REELBENCH: TOWARDS SYSTEMATIC EVALUATION OF MULTI-SHOT NARRATIVE VIDEO GENERATION UNDER IMAGE REFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-shot narrative video generation increasingly relies on image references, including asset and storyboard images, to maintain visual and narrative consistency across independently generated shots. Yet existing benchmarks primarily assess final video quality or rank model capabilities, offering limited insight into how different forms of image conditioning affect generation. We introduce ReelBench, an evaluation framework and dataset for image-guided multi-shot narrative video generation. ReelBench comprises 300 story-style pairs with associated asset and storyboard images across three visual styles: Pixar-style animation, live-action realism, and 2D anime. It supports controlled comparisons among image reference conditions within the same story and evaluates generated videos along seven dimensions and 40 subdimensions: instruction alignment, character quality, visual quality, physical plausibility, visual continuity, cinematic quality, and narrative coherence. Our experiments show that stronger visual anchoring through first-frame conditioning does not consistently improve overall quality, and that reference complexity affects physical plausibility more strongly than overall scores. ReelBench agrees with human overall preferences in over 90% of pairwise comparisons.

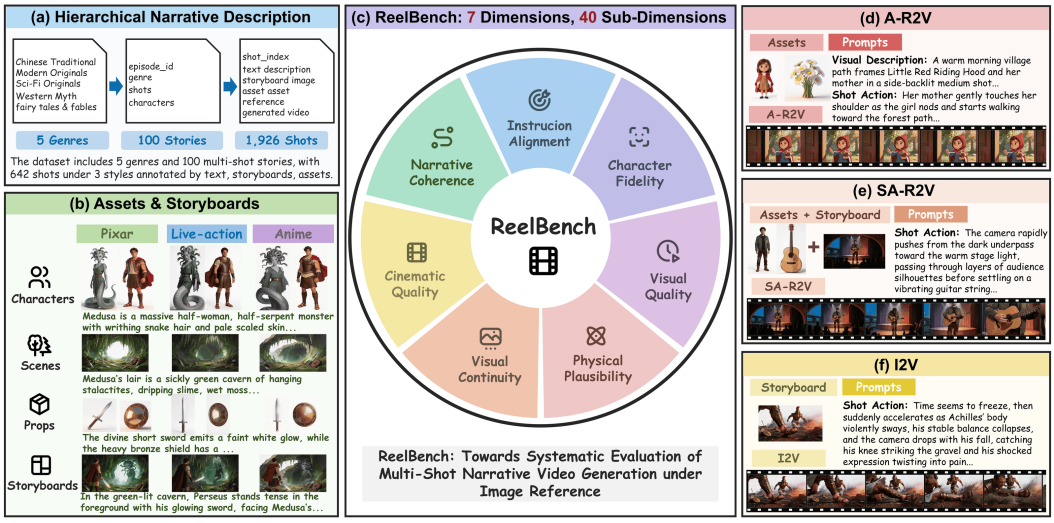

Figure 1: Overview of ReelBench and the controlled image reference conditions. ReelBench organizes narrative video generation data hierarchically from genre to story and shot, where each shot is paired with text descriptions, assets, storyboards, and generated videos. The benchmark evaluates image-reference-conditioned multi-shot narrative videos across 7 dimensions and 40 sub-dimensions, covering shot-level instruction alignment, character quality, visual quality, and physical plausibility, as well as sequence-level visual continuity, cinematic quality, and narrative coherence. We compare three controlled reference settings: A-R2V uses asset references only; SA-R2V combines assets and storyboards; and I2V uses the storyboard as an image reference or first-frame reference.

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.

1 INTRODUCTION

Recent video generation models such as Sora (Liu et al., 2024c; Brooks et al., 2024), Wan (Wan et al., 2025), and Vidu (Bao et al., 2024) have made it increasingly practical to construct multi-shot narrative videos from independently generated shots. These workflows support AI animation, short-form drama generation, interactive storytelling, and virtual-production previsualization (Zhou et al., 2024; Xu et al., 2025; Cheng et al., 2024). A typical pipeline decomposes a script into shot descriptions, generates each shot using image references, and assembles the results into a narrative sequence. Image references serve distinct roles in this process: asset images specify the appearance of characters, scenes, and props; storyboard images encode composition, subject placement, and cinematic intent; and a storyboard used as the first frame anchors the initial visual state of a shot (Zhang et al., 2025b).

Despite their central role in these pipelines, the effects of image references on multi-shot narrative video generation remain poorly understood. It is unclear which forms of reference improve generation, whether their benefits persist across visual styles and model capabilities, and how those benefits change with asset complexity. These questions directly inform reference design and model selection, but existing benchmarks are not designed to answer them systematically.

Benchmarks such as VBench (Huang et al., 2024), EvalCrafter (Liu et al., 2024b), and Video-Bench (Han et al., 2025) primarily assess individual shots, offering limited coverage of cross-shot viewing experience, cinematic quality, and narrative coherence. MSVBench (Shi et al., 2026) and UniVBench (Wei et al., 2026a) extend evaluation to multi-shot settings: MSVBench introduces hierarchical data organization and a hybrid evaluation framework, while UniVBench spans video understanding, generation, editing, and reconstruction. Studies of long-form generation further highlight challenges in character identity preservation and scene-level temporal consistency (Xie et al., 2026; Babu et al., 2025). Yet existing evaluations still center on final output quality or model ranking and rarely treat image references as controlled experimental variables.

Existing visual-story and multi-shot datasets pose a related limitation (Gupta et al., 2018; Huang et al., 2016; Li et al., 2019; Liu et al., 2024a; Yang et al., 2025b; Ye et al., 2024). ViStoryBench (Zhuang et al., 2025) constructs multi-shot story scripts, MSVBench (Shi et al., 2026) organizes hierarchical story-shot evaluation data, and UniVBench (Wei et al., 2026a) provides multi-shot videos for unified cross-task evaluation. However, they do not provide parallel samples that vary the use of asset images, storyboards, and storyboard first frames while holding the story and shot description fixed (Zhang et al., 2025a; He et al., 2025). Without such controlled samples, it is difficult to isolate how image reference conditions, visual style, and asset complexity affect generation quality.

Table 1: Comparison of ReelBench with existing benchmarks. Reference control types: ① asset images (characters, scenes, props); ② storyboard images (composition and subject placement); ③ storyboard-as-first-frame (I2V).

