts its top results without producing new admitted videos, the keyword generator is invoked again under the prompt in Fig. 10, conditioned on the most recent ten keywords already issued. The variant generator is asked to stay inside the category's focal vocabulary (the `variant_focus` string).

Variant expansion prompt (category-agnostic envelope)

System.
You are an advanced search expert. Given a list of existing keywords, produce novel non-duplicate variants that cover additional relevant videos. Stay strictly inside the focal vocabulary supplied by the user. Do not return queries that are paraphrases of an existing one or that target a single named work.
Output. A single JSON object `{"keywords":[...]}` and nothing else.

User.
Existing keywords (most recent ten):
- `{kw_1}`
- `{kw_2}`
- `...`
- `{kw_10}`

Generate `{num_variants}` novel non-duplicate keywords that focus on `{variant_focus}`.

Figure 10: Variant expansion prompt. `variant_focus` is the category-specific focal string; for example, *cinematography, editing, mise-en-scène and sound design* for Cinematic Arts and *stage performance, musical theater, stand-up comedy, dance and live performance art analysis* for Stage Performing Arts.

Human-vetting prompt. After the relevance judge admits a candidate, a final pass removes the three residual failure modes that metadata filtering cannot detect. The instruction shown to reviewers (and used as a verbatim system prompt when the same pass is delegated to a stronger LLM as a sanity check) is reproduced in Fig. 11.

Human-vetting prompt (final source-list cut)

System.
You are a final-stage source curator. The candidate has already passed the automated relevance judgment. Your job is to reject the residual three failure modes that metadata filtering cannot detect. Return a single decision JSON.
Reject (i) *narrator-only*. The on-screen visuals are a static talking head, generic stock footage, or a slide deck rather than the analyzed artifact itself.
Reject (ii) *tutorial drift*. The video is structured as a how-to guide that walks through a workflow rather than as an analytical essay about an artifact, even when the keywords match.
Reject (iii) *promotional cut-down*. The video is a trailer, showreel, advertisement, or fan compilation rather than commentary, even when long.
Output. A single JSON object
`{"keep": bool, "reject_reason": "none | narrator_only | tutorial_drift | promotional"}`.

User.
Title: `{title}`
Description: `{description}`
Visual sample (eight uniformly sampled frames): `{frames}`
Audio sample (first 30 seconds transcript): `{transcript}`
Decide whether the candidate should enter the construction pool.

Figure 11: Human-vetting prompt applied as a final cut over admitted candidates.

Keyword list summary. Table 7 reports the per-block size of the deterministic-fallback keyword list together with 3–5 representative entries per block. The fallback covers 273 keywords across the four categories: Cinematic Arts and Static Visual Arts use a flat 10-keyword controlled vocabulary each; Game Arts is split into a core controlled vocabulary, case-anchored studies, studio/director vocabulary, and a general-purpose expansion block; Stage Performing Arts is the largest list, organized into thirteen thematic blocks that span comedy, musical theater, opera, drama, dance, variety, world theater, and general performance analysis. The complete word-for-word lists are released with the benchmark code release.

D.2 Transcription

Every retained video is transcribed with Whisper-Large-v3 (Radford et al., 2023) into a JSON record that anchors the rest of the pipeline. Fig. 12 shows the schema. The full free-text transcript is consumed by the QA generation prompt of Phase C; the timestamped `segments` entries are routed to the clip-level captioning prompt of Phase B so that each 10 s window can see the narration that overlaps it; the `metadata` block lets the segmenter enforce the duration, resolution, and channel-count limits documented in Table 4.

D.3 Phase A: Clip Segmentation

Following (Wang et al., 2025b), each retained video is partitioned into non-overlapping ten-second clips that establish a uniform temporal granularity for downstream captioning, question generation, and evaluation. The maximum video duration is capped at 1,800 s (30 minutes) so that a single source contributes a bounded number of clips and a bounded number of QA pairs. Each clip carries a stable index that is reused as a stable handle by the captioning and clip-matching prompts and by the per-question `relevant_clips` field.

D.4 Phase B: Clip Captioning

Each ten-second clip is captioned by Keye-VL-1.5 (Yang et al., 2025a) sampling at one frame per second, conditioned on the temporally aligned narrator transcript. The caption covers visual attributes such as color, composition, motion, and scene context. Captions serve only as a construction artifact for downstream question generation and review; they are never exposed to evaluated models. The prompt envelope is reproduced in Fig. 13; subsequent clips additionally see the running chronological narrative of previous clips so that descriptions remain locally coherent. The system prompt is concatenated at runtime with one of four category-specific guidance blocks (Cinematic, Static Visual, Stage Performing, or Game) that biases the caption toward category-relevant evidence.

Table 7: Summary of the deterministic-fallback source-discovery keyword lists, with per-block totals and representative examples (3–5 per block, full list omitted).

Block	#KW	Representative keywords
Cinematic Arts		
Cinematography and lighting	4	cinematography shot size breakdown; dutch angle vs bird's eye; dolly zoom vertigo effect; high/low key lighting study
Composition and mise-en-scène	2	rule of thirds leading lines composition; mise en scene blocking symbolism
Editing	2	montage cross-cutting film editing; Kuleshov effect explanation
Sound	2	diegetic vs non-diegetic sound design; sound bridge ambient noise film
Static Visual Arts		
Fine art and art history	5	painting composition analysis; color theory visual art breakdown; chiaroscuro oil painting; impressionism / post-impressionism; art history renaissance and baroque
Photography	3	photography visual language analysis; rule of thirds photography composition; fine art photography essay depth of field
Concept and digital art	2	concept art design principles essay; digital art critique aesthetic analysis
Game Arts		
Controlled-vocabulary core	10	game visual storytelling environmental design; ray tracing real-time rendering aesthetic; cel shading hand-drawn game art; game CG cinematic animation; level design visual guidance
Case-anchored studies	30	Elden Ring art direction dark fantasy; Hollow Knight hand-drawn art; Red Dead Redemption 2 lighting; Cuphead 1930s cartoon animation; Disco Elysium painting style
Studio / director vocabulary	11	Naughty Dog cinematic game design; FromSoftware visual design philosophy; Nintendo art direction philosophy; Silent Hill visual symbolism
General game-art vocabulary	19	game UI/UX design aesthetic; game environment art world building; Unreal Engine 5 Nanite showcase; game lighting mood atmosphere; horror game visual atmosphere
Stage Performing Arts		
Stand-up comedy and specials	41	George Carlin comedy genius; Bo Burnham Inside special; John Mulaney comedy style; stand-up structure and callback craft
Sketch comedy and improv	16	SNL sketch comedy analysis; Key and Peele sketch genius; Monty Python comedy; UCB / Second City improv training
Musical theater	41	Sondheim musical genius; Hamilton cultural impact; Hadestown mythology; Lion King staging puppetry; Sweeney Todd analysis
Opera	11	Wagner Ring Cycle; Verdi dramatic analysis; Puccini La Bohème; Carmen opera; opera vs musical theater
Drama and theater	12	Shakespeare staging; Arthur Miller *Death of a Salesman*; Tennessee Williams *Streetcar*; immersive theater Sleep No More
Dance and ballet	8	Swan Lake interpretation; Pina Bausch dance theater; contemporary dance analysis; Nutcracker staging
Cabaret, drag, variety	9	RuPaul Drag Race performance; drag queen lip-sync; burlesque history; America's Got Talent best acts
Circus and magic	8	Cirque du Soleil show analysis; Penn and Teller magic explained; Derren Brown mentalism; acrobatics performance art
Physical comedy and clown	7	Charlie Chaplin comedy genius; Buster Keaton physical comedy; Marcel Marceau mime; Jacques Tati comedy style
Spoken word and poetry	5	spoken word poetry performance; poetry slam competition; TED talk performance technique; oral storytelling tradition
Roast and panel comedy	9	comedy roast best moments; late-night monologue analysis; Conan O'Brien comedy style; Graham Norton best moments
World theater traditions	10	kabuki theater visual analysis; Beijing opera performance; Noh theater mask symbolism; commedia dell'arte; Bharatanatyam
General performance analysis	10	stage presence technique; concert staging visual design; Super Bowl halftime analysis; Eurovision performance
Total	**273**	

Per-video transcription record (illustrative)

```
{
  "video_id": "kuleshov_effect_explained_eisenstein_breakdown",
  "source": "<provider>",
  "category": "cinematic_arts",
  "video_path": "data/videos/cinematic_arts/<video_file>.mp4",
  "transcription": {
    "model": "Whisper-Large-v3",
    "language": "en",
    "text": "In this segment we examine how Eisenstein layers two...",
    "segments": [
      { "id": 0, "start":  0.00, "end":  6.84,
        "text": "In this segment we examine how Eisenstein layers two
                 unrelated shots so the cut itself produces meaning." },
      { "id": 1, "start":  6.84, "end": 13.21,
        "text": "Notice that the close-up of the face does not change,
                 yet the read of the emotion shifts with the cutaway." },
      { "id": 2, "start": 13.21, "end": 19.05,
        "text": "This is the Kuleshov effect, the foundation of analytic
                 montage we will trace through the rest of the essay." }
    ],
    "created_at": "2026-04-08T11:42:18+00:00"
  },
  "metadata": {
    "duration_seconds": 612.5,
    "frame_rate":       29.97,
    "resolution":       "1280x720",
    "audio_channels":   2,
    "audio_sample_rate": 48000
  }
}
```

Figure 12: Schema of the per-video transcription record produced by the preprocessing stage. Each video yields one such JSON file. The top-level fields anchor the record to a category and a source file; the `transcription` block contains the full free-text transcript together with timestamped sentence-level segments used downstream by the clip-level captioning and QA stages; `metadata` stores container properties used by the segmenter to bound clip duration and resolution.

Clip description prompt (Phase B)

System.
You are an audiovisual arts analyst. Your task is to extract professional, fine-grained visual information from consecutive frames of a video essay clip.
General requirements.
1. Analyze the frame sequence chronologically and interpret it as one coherent visual segment.
2. Adjacent frames may look similar because they are continuous frames; do not misinterpret this as visual effects.
3. If on-screen text appears, quote the original text and provide an English translation when needed, then explain its contextual meaning.
4. Distinguish different people using concrete cues (clothing, position, posture, etc.).
5. Provide rich visual detail covering color, shape, texture, motion traits, and scene background.
Language. The `clip_description` value must be written in natural English only.
Output. A single JSON object `{"clip_description": "The clip starts with..., develops through..., and ends with..."}`.
At runtime the system prompt is concatenated with a category-specific guidance block (for Cinematic Arts: shot scale, camera angle, camera movement, light and color style, composition; the Static Visual, Stage Performing, and Game variants substitute their own controlled vocabulary).

User (first clip).
Transcript segment for the current clip (narration or commentary): `{transcript_segment}`
Please describe this first clip of the video.

User (subsequent clips, with context).
Previous clip descriptions (chronological): `{previous_descriptions}`
Transcript segment for the current clip (narration or commentary): `{transcript_segment}`
Please describe the current clip, using previous clips as context when relevant.

Figure 13: Clip description prompt used by Phase B of the construction pipeline (Section 3.3). The first-clip user template omits the previous-clip context; subsequent clips receive the running chronological narrative so that descriptions remain locally coherent. The system prompt is concatenated at runtime with one of four category-specific guidance blocks.

D.5 Phase C: Evidence Selection and Question Generation

Given the chronological clip captions and the full narrator transcript, the generator first identifies a single contiguous `relevant_clips` range for each candidate item. It then generates the question and correct answer conditioned on the selected evidence, producing 3 to 5 candidate pairs per video. The target proportion of `multi_select` questions is 44.67% (2 to 4 independent correct answer points) and the rest are `single_select` (one correct answer). The prompt envelope is reproduced in Fig. 14. As with captioning, the system prompt is augmented with one of four category-specific question-design guidance blocks at runtime. The evidence-selection step follows the contiguous-range constraint documented in Fig. 16 (failure 4).

QA generation prompt (Phase C)

System.
You are an exam designer for an audiovisual arts benchmark. Based on clip-level visual descriptions and transcript text, generate high-quality QA items for the target art domain.
Core requirements.
1. Select evidence first. For each item, identify a single contiguous `relevant_clips` range from the chronological clip descriptions and aligned transcript before writing the question. The item must be answerable from the narrator-removed evidence in that range.
2. Generate the question and correct answer from the selected evidence before generating distractors. Use a concrete visual or auditory anchor to locate the question without stating the relationship or effect that the solver must infer. Provide an accurate, professional, and well-reasoned correct answer.
3. Difficulty labels. Use `basic`, `intermediate`, or `advanced`.
4. Sub-domain labels. Assign a sub-domain from the category's controlled list.
5. Question type. Target a `multi_select` share of 44.67% across the generated corpus (2 to 4 independent correct answer points returned in `correct_answers`); the rest are `single_select` (one `correct_answer`).
Quantity. Generate 3 to 5 questions per video.
Language. All textual fields must be in English only.
Output. A single JSON object `{"questions":[...]}` with each entry listing `relevant_clips`, `question`, `question_type`, `correct_answer` or `correct_answers`, `difficulty`, and `sub_domain`.
At runtime the system prompt is concatenated with a category-specific question-design guidance block.

User.
Video metadata. Category: `{category}`; Video ID: `{video_id}`; Total clips: `{num_clips}`.
Clip visual descriptions (chronological): `{clip_descriptions}`.
Full transcript: `{full_transcript}`.
First select a contiguous evidence range for each item, then generate 3 to 5 questions and their correct answers from the selected ranges, targeting a `multi_select` share of 44.67% across the generated corpus. Each question must test visual understanding and should not be answerable from transcript text alone.

Figure 14: QA generation prompt used by Phase C. The full transcript and chronological clip descriptions support evidence selection first, followed by question and correct-answer generation from the selected range. Each item must be answerable from the narrator-removed evaluation clip. The system prompt is augmented with one of four category-specific question-design guidance blocks at runtime.

D.6 Phase D: Distractor Generation

The distractor generator turns each correct answer into a multiple-choice item by synthesising 3 to 7 plausible distractors per question. The prompt envelope is reproduced in Fig. 15. At runtime, seven distractor strategies are exposed: the four documented in Section 3.3 (*technical misread*, *over-simplification*, plus the equivalents of *factual error* and *conceptual confusion*) and three additional strategies (*scope error*, *temporal confusion*, *partial truth*) that were added during the iterative review loop to absorb failure modes the four-strategy form did not yet cover. The category-specific `scope_description` field shares the controlled vocabulary used by the source-curation prompts of Section D.1.

D.7 Quality Review Loop

This subsection documents the iterative review loop introduced in Section 3.4. The loop is applied independently to each art category. In every round, a pilot batch of QA pairs is generated under the current in-context prompt, domain-expert reviewers tag each sampled item under the four failure dimensions of Section 3.4, the new failure types of the round are consolidated, and the prompt is rewritten with additional rules. Rules take two forms. *Hard red lines* are short-tagged content

Distractor generation prompt (Phase D)

System.
You are a benchmark item reviewer specializing in plausible distractors for audiovisual-arts multiple-choice questions.
Distractor strategies (use a different strategy for each distractor).
1. *Technical misread*. Uses valid terminology but draws a wrong analysis.
2. *Over-simplification*. Sounds partially correct but misses the core insight.
3. *Concept swap*. Confuses closely related but distinct concepts.
4. *Cause-effect inversion*. Reverses the causal relation between technique and effect.
5. *Scope error*. Applies a correct technique analysis to the wrong scope, scene, or time range.
6. *Temporal confusion*. Misattributes timing, sequence, or duration of techniques.
7. *Partial truth*. Correct about one aspect but fundamentally wrong about the key insight.
Requirements. Each distractor uses a different strategy. Each distractor must look plausible on the surface and require domain knowledge to reject. Keep all distractors tightly grounded in domain-specific concepts and avoid semantic repetition. All distractor text must be in English.
Output. A single JSON object `{"distractors":[{"text":...,"misconception_type":...}, ...]}` containing exactly `{num_distractors}` entries.

User.
Category: `{category}`; Sub-domain: `{sub_domain}`; Question type: `{question_type}`.
Question: `{question}`.
Correct answer(s): `{correct_answer}`.
Scope for this category: `{scope_description}`.
Generate exactly `{num_distractors}` plausible distractors using a different strategy for each. `{multi_select_guidance}`.