Benchmark	Scale	Eval Granularity	Ref. Control	Multi-Shot	Cross-Shot	Style / Complexity
VBench (Huang et al., 2024)	–	16 dims, shot	×	×	×	×
CineTechBench (Wang et al., 2025)	120 clips	12 dims, shot	③	×	×	×
OpenS2V-Nexus (Yuan et al., 2025)	180 prompts	6 dims, shot	①	×	×	×
ViStoryBench (Zhuang et al., 2025)	80 stories	5 dims, story	①	✓	✓	✓
MSVBench (Shi et al., 2026)	20 stories	8 dims, shot+seq	①	✓	✓	×
UniVBench (Wei et al., 2026a)	200 videos	4 tasks, shot	①	✓	✓	×
	300 story-style pairs					
ReelBench (Ours)	**1,926 shots**	**7 dims, 40 sub**	①②③	✓	✓	✓

We introduce ReelBench, an evaluation framework and dataset designed to study image references in multi-shot narrative video generation. The dataset comprises 300 story-style pairs and 1,926 shots, with aligned text descriptions, asset images, storyboard images, and videos generated under multiple reference conditions. It spans three visual styles: Pixar-style animation, live-action realism, and 2D anime. ReelBench evaluates each story at both the shot and sequence levels using seven dimensions and 40 subdimensions.

We use ReelBench in two main experiments. A controlled reference-condition comparison evaluates SA-R2V, A-R2V, and I2V on the same 100 stories, isolating the effects of asset references, storyboard references, and storyboard first-frame conditioning. A cross-model evaluation examines how reference effectiveness varies with model capability, visual style, and asset complexity. An auxiliary storyboard-optimization analysis tests whether improvements in static reference quality transfer to

generated videos. Pairwise human preference studies show strong agreement between ReelBench scores and human judgments.

In summary, our main contributions are as follows.

- **Evaluation Framework.** ReelBench decomposes evaluation into shot-level and sequence-level tasks and uses dimension-specific structured prompts to produce fine-grained scores and diagnoses. Its automatic overall rankings agree with human preferences in over 90% of pairwise comparisons.
- **Dataset.** We construct 300 story-style pairs and 1,926 shots with associated text descriptions, asset images, storyboard images, and generation results across three visual styles and multiple reference conditions. The dataset supports controlled ablations and cross-model analysis, and both the dataset and evaluation code are publicly available.
- **Systematic Findings.** Stronger first-frame anchoring does not reliably improve generation quality, and the value of image references varies across models and visual styles. Increasing asset count has a modest effect on overall scores but a larger effect on physical plausibility. Improvements in static storyboard quality also fail to transfer consistently to generated videos, highlighting the need to account for motion execution and cross-shot state continuity when optimizing references.

2 Related Work

2.1 Single-Shot Video Evaluation

Early work relied on FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) to measure distributional differences, and CLIP-Score (Hessel et al., 2021) to measure image–text alignment, but these metrics lack fine-grained diagnostic capability. VBench (Huang et al., 2024) decomposed video quality into 16 disentangled dimensions, pioneering hierarchical video evaluation, while VBench-2.0 (Zheng et al., 2025) further expanded to creativity and physics dimensions. FETV (Liu et al., 2023) introduced fine-grained temporal-aware text categories, DEVIL (Liao et al., 2024) focused on motion evaluation, T2V-CompBench (Sun et al., 2025) targeted compositional text-to-video generation, and Video-Bench (Han et al., 2025) incorporated MLLMs as judges to improve human alignment. In cinematic evaluation, Stable Cinemetrics (Chatterjee et al., 2025) formalized film production control into 76 fine-grained nodes, while CineTechBench (Wang et al., 2025) covered seven cinematographic dimensions including shot scale, camera movement, and lighting. However, all these benchmarks evaluate isolated single-shot videos and therefore cannot capture cross-shot viewing experience, cinematic quality, and narrative coherence in multi-shot narrative videos.

2.2 Multi-Shot Story Video Evaluation

OpenS2V-Nexus (Yuan et al., 2025) introduces subject-reference conditioning, while ViStory-Bench (Zhuang et al., 2025) provides a story visualization benchmark with character reference images; they focus on subject-driven video generation and story image sequences, respectively. MSVBench (Shi et al., 2026) targets multi-shot video generation and converts evaluation traces into supervision for lightweight judges; UniVBench (Wei et al., 2026a) jointly evaluates video understanding, generation, editing, and reconstruction. MSAVBench (Wei et al., 2026b) evaluates multi-shot audiovisual generation through an adaptive hybrid evaluation framework. ReelBench focuses on the role of image references, comparing asset images, storyboards, and storyboard first frames under the same stories and shot descriptions to analyze their effects on generation quality.

2.3 Multi-Shot Story Datasets

Existing multi-shot narrative datasets share several limitations in data construction. OpenS2V-Nexus (Yuan et al., 2025) centers its evaluation around subject references but lacks episode-level story structure and storyboard images. ViStoryBench (Zhuang et al., 2025) contains 80 stories, 344 characters, and 509 character reference images, but outputs image sequences instead of videos, making it unable to evaluate motion execution, camera movement, and cross-shot viewing experience. MSVBench (Shi et al., 2026) organizes 20 stories, 136 character references, and 276 shot

references into a hierarchical dataset for ranking generation systems, but does not treat storyboard images themselves as upstream quality variables or provide parallel reference-condition ablations. UniVBench (Wei et al., 2026a) provides 200 multi-shot videos centered on human-created content for unified evaluation across understanding, generation, editing, and reconstruction, but lacks paired "script–asset–storyboard–video" pipelines. ReelBench, built upon 300 story-style pairs and 1,926 shots, organizes textual descriptions, asset images, storyboard images, and multi-condition generation results for each shot simultaneously, enabling the disentanglement of model differences, reference-condition differences, and upstream image-quality differences within the same narrative unit. This transforms the dataset from a benchmark leaderboard resource into an experimental platform for studying the mechanisms of image reference conditioning.

3 REELBENCH FRAMEWORK

The ReelBench framework consists of three components: an evaluation dataset supporting controlled experiments (Section 3.1), a storyboard quality evaluation module for auxiliary analysis (Section 3.2), and a multi-dimensional video quality evaluation module (Section 3.3). Together, these components cover the evaluation pipeline from image reference inputs to multi-shot video outputs.