Figure 15: Distractor generation prompt used by Phase D. Seven strategies are exposed at runtime; the four named in Section 3.3 (*technical misread*, *over-simplification*, plus the equivalents of *factual error* and *conceptual confusion*) are the originally documented core, while *scope error*, *temporal confusion*, and *partial truth* were added during the iterative review loop to absorb failure modes that the four-strategy form did not yet cover. The category-specific `scope_description` field shares the controlled vocabulary used by the source-curation prompts of Section D.1.

constraints written into the QA-Generation prompt that gate the stem and options at generation time. *Category filters* are short-tagged content constraints written into the source-curation human-vetting pass that gate the source clips before they enter Phase A.

Fig. 16 walks through one representative bad case per main-text failure dimension and shows the prompt-level rule added in response together with the regenerated form of the same item. The four rows of the figure correspond exactly to the four failure dimensions named in Section 3.4 (narrator-dependent answerability; ambiguous stems; weak or factually incorrect distractors; misaligned clip references), so a reader who has just read Section 3.4 can scan top to bottom and see the loop in action for each named dimension.

Game Arts trajectory across three rounds. A category-specific instance of the loop is reproduced in Fig. 17, which traces a single Game Arts item through the three prompt revisions that retired the round-by-round failure types. Round 1 used a generic instruction that asked the model to "analyze the design intent" of an arbitrary clip; pilot review surfaced two recurring failures, namely stems written in an abstract academic register that no longer commit to a specific design choice, and pairs in which the generation pipeline failed to settle on a unique key. Round 2 tightened the prompt with two hard requirements: every question must name a specific game in the stem, and every stem must point to a concrete on-screen design choice. This eliminated the abstract-register failure but introduced a second-order shortcut, since the named title is sufficient for a model with strong text priors to retrieve the answer from training-set memory without watching the clip. Round 3 added two anti-shortcut rules in response: *R1* forbids proper nouns (game titles, character names, locations) anywhere in the stem, options, or correct answer; *R2* requires each stem to identify a visible or audible element that locates the question in the clip, while leaving the relevant spatial, temporal, or expressive relationships for the solver to infer. After Round 3 no new failure type was raised in three consecutive rounds, and Game Arts exited the refinement loop.

D.8 Failure Taxonomy

We additionally conducted a round-by-round human audit of an early benchmark draft of 5,000 QA pairs. In each round, domain experts inspected the full corpus and, at the end of the round,

Failure 1: narrator-dependent answerability

Observed bad case (stem). "Why does the narrator argue that the dolly zoom in this scene is more emotionally effective than a simple zoom?"
Why it fails. The answer paraphrases a verbal claim from the narrator transcript. Once the narrator track is removed for evaluation, the clip itself contains no spoken argument, so a model that has watched the visual evidence cannot recover the answer.
Rule added (R3, *not transcript-only*). Every stem must be answerable from visual or audible evidence inside the narrator-removed clip. Questions that paraphrase narrator claims, summarize narrator opinions, or otherwise treat the transcript as ground truth are forbidden.
Regenerated stem. "In the shot centered on the protagonist, how does the changing spatial relationship between the figure and the room shape the emotional effect?"

Failure 2: ambiguous stems

Observed bad case (stem). "What is interesting about the design of the still life shown?"
Why it fails. The stem names no observable element of the artifact; many mutually inconsistent answers are equally valid. Source clips whose narration only restates generic art theory frequently produce such stems.
Rules added (R2, *grounded in AV evidence*; F4, *under-specified*). R2 requires each stem to identify a concrete visible or audible element that locates the question, without supplying the relationship or effect that the solver must infer. F4 filters out source clips whose narration is too generic to anchor any concrete observable, so the question generator never receives them.
Regenerated stem. "How does the placement of the porcelain bowl within the frame guide the viewer's attention in the still life shown?"

Failure 3: weak or factually incorrect distractors

Observed bad case (options). A. "creates rhythm through editing tempo"; B. "creates rhythm through editing pacing"; C. "creates rhythm through editing speed"; D. "creates rhythm through cut frequency". Four lexical paraphrases of the same idea; a reader cannot use domain knowledge to choose between them.
Rule added (Distractor prompt). Each distractor must use a distinct strategy from the seven-strategy taxonomy (*technical misread*, *over-simplification*, *concept swap*, *cause-effect inversion*, *scope error*, *temporal confusion*, *partial truth*); no two options may share a fifty-character prefix; option texts must be unique.
Regenerated options. A. (correct) "locks the audience into the protagonist's heartbeat by aligning each edit point with a percussive accent"; B. (*technical misread*) "shortens the average shot length to compensate for an underlit interior"; C. (*concept swap*) "cross-cuts between two unrelated scenes to build narrative parallelism"; D. (*cause-effect inversion*) "slows the apparent action by inserting establishing shots after each cut".

Failure 4: misaligned clip references

Observed bad case (`relevant_clips`). `[3, 8, 12]`. The matcher attached three disjoint clips even though the relevant evidence sits inside clip 8 alone; the evaluator who plays the concatenation of these three windows cannot reconstruct the original visual context.
Rule added (Clip Match prompt). The `relevant_clips` field must form a single contiguous range. When the evidence spans disjoint regions of the video, the evidence-selection step first chooses the most informative contiguous range; the question generator then writes the stem and correct answer using only evidence within that range.
Regenerated reference. `relevant_clips = [7, 8, 9]`, selected before generating a question about the visible setup that unfolds across those three consecutive ten-second windows.

Figure 16: Four representative failure modes uncovered during the quality review loop (Section 3.4), each shown with the bad case observed during pilot generation, the prompt-level rule added in response, and the regenerated form of the same item after the rule fired. The four rows correspond exactly to the four failure dimensions named in the main text. R/F-tagged rules are content-semantic red lines on the stem and options; the remaining two rules are schema-level constraints written into the Distractor and Clip Match prompts of the construction pipeline.

Round 1: initial generation

Q. Analyze the interplay between the environmental conditions and the sequential logic presented. Which two conclusions accurately characterize the intentionality of the design and its formal execution?

A. The environmental brightness functions as a psychological cue to suggest a sense of optimism and emotional reprieve for the participant during the final stages of the journey.
B. The implementation of a daylight setting serves as a strategic mechanism to heighten the visibility of hazards.
C. The transition suffers from a structural flaw where the absence of specific visual markers, such as the transport vehicle, breaks narrative continuity.
D. The exclusion of the C-130 transport plane at the start of the sequence is a calculated design choice to emphasize the survivors' isolation.

Key: `""` (unresolved)

Failure modes addressed in Round 2. abstract academic register; no concrete design choice; generation pipeline failed to commit a unique correct option.

Round 2: named-game decision focus

Q. In the "Blood Harvest" campaign of *Left 4 Dead*, the storefronts on the building facades carry the logos of the Richardson Atlantic Company and its subsidiaries. What player experience does this design choice most directly aim to produce?

A. (correct) A sense that the world is a real place where people lived and worked, rather than a generic zombie shooting gallery.
B. A way to help the player quickly locate the destructible weak points of each building.
C. A cue that collecting company logos is required to unlock map rewards.
D. A direct narrative critique of corporate headquarters as the central theme of the map.

Key: A; `game_referenced`: Left 4 Dead

Failure mode addressed in Round 3. the named title plus the named campaign in the stem are sufficient priors for a strong language model to retrieve the answer from training-set memory without watching the clip.

Round 3: visible-anchor, no proper nouns

Q. How does the arrangement of interior spaces and the player's route through them shape the experience of this level?

A. (correct) A prolonged sense of confinement and unease by delaying escape and increasing exposure to threats.
B. A methodical resource-scarcity loop, where the cramped rooms force players to ration ammo by limiting engagement distances.
C. A tactical chokepoint system that rewards reading enemy patrol paths and timing movement through each corridor.
D. A spatial memory challenge where vertical layering forces the player to mentally reconstruct the building layout to navigate efficiently.

Key: A; `visual_anchor`: "narrow corridors, locked-in spaces, vertical multi-level buildings". The internal title field is withheld from the evaluated model.

Stable form. the stem identifies the interior layout and the player's route as the focus of analysis without describing their spatial constraints or stating the intended effect.

Figure 17: Game Arts question evolution across three prompt revisions. Round 1 produces an abstract academic stem with no committed key; Round 2 enforces a named-game decision focus but introduces a title-recognition shortcut; Round 3 removes proper nouns and anchors the question to the interior layout and player movement, leaving their relationship and expressive effect to be inferred from the clip. Each box lists the failure mode that motivated the next revision.

consolidated the newly observed failures into a tag set that fed the next round of prompt revision. Across rounds, the audit surfaced eight failure tags spread across four severity tiers, each accompanied by a concrete prompt-level rule whose addition retired the failure. Table 8 reports the tag, severity, count, one-sentence description, and mitigating rule for each entry. The four severity tiers correspond to the order in which the audit retired the failures: CRITICAL and HIGH issues were eliminated by replacement from the QA pool, while MEDIUM and LOW issues were eliminated by a combination of strengthened generation and distractor prompts and programmatic post-hoc alignment. None of the listed failures remain in the released benchmark.

Table 8: Failure taxonomy uncovered by a multi-round human-in-the-loop review of benchmark construction. Each row gives the review tag, the severity bucket, the count of items affected, a one-sentence description of what the failure looks like, and the prompt-level rule added in response. Rows are sorted by severity; the four severity tiers and counts mirror the order in which the review retired them. Severity-CRITICAL and HIGH issues were eliminated by replacement from the QA pool; MEDIUM and LOW issues were eliminated by a combination of prompt-level rules and programmatic post-hoc alignment. None of the listed failures remain in the released benchmark.

Severity	review tag	Count	What the failure looks like	Mitigating rule added
CRITICAL	`INVALID_LABEL`	175	The `correct_label` points to an option index that does not exist in the options array.	Distractor-prompt schema check rejects items whose label set does not match the option set; residual items replaced from the QA pool.
CRITICAL	`EMBEDDED_MISMATCH`	60	The stem contains an inline option list whose labels disagree with the canonical options array.	QA-Generation rule forbids inline option lists in stems.
CRITICAL	`MULTI_0_ANSWER`	37	A multi-select item carries an empty `correct_answers` list.	QA-Generation rule requires at least two correct answers for any `multi_select` item.
HIGH	`ALL_SAME_PREFIX`	99	All options share a fifty-character prefix and become indistinguishable on a first read.	Distractor-prompt rule: no two options may share a fifty-character prefix; strategy diversity enforced across options.
HIGH	`DUPLICATE_OPTS`	85	Two or more options have identical text.	Distractor-prompt rule: option texts must be unique.
MEDIUM	`MULTI_1_ANSWER`	785	A multi-select item carries only one correct answer and degenerates to single-select.	QA-Generation rule strengthened: `multi_select` items must list 2 to 4 independent correct points.
MEDIUM	`MANY_OPTS_SIMILAR`	83	More than half of the option pairs share a fifty-character prefix.	Distractor-prompt rule: pairwise prefix-diversity threshold lowered to thirty percent of option pairs.
LOW	`ANSWER_TEXT_MISMATCH`	639	The `correct_answer` field is a paraphrase rather than the exact option text.	Programmatic post-hoc alignment of `correct_answer` to the exact option string.

In addition to the eight tagged failures of Table 8, the audit identified seven systemic issues that resist item-swap remediation and were instead addressed by full prompt-level refinement: low option discriminability, overly academic or verbose language, stems that do not precisely target artistic intent, imprecise distractors lacking distinct error strategies, occasional misclassification of text-vision alignment, inconsistent application of the visual-evidence requirement, and embedded option lists in stems. The corresponding prompt-level fixes are reflected in the QA-Generation and Distractor prompts of Sections D.5 and D.6, and in the four reviewer-facing dimensions of Fig. 16.

D.9 Human Evaluation Details

To support the human evaluation reported in Section 3.4, we built a self-contained web interface (Fig. 18) that lets each domain expert play the corresponding clip in place, read the question and options, and rate the item on the four Likert dimensions defined below.

Rater Pool. We invited four domain experts, who participated as unpaid collaborators, and assigned them to categories matching their formal training. Two raters major in Drama, Film and Literature and were assigned to *Stage Performing Arts* and *Cinematic Arts*; the remaining two major in Computational Media and Arts and were assigned to *Game Arts* and *Static Visual Arts*. Each category was independently rated by two raters, yielding two scores per item which we then average for the per-category means in Fig. 4.

Expert performance baseline. The expert performance baseline was collected separately from the four-person quality review above. Two domain experts, each with a master's degree in a relevant field, answered all 4,016 questions through disjoint category assignments: one covered Cinematic Arts and Stage Performing Arts, and the other covered Static Visual Arts and Game Arts. Each question received one response. Applying the same scoring rules used for models to their combined responses gave 87.18% Overall ACC, 90.98% Single CAA, and 78.00% Multi EM.

Non-expert comparison. In an auxiliary study, we randomly sampled 500 questions and asked two participants without formal training in the covered art domains to answer them. The non-expert participants, domain experts, and Claude-4.6-Opus were evaluated on the same subset with the same scoring protocol. Table 9 reports the individual non-expert scores rather than pooling two participants into a single estimate.

Table 9: Human and model results on the same random 500-question subset (%).

Evaluator	Overall ACC	Single CAA	Multi EM	Multi F1
Non-expert participant 1	54.60	65.57	36.44	74.16
Non-expert participant 2	42.80	51.74	23.89	68.32
Domain experts	85.40	90.35	78.14	93.70
Claude-4.6-Opus	49.20	55.44	33.60	70.45

Scoring Rubric. For every item, raters score the four dimensions on a 0–5 Likert scale, with the one-line definition and anchor descriptions shown beside each slider.

- **Holistic Quality** – whether the item, taken as a whole, exemplifies a high-quality benchmark question. Anchors: 5 = publication-ready; 3 = acceptable with minor edits; 1 = fundamentally flawed.
- **Visual Necessity** – whether answering requires watching the visual content rather than reading text alone. Anchors: 5 = stem points to a specific on-screen cue and cannot be answered without watching; 3 = cue exists but the transcript or a single frame suffices; 1 = no audiovisual anchor.
- **Mechanistic Trace** – whether the correct option exhibits a chain of the form *technique* $\rightarrow$ *effect* $\rightarrow$ *intent*. Anchors: 5 = names a specific craft object and binds it to a causal explanation; 3 = craft term appears with a vague effect; 1 = only aesthetic labels with no mechanism.
- **Answer Integrity** – whether the option set is well-engineered. Anchors: 5 = unique supported correct option with diverse distractors (or, for multi-select, a closed correct subset); 3 = one weak distractor; 1 = wrong, duplicated, or missing correct option.

Raters receive a brief calibration walkthrough on three pre-selected pairs per category before scoring begins, and are instructed to reserve the extreme anchors (0 and 5) for unambiguous cases.

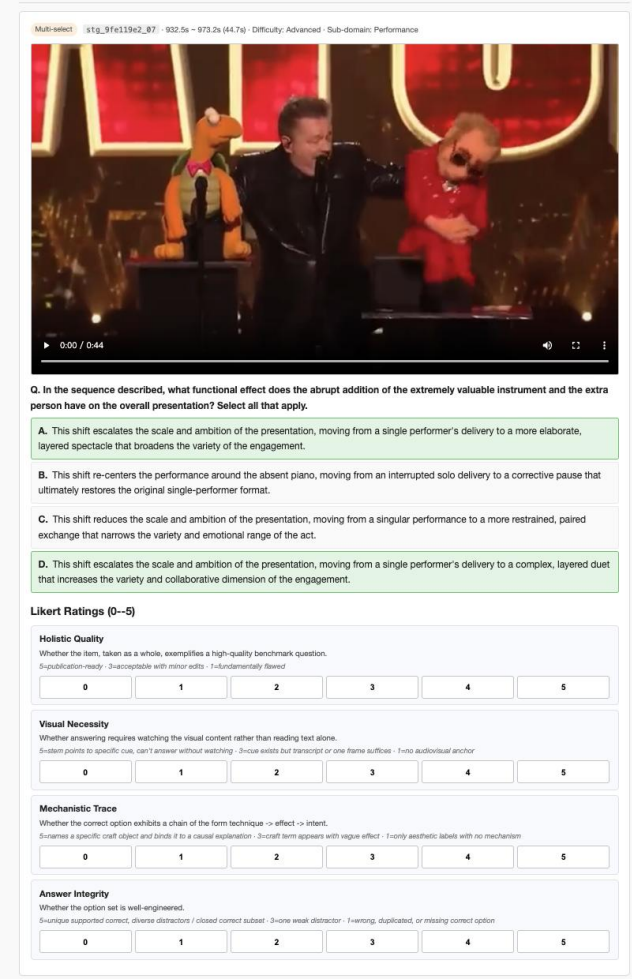


Figure 18: Human evaluation interface used by domain experts. The top panel shows the corresponding clip, the question, and the answer options with correct option(s) highlighted in green; the bottom panel shows the 0–5 Likert sliders for the four quality dimensions (Holistic Quality, Visual Necessity, Mechanistic Trace, Answer Integrity), each accompanied by its anchor descriptions.