3.1 DATASET CONSTRUCTION

ReelBench organizes its data at three levels: genre, story, and shot. Each story belongs to one narrative category and contains an ordered sequence of 6–8 shots. Each shot is the smallest evaluation unit and is paired with a shot description, asset images, and a storyboard image. This hierarchy supports shot-level diagnosis, story-level aggregation, and evaluation across adjacent shots.

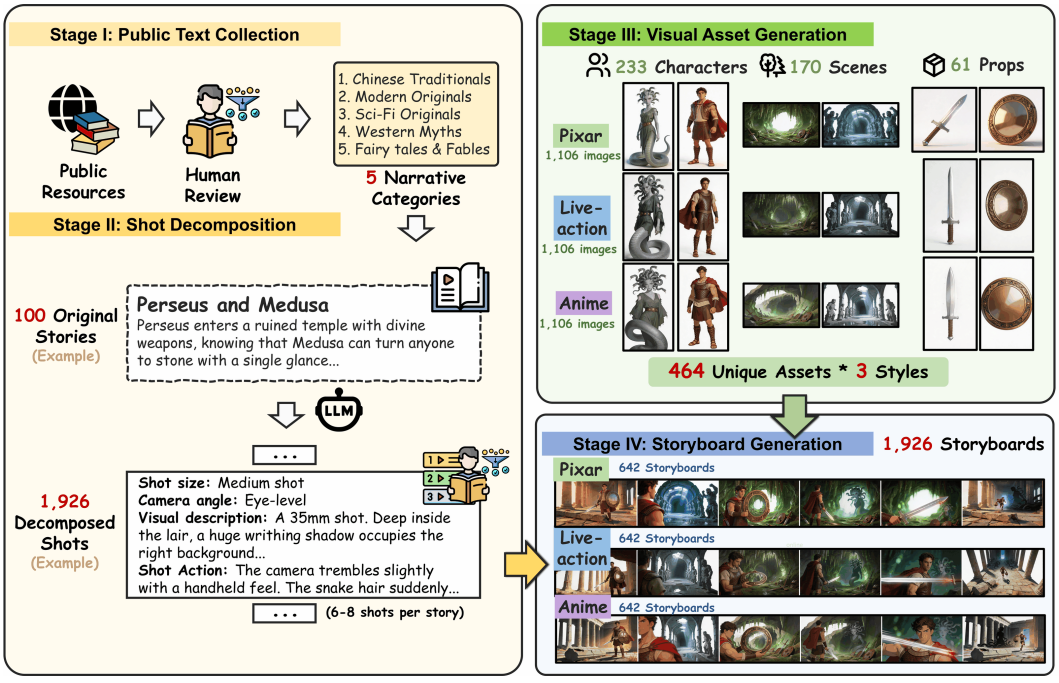

Figure 2: Dataset construction pipeline of ReelBench. The pipeline transforms publicly collected and manually curated stories into shot-level text descriptions, structured visual assets, storyboard references, and image-reference-conditioned video generations for evaluation.

Story Collection. The dataset spans five narrative categories: Chinese traditional stories, modern original stories, science-fiction original stories, Western mythology, and fairy tales and fables. We collect source stories from publicly available text resources and manually curate them for narrative completeness and coverage across categories.

Shot Decomposition. Gemini 3.0 Flash decomposes each story into 6–8 shot descriptions specifying actions, camera directions, scene settings, and visual style. We manually review the resulting descriptions for narrative coherence and suitability for video generation.

Asset Description Extraction. Gemini 3.0 Flash also extracts structured descriptions of the characters, scenes, and key props in each story. These descriptions establish the identity, appearance, and semantic constraints used to construct the visual references.

Visual Asset Generation. Doubao-Seedream-5.0 generates asset images from the character descriptions. The asset images and corresponding shot descriptions then condition storyboard generation, producing frames that instantiate the intended composition, character placement, and spatial layout. Style prompts are used to generate both asset and storyboard images in each of the three visual styles. We manually screen both image types for reference quality and usability before video generation.

Video Generation. We then generate each shot with multiple video generation models under three reference conditions: SA-R2V, A-R2V, and I2V. Section 4.1 provides the model and generation configurations.

3.2 Auxiliary Storyboard Quality Evaluation

We evaluate the static quality of storyboard images to analyze whether their optimization improves downstream video generation. Evaluation prompts and dimension definitions are provided in Appendices A.1 and A.3, respectively; optimization procedures and results are in Appendix E.

3.3 Multi-Dimensional Video Quality Evaluation

ReelBench evaluates video quality using seven dimensions and 40 subdimensions, grouped into shot-level and sequence-level criteria. The shot-level dimensions are *Instruction Alignment*, *Character Quality*, *Visual Quality*, and *Physical Plausibility*; the sequence-level dimensions are *Visual Continuity*, *Cinematic Quality*, and *Narrative Coherence*. Appendix A provides the complete prompts and subdimension definitions.

Holistic scoring can disperse a multimodal evaluator's attention across distinct aspects of video quality. ReelBench therefore assigns each dimension a separate task with a structured prompt and output schema. Shot-level tasks receive one shot, whereas sequence-level tasks receive a pair of adjacent shots. The resulting scores are aggregated at the story level, while the dimension-specific outputs localize errors in instruction execution, character identity, visual quality, physical interaction, and cross-shot narrative expression. Appendix A.2 provides the aggregation formula and weights.

(a) Shot-Level Quality

Instruction Alignment. (IA) Assesses how faithfully a video follows its shot description, including the specified actions, camera movement, scene setting, and visual style.

Character Quality. (CQ) Assesses characters and other subjects in terms of identity consistency, anatomical structure, clothing and props, facial expression and gaze, subject prominence, and rendering quality.

Visual Quality. (VQ) Assesses technical quality and temporal stability, including sharpness, artifacts, exposure and color consistency, frame-to-frame stability, motion smoothness, flicker, deformation, material rendering, and dynamic expressiveness.

Physical Plausibility. (PP) Assesses whether motion and interactions follow physical and commonsense constraints, including gravity, collision, contact, perspective, object penetration, lighting, and object-state consistency. These criteria also apply to stylized content while accounting for its intended visual conventions.

(b) Sequence-Level Coherence

Visual Continuity. (VC) Assesses whether character performance, visual detail, motion appeal, and visual atmosphere form a coherent viewing experience across adjacent shots.

Cinematic Quality. (Cine) Assesses cinematic expression across a shot sequence, including shot rhythm, visual guidance, compositional richness, scene layering, and execution of the intended cinematic design.

Narrative Coherence. (NC) Assesses whether a sequence conveys the intended narrative through clear visual storytelling, narrative relationships, and logical information progression across shots. It applies to linear narratives, parallel editing, montage, and flashbacks without prescribing a particular narrative form.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Reference Conditions. We evaluate three image reference conditions. SA-R2V (*Storyboard-and-Asset conditioned R2V*) uses both storyboard and asset images; A-R2V (*Asset-conditioned R2V*) uses asset images only; and I2V (*Image-to-Video*) uses a storyboard as either the first-frame condition or a strong visual reference, depending on the model interface. In the controlled comparison, all three conditions use the same stories, shot descriptions, and evaluation protocol; only the reference input and corresponding generation interface differ.

Dataset and Samples. The controlled comparison uses the 100-story Pixar-style subset of ReelBench and generates each story under SA-R2V, A-R2V, and I2V. The cross-model evaluation is stratified by visual style, story genre, and asset count. It samples 20 stories for each of Pixar-style animation, live-action realism, and 2D anime, yielding up to 60 story-style pairs per model. Because few open-source models support both multi-reference and I2V interfaces, the cross-model study is not fully crossed. It characterizes performance across the available combinations of models and reference interfaces rather than serving as a strict leaderboard.

Model Configurations. Doubao-Seedream-5.0 produces the asset and storyboard images. The controlled comparison fixes Vidu Q2 (Bao et al., 2024) as the video generation model to isolate the effect of the reference condition. The cross-model evaluation covers the combinations of models and reference interfaces listed in Table 3: Seedance 2.0 (Seedance et al., 2026), Vidu Q2 (Bao et al., 2024), Vidu Q3, Wan2.2-TI2V-5B (Wan et al., 2025), LTX-Video 2.3 (HaCohen et al., 2024), SkyReels-V3 (Li et al., 2026), and LTX-Video I2V. Vidu Q3 is evaluated only with I2V because its SA-R2V and A-R2V interfaces were unavailable at the time of the experiments. Closed platforms may reject a small number of generations through safety or copyright filters; we exclude these rejected generations from the corresponding model-specific analysis rather than replacing them with outputs from other models. All open-source models run in the same environment with four 80 GB NVIDIA A800 GPUs.

Evaluation Pipeline. Qwen3.5-Plus (Yang et al., 2025a) evaluates individual shots and adjacent shot pairs using the structured prompts in Section 3.3. Shot-level and adjacent-pair scores are aggregated into story-level results as described in Appendix A.2. On fixed subsets, SA-R2V retains the highest mean score in all three runs of the controlled experiment, while repeated cross-model scores have a Spearman correlation of 0.933 with the original scores. Appendix C.4 reports the repeatability settings and complete results.

4.2 CONTROLLED IMAGE REFERENCE CONDITION COMPARISON

This controlled experiment holds the stories, shot descriptions, video model, and evaluation protocol fixed across 100 stories while varying the image reference condition. It therefore isolates the effects of asset references, storyboard references, and storyboard first-frame conditioning on multi-shot narrative video quality.

With Vidu Q2 on the Pixar-style subset, SA-R2V achieves an overall score of 70.40, compared with 67.47 for A-R2V and 66.95 for I2V. Story-paired tests show that SA-R2V significantly outperforms both alternatives, whereas the difference between A-R2V and I2V is not significant. This pattern supports the complementary roles of storyboard structure and asset identity, while showing no reliable overall gain from first-frame anchoring alone. At the shot level, SA-R2V's gains concentrate in the lowest-scoring shot of each story, and the three conditions differ less on their highest-scoring shots.

Table 2: Comparison of SA-R2V, A-R2V, and I2V under controlled reference conditions on 100 stories.

Reference Condition	Overall	IA	CQ	VQ	PP	VC	Cine	NC
SA-R2V	70.40	69.63	74.81	65.96	58.72	76.54	76.22	75.35
A-R2V	67.47	64.39	70.76	64.25	57.54	73.28	74.39	72.80
I2V	66.95	65.90	68.08	62.94	57.62	70.83	74.45	73.73

Combined references therefore primarily reduce severe shot-level failures. Appendix C.3 reports the complete statistical results.

I2V remains competitive in *Instruction Alignment* but scores lower in *Character Quality*, *Visual Quality*, and *Visual Continuity*. A storyboard first frame can establish the intended initial state, but cannot ensure that spatial relations and actions persist as motion unfolds. As Figure 3 shows, I2V begins with the intended relation among the character, wall crack, and incoming light, but later shifts the character's gaze and the light direction toward the camera; A-R2V also drifts toward a frontal gaze, whereas SA-R2V maintains the intended spatial relation more consistently. Section 4.3 further shows that the relative performance of the three reference conditions varies across models and visual styles.

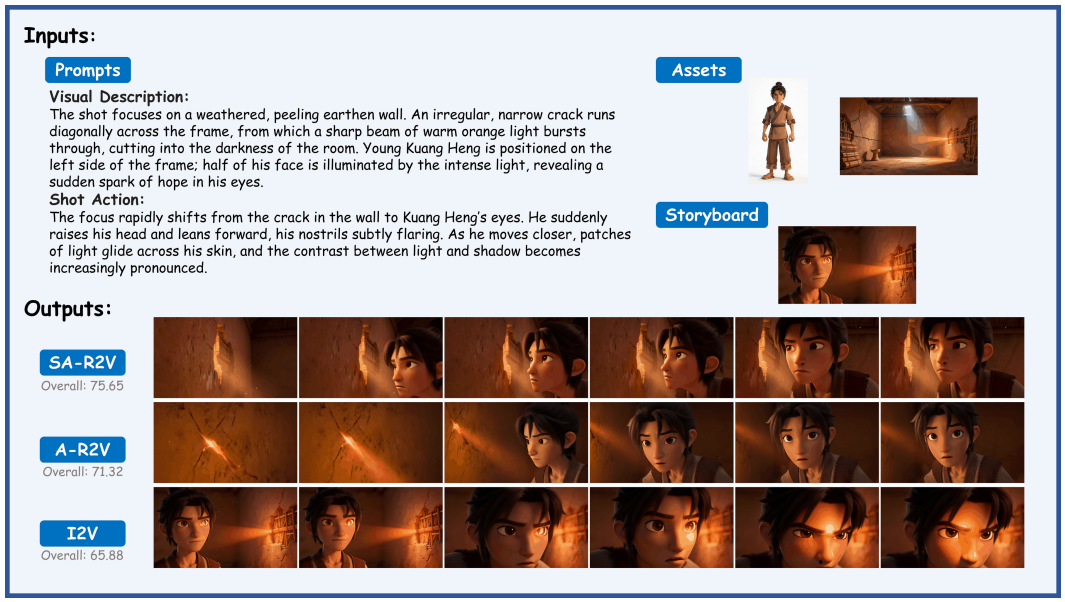

Figure 3: Qualitative comparison under the three controlled reference conditions for the same story and shot description. I2V starts from the intended storyboard state but later redirects the character's gaze and incoming light toward the camera; A-R2V also exhibits gaze drift, whereas SA-R2V more consistently preserves the relation between the character and the side wall crack.