E Detailed Evaluation Metrics

E.1 Single-Select Evaluation: Chance-Adjusted Accuracy

For single-select questions, the number of answer options varies across instances. As a result, raw accuracy is not directly comparable across questions with different random-guess baselines. To account for this, we report *chance-adjusted accuracy* (CAA), which measures performance relative to random guessing.

For the i-th single-select question, let K_i be the number of answer options and let $a_i = \mathbf{1}[\hat{y}_i = y_i]$ denote whether the model prediction $\hat{y}_i$ matches the ground-truth answer y_i. We define

$$c_i = \frac{1}{K_i}, \qquad \mathrm{CAA}_i = \frac{a_i - c_i}{1 - c_i} = \frac{\mathbf{1}[\hat{y}_i = y_i] - \frac{1}{K_i}}{1 - \frac{1}{K_i}}. \tag{2}$$

Here, c_i is the random-guess accuracy for question i. This normalization gives $\mathrm{CAA}_i = 1$ for a correct prediction, has expected score 0 under uniform random guessing, and yields negative scores for worse-than-chance performance.

Given N_{single} single-select questions, the reported score is

$$\mathrm{CAA} = \frac{1}{N_{\mathrm{single}}} \sum_{i=1}^{N_{\mathrm{single}}} \mathrm{CAA}_i. \tag{3}$$

In practice, this metric makes results more comparable across questions with different option counts, since achieving the same raw accuracy on a question with more options reflects stronger performance.

E.2 Multi-Select Evaluation: Exact Match, Precision, Recall, and F1

The primary multi-select metric is exact-match (EM) accuracy. For ground-truth option set Y_j and predicted set $\hat{Y}_j$, it is

$$\mathrm{EM} = \frac{1}{N_{\mathrm{multi}}} \sum_{j=1}^{N_{\mathrm{multi}}} \mathbf{1}[\hat{Y}_j = Y_j]. \tag{4}$$

EM gives credit only for the complete correct set. To diagnose partially correct predictions, we additionally report precision, recall, and F1 on the predicted answer set, using macro averages for the supplementary results.

For each multi-select question q_j, let $Y_j \subseteq \mathcal{O}_j$ denote the set of ground-truth correct options and let $\hat{Y}_j \subseteq \mathcal{O}_j$ denote the model-predicted set. We first define

$$\mathrm{TP}_j = |\hat{Y}_j \cap Y_j|, \qquad \mathrm{FP}_j = |\hat{Y}_j \setminus Y_j|, \qquad \mathrm{FN}_j = |Y_j \setminus \hat{Y}_j|. \tag{5}$$

The per-question precision, recall, and F1 are then

$$P_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j}, \qquad R_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FN}_j}, \qquad F1_j = \frac{2P_jR_j}{P_j + R_j}. \tag{6}$$

For macro averaging, we average the per-question scores:

$$M_{\mathrm{macro}} = \frac{1}{N_{\mathrm{multi}}} \sum_{j=1}^{N_{\mathrm{multi}}} M_j, \qquad M \in \{P, R, F1\}. \tag{7}$$

For completeness, micro averaging instead aggregates counts across all multi-select questions,

$$\mathrm{TP} = \sum_{j=1}^{N_{\mathrm{multi}}} \mathrm{TP}_j, \qquad \mathrm{FP} = \sum_{j=1}^{N_{\mathrm{multi}}} \mathrm{FP}_j, \qquad \mathrm{FN} = \sum_{j=1}^{N_{\mathrm{multi}}} \mathrm{FN}_j, \tag{8}$$

and then compute

$$P_{\mathrm{micro}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad R_{\mathrm{micro}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad F1_{\mathrm{micro}} = \frac{2P_{\mathrm{micro}}R_{\mathrm{micro}}}{P_{\mathrm{micro}} + R_{\mathrm{micro}}}. \tag{9}$$

Table 10: Zero-shot evaluation results on MUSEBENCH. For each art category, we report overall accuracy (ACC, %), chance-adjusted accuracy (CAA, %) for single-select questions, and precision / recall / F1 (%) for multi-select questions. Best results are in **bold**, second-best are underlined.

Model	Cine.					Static					Stage					Game					Overall
	ACC	CAA	P	R	F1	ACC	CAA	P	R	F1	ACC	CAA	P	R	F1	ACC	CAA	P	R	F1	ACC
Proprietary MLLMs																					
GPT-5.4 (Singh et al., 2025)	47.58	56.50	77.52	71.05	74.14	49.49	54.24	70.03	68.71	69.36	51.68	56.43	76.32	68.60	72.25	30.11	32.00	59.43	53.57	56.35	44.57
Claude-4.6-Opus (Anthropic, 2026)	50.20	63.26	**79.60**	70.68	**74.87**	52.93	58.51	**73.36**	72.60	**72.98**	**57.77**	**62.65**	**79.69**	74.87	**77.21**	**32.84**	34.07	**63.95**	55.78	59.59	**48.28**
Gemini-3.1-pro-preview (Team et al., 2023)	34.34	43.16	41.48	43.29	42.37	39.70	42.70	42.63	39.07	40.78	43.86	49.50	47.51	46.01	46.75	29.91	38.72	44.05	45.43	44.73	36.90
Grok-4.1 (xAI, 2025)	18.18	14.19	29.14	32.21	30.60	25.86	19.70	31.15	37.09	33.86	25.28	15.90	32.38	36.49	34.31	13.10	3.20	23.55	29.53	26.20	20.54
Qwen-3.5-Plus (Bai et al., 2025)	**51.11**	**68.88**	65.89	58.32	61.87	51.62	64.36	49.28	48.06	48.67	55.23	60.69	57.43	55.48	56.44	31.67	**38.80**	44.53	45.06	44.79	47.26
Doubao-Seed-1.8-Pro	48.38	62.10	78.07	69.01	73.26	**54.75**	**65.63**	70.88	66.22	68.47	50.56	56.84	74.28	65.35	69.62	31.28	32.86	60.99	54.99	57.84	46.12
GLM-4.5v (Hong et al., 2025)	23.03	16.17	33.52	35.69	34.57	13.94	-2.97	29.70	35.74	32.44	22.64	13.60	33.04	34.17	33.60	9.19	-4.34	21.26	27.90	24.13	17.13
Kimi-K2.5 (Team et al., 2026)	19.49	23.06	7.54	10.71	8.85	25.15	24.05	7.07	10.90	8.58	24.87	23.62	7.98	12.37	9.70	10.46	0.35	5.80	10.56	7.49	19.92
Open Source General Purpose MLLMs																					
Qwen3.5-397B-A17B (Bai et al., 2025)	50.91	62.65	56.88	55.30	56.08	49.60	57.45	47.30	50.83	49.00	49.44	56.60	42.30	45.13	43.67	29.62	35.79	30.33	35.95	32.90	44.77
Qwen2.5-Omni-7B (Xu et al., 2025)	34.55	36.47	65.45	63.97	64.70	39.19	40.76	58.47	62.30	60.33	38.38	34.43	62.81	60.44	61.60	19.16	8.31	51.79	50.92	51.36	32.69
InternVL3-78B (Chen et al., 2024b)	35.25	47.98	62.05	55.54	58.62	43.43	48.79	60.39	57.10	58.70	47.01	57.25	62.96	58.03	60.39	26.00	31.59	52.23	48.80	50.46	37.80
InternVL3-8B (Chen et al., 2024b)	40.71	43.41	67.64	72.12	69.81	40.71	36.23	64.54	70.64	67.45	33.60	30.48	59.08	66.23	62.45	17.79	6.66	59.40	45.06	51.25	33.07
LLaVA-OneVision-7B (Li et al., 2024a)	17.68	22.01	40.32	29.79	34.27	25.66	25.77	40.99	32.98	36.55	24.26	25.09	41.89	31.62	36.04	14.27	10.24	44.20	32.33	37.34	20.42
MiniCPM-o (Yu et al., 2026)	35.15	35.72	58.04	56.24	57.12	35.96	32.92	52.02	51.47	51.74	36.04	31.49	53.72	53.11	53.41	18.67	6.14	42.12	43.34	42.72	31.35
Gemma-4-E4B (Team et al., 2024)	30.71	39.40	60.79	47.25	53.17	33.33	32.55	56.03	47.69	51.52	30.76	31.44	60.13	47.02	52.77	16.03	10.30	45.12	39.85	42.32	27.61
Open Source Video-Specific MLLMs																					
VideoLLaMA2 (Cheng et al., 2024)	22.42	31.35	39.80	32.86	36.00	21.82	18.07	35.35	30.02	32.47	24.87	25.34	48.60	37.02	42.02	12.51	5.40	35.71	28.27	31.55	20.34
VideoLLaMA3 (Zhang et al., 2025a)	27.88	34.37	56.67	55.28	55.97	28.79	24.55	48.82	51.07	49.92	32.28	33.76	53.29	50.83	52.03	20.04	14.02	46.65	47.61	47.12	27.19
Video-R1 (Feng et al., 2025)	23.74	30.50	55.80	61.21	58.38	30.10	28.70	53.53	65.63	58.97	34.92	38.49	56.04	62.37	59.03	18.48	13.87	47.81	51.12	49.41	26.72
LongVU (Shen et al., 2024)	14.85	14.40	54.52	**87.69**	67.24	15.66	6.50	52.39	**86.35**	65.21	14.82	5.46	53.03	**89.52**	66.60	14.17	7.75	53.60	**91.65**	**67.64**	14.87
VideoRFT (Wang et al., 2025a)	23.74	26.50	54.44	50.10	52.18	31.21	30.14	51.28	54.47	52.83	33.10	36.04	54.47	51.23	52.80	16.81	9.01	46.46	46.32	46.39	26.12
VideoChat-R1 (Li et al., 2025)	25.86	35.87	53.18	47.66	50.27	31.11	29.05	55.69	52.83	54.22	30.56	33.04	53.46	42.02	47.06	17.11	6.46	44.80	42.78	43.77	26.07
VideoChat2 (Li et al., 2024b)	15.96	17.20	44.59	31.77	37.11	19.29	14.35	34.36	32.44	33.38	21.42	18.67	41.18	32.98	36.63	14.57	10.45	38.05	28.76	32.76	17.78
Video-XL-2 (Shu et al., 2024)	20.30	28.63	42.59	32.13	36.63	27.07	29.29	42.47	35.12	38.45	31.88	40.42	43.73	33.33	37.83	17.69	19.17	41.75	31.42	35.86	24.18
AKS (Tang et al., 2025)	16.06	17.61	34.57	24.95	28.98	22.83	21.46	33.56	30.01	31.68	25.79	28.01	45.22	32.85	38.06	12.81	6.34	34.54	28.74	31.38	19.30
Q-Frame (Zhang et al., 2025b)	19.49	13.81	57.08	62.78	59.80	22.73	13.83	53.80	62.64	57.89	17.06	3.30	55.35	72.98	62.96	15.84	8.19	54.96	60.80	57.73	18.75
LongVT (Yang et al., 2025b)	17.58	17.99	32.34	30.70	31.50	26.26	21.61	46.06	41.90	43.88	25.18	21.92	49.74	40.09	44.39	13.29	5.02	25.93	25.30	25.61	20.52
Video-CCAM (Fei et al., 2024)	18.18	23.38	38.76	28.78	33.04	21.21	18.31	37.63	32.25	34.73	20.30	16.92	37.76	29.30	33.00	10.65	1.05	38.78	29.94	33.79	17.53
TimeChat (Ren et al., 2024)	14.14	12.27	49.04	63.50	55.34	16.87	9.66	42.52	57.33	48.83	14.31	4.61	43.50	52.76	47.69	12.41	5.04	45.60	49.31	47.38	14.42