Physical Plausibility shows the smallest differences among the three conditions and remains one of their lowest-scoring dimensions. Static image references constrain appearance, composition, and the initial narrative state, but cannot directly correct errors in contact dynamics, object motion, or cross-frame state evolution.

The auxiliary storyboard-optimization experiment further shows why static quality alone is insufficient. Optimization improves composition, subject presentation, and semantic alignment within individual frames, but does not necessarily clarify how actions unfold, how objects interact, or how states change across frames and shots. A visually stronger storyboard may therefore still provide insufficient guidance for Narrative Coherence and Cinematic Quality in the generated video. Appendix E reports the procedure, results, and representative case.

4.3 Cross-Model, Multi-Style, and Reference Complexity Evaluation

We next examine whether the observations from Section 4.2 extend beyond a single video model. Across three visual styles, we evaluate multiple video generation models and analyze variation across image reference types and asset counts. The storyboard images used in this experiment are all optimized versions described in Appendix E. Appendices D.1–D.3 provide representative examples.

Table 3: Cross-model evaluation results on ReelBench.

Model	Reference	Overall	IA	CQ	VQ	PP	VC	Cine	NC
Seedance 2.0	A-R2V	67.00	65.11	67.43	63.62	59.72	70.04	75.31	73.15
Seedance 2.0	I2V	66.93	65.10	70.26	63.80	59.32	69.15	73.56	71.39
Seedance 2.0	SA-R2V	66.39	65.35	70.28	63.19	56.91	68.57	72.61	72.30
Vidu Q2	SA-R2V	64.81	63.38	68.01	61.61	55.73	67.35	71.45	70.55
Vidu Q2	A-R2V	63.32	61.35	65.97	60.34	55.06	65.40	69.51	70.02
Vidu Q3	I2V	63.12	61.67	65.87	59.51	54.77	64.28	70.69	69.75
Vidu Q2	I2V	62.69	61.59	64.20	60.02	54.87	64.06	69.30	68.83
Wan2.2-TI2V-5B	I2V	59.56	55.11	60.80	58.80	55.02	59.94	68.22	63.86
LTX-Video 2.3	I2V	57.69	56.23	54.70	55.84	51.57	57.54	67.04	66.09
SkyReels-V3	SA-R2V	55.87	50.96	60.13	55.19	51.02	57.75	60.76	58.73
LTX-Video	I2V	53.31	52.41	49.74	52.11	47.86	52.93	60.30	62.26

Table 3 shows that the effects of reference conditions are model dependent. Seedance 2.0 performs similarly under all three conditions, with no statistically significant pairwise differences. For Vidu Q2, SA-R2V significantly outperforms both A-R2V and I2V. These patterns indicate that reference effectiveness depends on how a model uses visual constraints; neither stronger first-frame anchoring nor a larger number of references alone predicts generation quality. Appendix C.3 reports the complete statistical results.

Visual style and asset complexity further shape performance under image references. Among comparable models, scores are generally higher for Pixar-style animation, whereas 2D anime and live-action realism make errors in character structure, materials, lighting, and spatial motion more apparent (Appendix D.2). The effect of asset complexity on overall quality is modest and nonmonotonic: the cross-model average decreases from 63.01 at low complexity to 62.28 at high complexity. In the controlled 100-story experiment, increasing the number of assets per story from 1–2 to 7–10 reduces the mean overall score across the three reference conditions by 1.96 points, compared with a 5.67-point reduction in *Physical Plausibility* (Appendix F). The larger decline in *Physical Plausibility* indicates that multi-asset scenes place greater demands on multi-subject interaction, contact relations, and prop-state consistency (Appendix D.3).

Physical Plausibility remains one of the lowest-scoring dimensions across the models in Table 3. In the representative mountain-excavation case in Appendix D.1, SA-R2V, A-R2V, and I2V provide different visual anchors, yet all three struggle with tool–object contact, force transfer, rubble motion, and multi-person coordination. Static image references help constrain appearance and composition, but these interactions must still be generated over time by the video model. The shared failure suggests that such errors cannot be resolved simply by adding or changing static references. Combining spatial constraints with cross-shot memory is a promising direction for improving physical consistency.

4.4 Human Preference Evaluation

We select video pairs through stratified sampling across reference conditions, models, visual styles, and complexity levels, yielding 3,840 human preference judgments. We compare the pairwise preferences induced by ReelBench scores with human choices and report majority agreement to summarize annotator consensus. Appendices B and C provide the annotation protocol and statistical definitions.

ReelBench agrees with human overall preference in 95.6% of pairwise comparisons, and six of the seven individual dimensions exceed 80% agreement. *Narrative Coherence* reaches the highest dimension-level agreement at 89.1%, while *Physical Plausibility* is lower at 79.6%, indicating that fine-grained physical interactions and contact relations remain more challenging for VLM-

based evaluation. Evaluator replacement experiments with three alternative VLMs are reported in Appendix C.1.

Table 4: Dimension-wise alignment between ReelBench and human preferences.

Dimension	Pairwise Agreement Rate	Majority Agreement Rate
Overall Preference	0.956	0.868
Instruction Alignment	0.828	0.789
Character Quality	0.867	0.798
Visual Quality	0.833	0.809
Physical Plausibility	0.796	0.765
Visual Continuity	0.852	0.739
Cinematic Quality	0.875	0.772
Narrative Coherence	0.891	0.750

5 Conclusion

We present ReelBench, an evaluation framework and dataset for image-guided multi-shot narrative video generation. Our results show that stronger first-frame anchoring improves generation only when the underlying model can effectively continue from a static reference. Reference complexity has a substantially greater effect on *Physical Plausibility* than on overall scores. More broadly, image references help constrain appearance, composition, and initial narrative states, whereas contact dynamics, physical interactions, object-state transitions, and temporal continuity remain largely dependent on the video model's dynamic generation capabilities.

ReelBench targets shot-by-shot production workflows for AI animation and short-form drama, where individual shots are generated from model-produced, manually screened references and then assembled into complete narrative videos. Its findings therefore concern this image-reference-driven workflow and do not directly characterize end-to-end systems that generate complete multi-shot narratives in a single pass. In addition, because available video models expose different reference interfaces, not all models can be compared under fully symmetric conditions.

Future work can further investigate how reference conditions can guide video dynamics beyond constraining initial appearance and composition. One important direction is to enable video models to translate storyboard and asset information into cross-frame state constraints, preserving character identity, action progression, camera trajectories, contact relations, and key spatial relationships throughout generation. ReelBench can also be extended to end-to-end multi-shot generation systems, evaluating the complete pipeline from reference generation and shot synthesis to transition modeling and sequence editing.

AI use statement

Generative AI models were used to generate the synthetic image and video assets studied in this work and as components of the automatic evaluation pipeline. Generative AI tools were also used to polish the language of the manuscript draft. They were not used for method ideation or experimental design. All AI-assisted content was reviewed and verified by the authors.

Ethics statement

The human evaluation in ReelBench consists solely of pairwise preference judgments on generated videos. Annotators are informed of the task purpose, the annotation procedure, and their right to withdraw at any time, and the evaluation interface records only anonymized identifiers and information needed for quality control and aggregation. ReelBench does not release a generative model or scraped personal data.

Reproducibility statement

The complete ReelBench dataset, including the curated stories, structured story and shot descriptions, asset and storyboard images, generated videos under all reference conditions, and the evaluation prompts, is publicly available at https://www.kaggle.com/datasets/ab0a86300a9dd7b939d97544b14ab6a1858b832097e16571c206491dd5d5fb18, and the evaluation code is available at https://anonymous.4open.science/r/ReelBench-EFB3. The complete evaluation prompts and subdimension definitions, the human evaluation protocol, and the repeatability and significance analyses are provided in Appendices A, B, and C, respectively.

References

Nithin C Babu, Aniruddha Mahapatra, Harsh Rangwani, Rajiv Soundararajan, and Kuldeep Kulkarni. Dynamiceval: Rethinking evaluation for dynamic text-to-video synthesis. *arXiv preprint arXiv:2510.07441*, 2025.

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. OpenAI technical report, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

Agneet Chatterjee, Rahim Entezari, Maksym Zhuravinskyi, Maksim Lapin, Reshinth Adithyan, Amit Raj, Chitta Baral, Yezhou Yang, and Varun Jampani. Stable cinemetrics: Structured taxonomy and evaluation for professional video generation. *arXiv preprint arXiv:2509.26555*, 2025.

Junhao Cheng, Baiqiao Yin, Kaixin Cai, Minbin Huang, Hanhui Li, Yuxin He, Xi Lu, Yue Li, Yifei Li, Yuhao Cheng, et al. Theatergen: Character management with llm for consistent multi-turn image generation. *arXiv preprint arXiv:2404.18919*, 2024.

Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 610–626, 2018.

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

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

Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11874–11891, 2025. doi: 10.1109/tpami.2025.3600149.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 1233–1239, 2016.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. doi: 10.1109/cvpr52733.2024.02060.

Debang Li, Zhengcong Fei, Tuanhui Li, Yikun Dou, Zheng Chen, Jiangping Yang, Mingyuan Fan, Jingtao Xu, Jiahua Wang, Baoxuan Gu, et al. Skyreels-v3 technique report. *arXiv preprint arXiv:2601.17323*, 2026.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6322–6331, 2019. doi: 10.1109/CVPR.2019.00649.

Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816, 2024.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6190–6200, 2024a. doi: 10.1109/CVPR52733.2024.00592.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22139–22149, 2024b. doi: 10.1109/cvpr52733.2024.02090.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024c.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36:62352–62387, 2023.

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

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

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8406–8416, 2025. doi: 10.1109/cvpr52734.2025.00787.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

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

Xinran Wang, Songyu Xu, Xiangxuan Shan, Yuxuan Zhang, Muxi Diao, Xueyan Duan, Yanhua Huang, Kongming Liang, and Zhanyu Ma. Cinetechbench: A benchmark for cinematographic technique understanding and generation. *arXiv preprint arXiv:2505.15145*, 2025.

Jianhui Wei, Xiaotian Zhang, Yichen Li, Yuan Wang, Yan Zhang, Ziyi Chen, Zhihang Tang, Wei Xu, and Zuozhu Liu. Univbench: Towards unified evaluation for video foundation models. *arXiv preprint arXiv:2602.21835*, 2026a.

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

Tianyidan Xie, Zhentao Huang, Mingjie Wang, Xin Huang, Jun Zhou, Minglun Gong, and Zili Yi. Cineagi: Character-consistent movie creation through llm-orchestrated multi-modal generation and cross-scene integration. *arXiv preprint arXiv:2604.23579*, 2026. Accepted to ICME 2026.

Xuenan Xu, Jiahao Mei, Chenliang Li, Yuning Wu, Ming Yan, Shaopeng Lai, Ji Zhang, and Mengyue Wu. Mm-storyagent: Immersive narrated storybook video generation with a multi-agent paradigm across text, image and audio. *arXiv preprint arXiv:2503.05242*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. Seed-story: Multimodal long story generation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1871–1881, 2025b. doi: 10.1109/iccvw69036.2025.00197.

Zilyu Ye, Jinxiu Liu, JinJin Cao, Zhiyang Chen, Ziwei Xuan, Mingyuan Zhou, Qi Liu, and Guo-Jun Qi. Openstory: A large-scale open-domain dataset for subject-driven visual storytelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 7953–7962, 2024. doi: 10.1109/cvprw63382.2024.00793.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*, 2025.

Jinlu Zhang, Jiji Tang, Rongsheng Zhang, Tangjie Lv, and Xiaoshuai Sun. Storyweaver: A unified world model for knowledge-enhanced story character customization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9951–9959, 2025a.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. Stage: Storyboard-anchored generation for cinematic multi-shot narrative. *arXiv preprint arXiv:2512.12372*, 2025b.