Compared with exact-match accuracy, these set-based metrics provide a more informative diagnosis of model behavior. In particular, precision captures the tendency to select incorrect distractors, while recall reflects whether the model misses valid analytical perspectives. F1 balances both aspects into a single score.

E.3 Overall Accuracy

Overall ACC is the fraction of fully correct answers across both question formats:

$$\mathrm{ACC} = \frac{\sum_{i=1}^{N_{\text{single}}} a_i + \sum_{j=1}^{N_{\text{multi}}} \mathbf{1}[\hat{Y}_j = Y_j]}{N_{\text{single}} + N_{\text{multi}}}. \tag{10}$$

Each question has equal weight. CAA is reported separately and is not used to compute Overall ACC. All tabulated scores are expressed as percentages.

F Complete Experimental Results

Compute resources. All open source MLLMs reported in Tables 1 and 10 are evaluated on a single internal cluster equipped with NVIDIA A800 GPUs, while proprietary systems (e.g., GPT-5.4, Claude-4.6-Opus, Gemini-3.1-pro-preview) are queried through their official APIs. Each system is evaluated zero-shot on the full MUSEBENCH test set in a single pass.

Table 10 expands each art category into its full per-category ACC, CAA, precision, recall, and F1 columns. Two patterns are worth re-emphasizing at this resolution. First, per-category ACC orders all four art categories almost identically across rows: Stage Performing Arts and Cinematic Arts are most accessible (Claude-4.6-Opus 57.77% Stage ACC and 50.20% Cinematic ACC; GPT-5.4 51.68% and 47.58%; Doubao-Seed-1.8-Pro 50.56% and 48.38%) while Game Arts trails by 15 to 25 points for nearly every system (Claude 32.84%, GPT-5.4 30.11%, Doubao 31.28%, Qwen3.5-397B-A17B 29.62%). Second, the precision-over-recall asymmetry observed in the main results holds within every category: for instance, GPT-5.4 on Cinematic multi-select reaches 77.52% precision against 71.05% recall, and Claude reaches 79.60% vs. 70.68%, while LLaVA-OneVision-7B on Static multi-select shows the most extreme version of the same skew (40.99% vs. 32.98%). Video-R1 again breaks the trend, with recall (61.21% Cinematic, 65.63% Static, 62.37% Stage, 51.12% Game) consistently above precision in all four categories.

F.1 Language and Option-Set Shortcut Controls

We ran two additional input controls for VideoLLaMA2 and Qwen2.5-Omni-7B: options only, and transcript plus question and options. The first probes answer-option priors; the second tests whether the source narration can support answers without audiovisual input. Decoding settings, item order, option order, and scoring were held fixed across these controls. Table 11 reports only the added conditions from this auxiliary experiment. These scores are standalone diagnostics and are not used to compute differences against the updated full-input leaderboard.

Table 11: Additional language and option-set controls from the auxiliary experiment (%).

Model	Input	Overall ACC	Single CAA	Multi EM
VideoLLaMA2	Options only	11.05	0.39	0.05
VideoLLaMA2	Transcript + question + options	11.30	0.92	0.12
Qwen2.5-Omni-7B	Options only	17.83	9.53	6.39
Qwen2.5-Omni-7B	Transcript + question + options	21.47	14.51	9.73

G Broader Impact

MuseBench targets a capability that is currently underrepresented in MLLM evaluation, namely intent-level reasoning about audiovisual artistic expression across cinematic, static visual, stage performing, and game arts. By exposing the gap between expert artistic understanding and current models, the benchmark can guide future training corpora and instruction tuning toward richer cultural and stylistic supervision. It can also serve as a standardized probe for arts education tools, accessibility applications such as audio description for visually impaired audiences, and content analysis pipelines used by archivists, educators, and independent critics.

The gold answers are anchored to interpretations presented in individual source video essays. They should therefore be read as source-supported analyses within a controlled answer space, rather than as exhaustive or uniquely correct interpretations of an artwork. The source-language and regional distribution in Table 6 also limits which artistic traditions are represented. We report those distributions so users can interpret model scores within the benchmark's present scope.

H ADDITIONAL EXAMPLES OF MUSEBENCH

This section showcases additional samples spanning all four categories of MUSEBENCH. Each card displays the eight uniformly sampled frames that form the visual prompt, the multiple-choice question, and all answer options, with the correct option(s) highlighted in red. The header bar reports the question type (single- or multi-select), the number of options, and the sub-domain. Across categories, the pairs consistently demand inferential reasoning grounded in the on-screen evidence rather than surface recognition: models must explain *why* a creative choice produces a particular effect, not merely *what* appears in the frame.

Game Arts. Fig. 19 presents Game Arts pairs that probe analytical reading of boss design, level pacing, narrative cinematography, and audiovisual signaling, requiring models to reason about the design intent behind each visual choice.