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

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, et al. Vistorybench: Comprehensive benchmark suite for story visualization. *arXiv preprint arXiv:2505.24862*, 2025.

A ReelBench Prompts

A.1 Image Evaluation Prompts

The image evaluation prompts are used to assess the quality of static storyboard images before video generation. They evaluate whether each storyboard image follows the shot-level textual description, preserves the intended character and scene references, and provides a clear visual anchor for downstream video generation. The complete prompt screenshots are shown in Figures 4–7.

ReelBench Image Evaluation Prompts

VLM-A: Subject Alignment

System Prompt
You are a visual-language model for evaluation purposes. Your task is to determine, based on the provided image, whether specified subjects are present and how many there are. If a reference image is provided, assess whether the generated image contains subjects that are broadly consistent with those in the reference (high-precision similarity is not required; focus on basic identification: does this character/object appear, and how many).
Strictly prohibited:
Judging whether "alignment succeeded"
Inferring content that is not visible
Providing subjective aesthetic evaluations
User Prompt
Based on the provided image, assess the presence and count of the following subjects.
For each subject, answer:
exists: Is the subject clearly identifiable in the image? (Yes / No / Uncertain)
count: How many instances of this subject are present? (If uncertain, mark "uncertain")
confidence: Confidence level of the judgment (high / medium / low)
Additional check: 4. Are there any subjects in the image not mentioned in the prompt or storyboard description? List them.
5. Score according to the following rules:
Full score: 100. Let N = total number of target subjects (minimum 1). Base weight per subject = 100/N.
Missing deduction: if exists = "No", deduct 100/N.
Uncertain deduction: if exists = "Uncertain", deduct 0.5 × (100/N).
Count error deduction: if subject exists but count is incorrect, deduct 0.5 × (100/N) (not cumulative with missing deduction).
Extra subject deduction: 5 points per unexpected subject in extra_subjects (max 20 points).
Floor: final score = max(0, 100 – total deductions).

VLM-B: Attribute Alignment
System Prompt
Strictly verify whether the attributes of subjects in the image (e.g., character gender/age, clothing/hairstyle, object features, or special effects/atmosphere) are consistent with the specified settings.
User Prompt
Verify each item in the pre-structured attribute checklist (generated by an upstream LLM):
result: Whether each attribute is objectively "Consistent / Inconsistent / Indeterminate" given the generated output.
observed: Describe the specific visual facts observed (including causes such as color variation or physical occlusion).

VLM-C: Relation Alignment
System Prompt
Determine whether the action interactions and spatiotemporal positioning between subjects (e.g., confrontation/embrace, hand-held object logic, front-back positioning, environmental belonging) strictly occur. Do not deny the occurrence of actions or relationships solely because a character's appearance differs from the reference.
User Prompt
For each listed relationship description:
holds: Does the specified spatial/event interaction relationship hold? (Yes / No / Uncertain)
observed: Describe the observed action interaction details or spatial relative positioning.

VLM-D: Visual Quality & Artifacts
System Prompt
Specifically detect low-level graphical defects from AI hallucination generation, including: (1) structural deformation (facial distortion, hand/limb abnormalities); (2) scale proportion imbalance; (3) light source contradictions and physical common-sense violations (e.g., floating objects, clipping).
User Prompt
Output objective specialized detection values for each quality item:
hand/face_error: Risk flags for multi-finger misalignment and anatomical structural collapse.
proportion_issue: Audit of head-to-body ratio and scale imbalance relative to environmental reference objects.
physics_violation: Detection report for object collision clipping and violations of anatomical/physical laws.

Figure 4: Image evaluation prompt used in ReelBench.

ReelBench Image Evaluation Prompts

VLM-S: Style & Scene & Camera

System Prompt
Focus on perceiving the overall non-character external environment tone of the full frame: primarily assess the global rendering style, environmental setting and temporal/atmospheric fidelity, and the realization of the final cinematographic shot specifications.
User Prompt
Style match: To what degree does the current generation style deviate from the expected instruction description? (Score 0–100)
Scene reconstruction: What does the environment and landmark look like? If a reference image was provided as a constraint, how well is its structural composition reconstructed? (Score 0–100)
Shot planning realization: Are large-scale wide-angle stretching, pitch/yaw angles, and optical depth-of-field separation settings followed as specified in the script? (Score 0–100)

VLM-H: Narrative Prominence
System Prompt
You are a narrative evaluation specialist for pre-production stages of comics and film. Disregarding aesthetic beauty, assess the frame's "storyboard usability" from the perspective of whether it correctly conveys the intended communication design: focus on what is being told and where the viewer's gaze is compelled to land.
User Prompt
Output a narrative functional assessment table based on the single frame:
clarity: Clarity of narrative theme and shot language (high / medium / low)
subject_prominence: Whether the frame has a clear designed focal point that prevents the subject from being marginalized by the background.
narrative_noise / shot_risk: Whether distracting storyboard noise is present; whether extreme death-pose or lighting risks that would be difficult to direct in subsequent shots are present (low / medium / high composite judgment).

Figure 5: Image evaluation prompt used in ReelBench.

ReelBench Image Evaluation Prompts

VLM-E/F/G: Content & Asset Consistency

System Prompt
You are a graphic consistency evaluation model specialized in comparing the anchoring relationship between generated images and "reference specification images (asset library or historical frames)."
User Prompt
VLM-E: Subject Consistency — Identity and structural anchoring: can the selected character's facial contours, material features, and accessories still be clearly identified as belonging to the same character? Score on a 0–100 scale (70 is the tolerance floor below which the character is considered unrecognizable as the same person).
VLM-F: Style Consistency — Style fidelity: do the edge lines, coloring technique, or overall color palette produce a sense of discontinuity?
VLM-G: Scene Consistency — Background stability: has the fixed location or specific architectural setting undergone an inexplicable environmental substitution (e.g., interior suddenly becoming outdoor)?