Cinematic Arts. Fig. 20 highlights Cinematic Arts pairs that span shot composition, lighting, editing rhythm, and dialogue pragmatics, emphasizing inferential judgments about how a director's choice yields a particular dramatic or compositional effect.

Stage Performing Arts. Fig. 21 presents Stage Performing Arts cases that test understanding of staging, choreography, lighting design, and audio-visual coordination in live performance, where the answer hinges on how staged elements jointly construct meaning.

Static Visual Arts. Fig. 22 features Static Visual Arts pairs covering composition, color theory, technique attribution, and aesthetic evaluation across painting, photography, and illustration, requiring fine-grained discrimination among visually similar artistic choices.

Game Arts

Single-Select | 4 Options Advanced | Interactive Visuals



Q: During the boss encounter shown, how do changes in the arena guide the player's next move?

A. Baiting the boss into descending by staying near the rising hazard threshold

B. Chasing the boss quickly to higher platforms before the safe space disappears

C. Memorizing the platform spawn pattern to predict the next safe route upward

D. Timing jumps between platforms to avoid telegraphed attacks from above

(a) *Game Arts* sample 1.

Game Arts

Single-Select | 5 Options Advanced | CG



Q: How does the segment frame the contrasting outlooks of the two central characters?

A. It presents Joel's distrust as the emotionally safer position, while Ellie's faith in people is treated as naïve, so the segment ends up endorsing the hard lesson that kindness usually gets people killed.

B. It frames Joel's distrust as the more justified worldview, while Ellie's trust in others is treated as immaturity, so her growth reads as learning to accept that goodwill has little place in their world.

C. It suggests that Joel's skepticism and Ellie's hope are both presented as incomplete, with the text mainly using their contrast to show how each of them needs the other to find a workable balance.

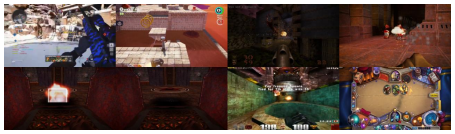
D. It frames Joel's distrust as the more justified worldview, while Ellie's trust in others is treated as a dangerous weakness, so her growth reads as learning to accept that goodwill has little place in their world.

E. It argues that Ellie's belief in others is the healthier outlook, while Joel's distrust is portrayed as a sad but understandable defense, making the segment primarily sympathetic to her optimism.

(b) *Game Arts* sample 2.

Game Arts

Single-Select | 8 Options Advanced | Interactive Visuals



Q: What gameplay purpose does the relationship between the explosion's visible size and its area of impact serve?

A. So the invisible splash damage is adjusted to match the fireball's appearance, rather than the fireball being built to match the hidden area.

B. So the larger-looking fireball automatically tells players that the impact reaches a wider zone than the engine actually applies.

C. So players can understand the rocket's travel path more clearly and predict where it will land before the target is struck.

D. So players can judge the blast's danger accurately and make reliable choices from the visual cue alone.

E. So the visible explosion and the hidden splash area both reinforce how dangerous the hit feels, even if the visual size is only approximate.

F. So the fireball signals the rocket's overall visual intensity rather than the exact area where the impact effect is applied.

G. So a more dramatic fireball naturally makes the attack more believable and therefore more effective in gameplay.

H. So people can assess the threat correctly and make dependable decisions based on what they see.

(c) *Game Arts* sample 3.

Game Arts

Single-Select | 6 Options Advanced | CG



Q: What atmosphere do the architecture, surface materials, and lighting create in the environments shown?

A. The visual strategy emphasizes a deliberate lack of history to immerse the observer in an atmosphere of artificial hyperreality, creating an oppressive and soulless aesthetic through uniform facades and stark lighting.

B. The visual strategy emphasizes a deliberate lack of history to immerse the observer in an atmosphere of artificial hyperreality, creating an oppressive and soulless aesthetic through uniform facades and shadow-drenched lighting.

C. The visual strategy emphasizes a deliberate lack of history to immerse the observer in an atmosphere of artificial hyperreality, creating an oppressive and soulless aesthetic through modern facades and razor-sharp shadows.

D. The visual strategy emphasizes a deliberate lack of history to immerse the observer in an atmosphere of artificial hyperreality, creating a sleek and aspirational aesthetic through uniform facades and stark lighting.

E. The visual strategy emphasizes the city's architectural heritage to immerse the observer in a richly historical atmosphere, creating a grounded and lived-in aesthetic through varied facades and warm natural lighting.

F. The visual strategy emphasizes the absence of visual variety to immerse the observer in a clean hyperreal atmosphere, creating a sterile and impersonal aesthetic through identical facades and intense illumination.

(d) *Game Arts* sample 4.

Figure 19: Additional *Game Arts* samples from MUSEBENCH.

Cinematic Arts

Multi-Select | 4 Options — Advanced | Mise-en-scène



Q: How does the treatment of spatial depth in the shots shown shape the relationship between the figures and their surroundings? *(Multi-select)*

A. It generates a crowded atmosphere that makes opposing figures appear more trapped and confrontational.

B. It makes subjects appear to move more fluidly through the background, so the visual effect of spatial compression produces a sense of ease and forward momentum rather than friction.

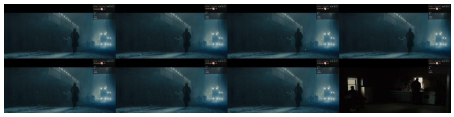
C. It reduces the perceived velocity of movement that occurs toward or away from the viewer.

D. It transforms the distance between subjects into a form of psychological pressure rather than a sense of open space.

(a) *Cinematic Arts* sample 1.

Cinematic Arts

Multi-Select | 6 Options — Advanced | Editing & Pacing



Q: What effects does the arrangement of light and dark areas create in the images shown? *(Multi-select)*

A. It facilitates a smooth transition for the viewer's focus to move progressively from the foreground details toward the distant horizon.

B. It maximizes the perception of spatial depth by utilizing high-contrast transitions to clearly delineate the boundaries between different structural planes.

C. It produces a rhythmic, harmonious feeling that gives the content a sense of musicality.

D. It makes the central figure stand out by placing them in a darker middle area between brighter foreground and background elements.

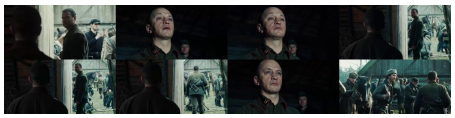
E. It establishes a specific narrative tempo that aligns the visual pace with the internal psychological development of the central figure.

F. It maintains a sense of equilibrium by ensuring the brightest values are concentrated in the right-side background and left-side foreground.

(b) *Cinematic Arts* sample 2.

Cinematic Arts

Multi-Select | 4 Options — Advanced | Mise-en-scène



Q: In the closing utterance, what inference best captures the interaction between the speaker's formal wording and the pragmatic force of the statement? *(Multi-select)*

A. The formal expression of gratitude creates a sharp contrast that reveals the speaker's underlying resentment or moral detachment rather than sincere appreciation.

B. The acknowledgment of 'honor' functions ironically, using the polished phrasing to expose a sense of hypocrisy rather than offer genuine appreciation.

C. The formal expression of gratitude reinforces the speaker's genuine commitment to the collective, using the polished phrasing to signal loyalty to the group rather than personal detachment.

D. The formal expression of gratitude maintains professional distance from personal emotion, ensuring the speaker's composure signals institutional discipline rather than any undertone of irony or resentment.

(c) *Cinematic Arts* sample 3.

Cinematic Arts

Single-Select | 5 Options — Advanced | Mise-en-scène



Q: How does the timing of cuts between the speakers and their reactions shape the exchange shown?

A. By breaking the conventional rhythm of exchange, it continuously reinforces the confrontation and pressure between the two parties, transforming relational conflict into psychological tension the audience can directly perceive.

B. By cutting earlier to the other party's response, the emphasis is not on intensifying the conflict but on more efficiently delivering relational background and expository information.

C. By inserting the other party's reaction shot before one side has finished speaking, it preserves the confrontational atmosphere while letting the audience more clearly perceive the suppressed emotional tension between the characters.

D. Through this preemptive reaction-shot organization, the hostility in the dialogue is extended into the central driving force of the entire narrative structure, hinting that subsequent conflict will inevitably escalate into total opposition.

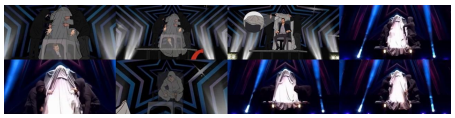
E. By interrupting the originally coherent narrative rhythm, the informational focus within the same scene is shifted earlier onto the other party's reaction, allowing the audience to grasp each side's stance and emotional shift more quickly.

(d) *Cinematic Arts* sample 4.

Figure 20: Additional *Cinematic Arts* samples from MUSEBENCH.

Stage Performing Arts

Single-Select | 5 Options — Advanced | Performance



Q: How does the staging of the action around the chair create the illusion shown?

A. It demonstrates the use of a proxy to facilitate a misattribution of agency, ensuring the perceived continuity of the primary subject's presence.

B. It demonstrates the use of a surrogate performer to sustain the deception, ensuring the perceived continuity of the magician's identity despite the assistant doing the actual visible action.

C. It demonstrates the use of a concealed assistant to maintain the illusion, ensuring the audience keeps attributing the visible actions to the magician at the center.

D. It demonstrates the use of a hidden stand-in to produce the illusion, ensuring the apparent subject is seen as fully controlling the sequence from behind the chair.

E. It demonstrates that the magician's direct movements are the essential mechanism of the effect, ensuring the audience interprets the blanket as merely covering his visible performance.

(a) *Stage Performing Arts* sample 1.

Stage Performing Arts

Single-Select | 5 Options — Advanced | Performance



Q: How does the background performer's response to the closing sung lines shape the meaning of this moment?

A. It functions as a gesture of deep reverence for the speaker's conviction, signaling the figure's shift from a position of authority to one of humble support for the next generation.

B. It serves as a visual acknowledgment of moral failure, transforming the spoken sentiment into a personal indictment that the figure accepts without defense.

C. It represents a public pledge of reform, signaling that the figure is committing to use his remaining influence to address the moral failures highlighted in the closing reflection.

D. It serves as a display of dignified composure, suggesting that while the figure hears the criticism, he maintains stoic neutrality to preserve the solemnity of the moment without taking a personal stance.

E. It acts as a solemn tribute to the lost potential described in the lyrics, framing the spoken sentiment as a collective expression of mourning that the figure shares with the entire ensemble.

(b) *Stage Performing Arts* sample 2.

Stage Performing Arts

Single-Select | 4 Options — Advanced | Stage Design



Q: How does the change in staging between the central character's workplace and home convey her agency?

A. The visual approach divides Jessica's life into two different power structures where, despite the public area concentrating her authority, the domestic setting reveals that the burden of labor in the home ultimately becomes the source of her most significant personal agency.

B. The visual approach divides Jessica's life into two different power structures: the public-facing area concentrates attention on her as commanding and sovereign, while the home area disperses attention across family objects and tasks to show that, despite her greatest labor being there, it is not fully under her control.

C. The visual approach divides Jessica's life into two different power structures where the home area, despite dispersing attention across objects and tasks, still frames those demands as an extension of Jessica's authority rather than as a domain operating beyond her control.

D. The visual approach divides Jessica's life into two different power structures: the public-facing area concentrates attention on her as commanding and sovereign, while the home area shows that the domestic sphere, despite its demands, is where she exercises the most practical day-to-day authority.

(c) *Stage Performing Arts* sample 3.

Stage Performing Arts

Single-Select | 6 Options — Advanced | Performance



Q: What parallel in the development of the two acrobatic traditions is illustrated by the performances presented?

A. It argues that the Roman and Greek roots preserved a ground-level form of acrobatics, while the Chinese circus later transformed it into a theatrical dance tradition without any comparable aerial component.

B. It posits that Western and Chinese traditions followed a similar trajectory by elevating ground-based physical feats onto elevated ropes, establishing a shared dramatic spectacle across distinct historical periods.

C. It concludes that both lineages originated in the same historical era and spread together through direct cultural borrowing, making their later similarities the result of one unified performance tradition.

D. It suggests that the key parallel between the two traditions lies in their shared use of audience-facing spectacle rather than in any specific technical transition from ground-based to elevated performance.

E. It maintains that the decisive change was the introduction of medieval fair entertainment in Europe and imperial court pageantry in China, which then independently produced the later high-wire tradition.

F. It claims the comparison is mainly about how both traditions used rope-based performance to display balance and risk, rather than about any historical shift in the type of acrobatics being performed.

(d) *Stage Performing Arts* sample 4.

Figure 21: Additional *Stage Performing Arts* samples from MUSEBENCH.

Static Visual Arts

Single-Select | 4 Options — Advanced | Fine Art



Q: What role do the light and color principles demonstrated in this segment play in artistic practice?

A. They are treated as technical prerequisites that ensure artists can depict light and color with scientific accuracy, so that materials and natural environments in the finished work look as close to reality as possible.

B. They are introduced as a preliminary framework for organizing compositional structure, so artists can balance perspective, strengthen spatial depth, and arrange the overall layout of a scene.

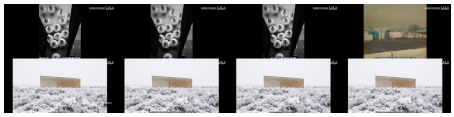
C. They are presented as a basis for mastering the perceptual impact of light and color on viewers, so that artists can design images that consistently trigger specific emotional and attentional responses regardless of the subject matter.

D. They are presented as the foundation for understanding how to use light and color more deliberately, so artists can steer visual focus, shape atmosphere, and accurately depict materials and outdoor scenes.

(a) *Static Visual Arts* sample 1.

Static Visual Arts

Single-Select | 4 Options — Advanced | Photography



Q: How does the arrangement of shapes and perspective guide the viewer's reading of the composition shown?

A. It creates a compositional illusion that, when viewed purely as structure and object, forces the representation of people to depend on imagination rather than direct observation.

B. It creates a compositional illusion that guides the viewer's eye toward the dancers as a central focal point, emphasizing their physical movement and presence over the geometric structure.

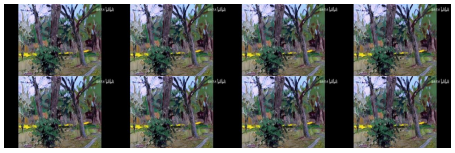
C. It creates a compositional illusion that prompts the viewer to reconstruct the region's lost social history through a fictional lens, using the geometry to imagine the more populated and joyful past of the site.

D. It creates a compositional illusion that heightens the historical contrast between the site's joyful past and its current stillness, prompting the viewer to document the visible physical decay rather than engage with the geometry.

(b) *Static Visual Arts* sample 2.

Static Visual Arts

Single-Select | 4 Options — Advanced | Photography



Q: Which principle best explains the relationships among colors in the composition shown?

A. Chromatic identity is fixed by the most immediately visible pairing; a single value functions as a stable anchor within one dominant contrast while excluding other complementary tensions.

B. Chromatic identity is inherently fluid; a single value functions as a transitional element within one prioritized contrast while simultaneously reinforcing a single, exclusive complementary tension.

C. Chromatic identity is inherently fluid; a single value functions as a transitional element within one prioritized contrast while simultaneously participating in multiple, overlapping neighboring relations.

D. Chromatic identity is inherently fluid; a single value functions as a transitional element within one prioritized contrast while simultaneously participating in multiple, overlapping complementary tensions.

(c) *Static Visual Arts* sample 3.

Static Visual Arts

Single-Select | 6 Options — Advanced | Photography



Q: For the plant-breeding artwork shown, what makes the creative process difficult to carry out as a fixed sequence of steps?

A. The procedure is resistant to standard execution because the breeding results are determined by environmental conditions alone, so the observed variation mainly reflects external growth factors rather than the crossing process itself.

B. The key constraint is the presence of many intermediate selection states, which means the project must keep sampling until the eventual prototype can be standardized through a fixed regulatory framework.

C. The inherent constraint is the absence of a predefined target genotype, which makes the workflow exploratory and highly dependent on finding promising offspring among many unusable candidates instead of applying a linear breeding schedule.

D. The defining constraint is the inability to forecast hybridization results in advance, but the repetitive approach it forces serves mainly to accumulate enough discarded specimens to statistically identify a stable line rather than to discover unexpected favorable variants.

E. The main issue is that the plants already contain a random genome, so the hybridization step itself becomes standardized only after the desired variety has been isolated and formally approved.

F. The central limitation is the inability to forecast hybridization results in advance, which produces many discarded specimens and forces a repetitive approach that searches for unexpected favorable variants rather than following a fixed development path.

(d) *Static Visual Arts* sample 4.

Figure 22: Additional *Static Visual Arts* samples from MUSEBENCH.