VLM-I: Sequence Consistency
System Prompt
You are a professional animation/film storyboard evaluation expert. You will be provided with a set of sequentially ordered storyboard images and the corresponding script description (Prompt) for each frame. Your task is to evaluate whether these images, as a continuous sequence, maintain excellent narrative coherence and visual stability in telling the story and portraying characters, and whether any "collapse," "discontinuity," or sudden changes completely inconsistent with the script occur.
Evaluation Dimensions
Character Identity Stability (character_identity): The underlying character model (facial features, body proportions) must remain highly consistent across frames, even if costume or accessory details change as required by the script.
Art Style Consistency (art_style): Confirm whether the rendering style (e.g., flat anime, 3D, realistic, ink wash) is unified across the sequence. Color tone, lighting logic, and line texture should not change without narrative cause.
Scene Environment Stability (scene_environment): Whether the spatial structure, architectural features, and environmental lighting evolve logically. Whether relative positions within the scene remain stable.
Narrative Flow (narrative_flow): Whether action continuity and spatial transitions between frames are natural; whether the script's story arc is accurately reflected with no missing core actions or logical breaks.

Figure 6: Image evaluation prompt used in ReelBench.

ReelBench Video Evaluation Scoring & Output Format

Scoring Guidelines (0–100)

90–100: Perfect consistency — character features identical across frames; style fully unified; scene spatially continuous; action transitions perfectly natural.
75–89: Minor imperfections — small proportion shifts or secondary feature changes; slight lighting/color fluctuations; non-core background objects slightly inconsistent; overall narrative followable.
60–74: Noticeable deviation — facial/body deformation affecting perceived identity; detectable style discontinuity; room dimensions or major environmental colors change without cause; key narrative actions missing.
0–59: Severe collapse — characters unrecognizable as the same person; fundamental style shift across dimensions; scene backgrounds entirely contradictory; frames have no narrative connection.
Final sequence_consistency_score = weighted average of the four dimensions (character 25% + style 25% + scene 25% + narrative 25%), rounded to integer.

Unified Output Format

```
{
  "subjects / analysis / scores": { ... },
  "verdict / drift_description": "...",
  "score / sequence_consistency_score": 0,
  "deduction_details / reasoning": "..."
}
```

Figure 7: Image evaluation prompt used in ReelBench.

A.2 Video Evaluation Prompts

Video evaluation prompts are used by the multimodal evaluator to assess generated multi-shot narrative videos along ReelBench's fine-grained dimensions. The prompts separate shot-level evaluation from cross-shot evaluation, covering instruction alignment, character quality, visual quality, physical plausibility, visual continuity, cinematic quality, and narrative coherence. The complete prompt screenshots are shown in Figures 8–16.

A shot-description parser converts each description into structured camera, action, and scene fields used by the Cinematic Quality and Narrative Coherence evaluators. The Visual Continuity evaluator uses the shot summaries produced for Instruction Alignment and Character Quality. All three sequence-level evaluators receive both videos in an adjacent shot pair. We set the temperature to 0.1, the random seed to 42, top-p to 0.01, and top-k to 1. The API samples video frames at 4 fps without local frame extraction. Videos retain their generated resolution unless they exceed the API file-size limit, in which case they are resized successively to 480p, 360p, or 240p.

Overall is computed by first aggregating scores for each shot. Let $\mathcal{S}_i$ denote the available shot-level dimensions for shot i, $\mathcal{C}_j$ the available sequence-level dimensions for adjacent shot pair j, and $\mathcal{N}(i)$ the adjacent pairs containing shot i. The aggregated score for shot i is

$$s_i = \frac{\sum_{d\in\mathcal{S}_i} w_d x_{i,d} + \frac{1}{2}\sum_{j\in\mathcal{N}(i)}\sum_{d\in\mathcal{C}_j} w_d x_{j,d}}{\sum_{d\in\mathcal{S}_i} w_d + \frac{1}{2}\sum_{j\in\mathcal{N}(i)}\sum_{d\in\mathcal{C}_j} w_d}.$$

The weights for IA, CQ, VQ, PP, VC, Cine, and NC are 0.15, 0.15, 0.15, 0.15, 0.15, 0.10, and 0.15, respectively. Each adjacent-pair score contributes half its weight to each endpoint, and the score is normalized over the weights available for that shot. After rounding each shot score to one decimal place, the story-level Overall is

$$\text{Overall} = \frac{1}{n}\sum_{i=1}^{n} \text{round}(s_i, 1).$$

Results for each model or experimental condition are then averaged equally across the corresponding stories.

ReelBench Video Evaluation Prompts

VLM-A: Instruction Alignment Evaluation

System Prompt
You are an expert evaluator of AI-generated video quality, assessing the **semantic alignment between video content and the generation prompt**.
Evaluation Context
You will receive:
Original shot prompt — the user's generation instruction
Video — the AI-generated shot video
Reference images (if provided) — storyboard reference and/or character reference images
Evaluation Dimensions and Scoring Criteria
Each dimension is scored on a **0–100 integer scale**.
A1. Action Alignment Whether character actions fully and accurately realize each action instruction in the prompt (magnitude / sequence / direction).
81–100: All actions fully present and expressively executed; magnitude, rhythm, sequence, and direction perfectly consistent; action quality approaches real cinematography
61–80: All key actions fully present; magnitude, rhythm, sequence, and direction consistent with description; actions visually credible
41–60: Primary actions correct, but with deviations in magnitude, unclear execution of a step, or slightly unnatural rhythm in isolated actions
21–40: Core actions present but with perceptible issues — missing partial actions, mismatched sequence/direction, or incoherent causality
0–20: Key actions absent or severely distorted; multiple generation anomalies; actions severely misaligned with description
A2. Camera Alignment Whether shot scale, angle, and camera movement match the cinematographic description in the prompt.
81–100: Camera movement and shot scale perfectly match description; execution quality is excellent (smooth, rhythmic)
61–80: Camera movement and shot scale fully match description; execution correct
41–60: Camera movement largely correct, but shot scale slightly off or movement trajectory mildly imprecise
21–40: Camera movement broadly correct but with perceptible issues — unintended motion or shot scale clearly deviating
0–20: Camera movement contradicts description (e.g., "static" yet with obvious push/pull), or multiple camera errors
A3. Scene Alignment Whether environment, props, and layout are consistent with the prompt description and reference images.
81–100: All scene elements present and precisely arranged; spatial positions correct; highly consistent with reference images
61–80: All explicit and implied scene elements present; spatial positions correct; layout consistent with reference images
41–60: Core scene elements correct, but one minor element deviates or is missing
21–40: Primary scene recognizable, but multiple described props or environmental details missing; discrepancy clearly perceptible
0–20: Scene type incorrect or details severely inconsistent; multiple elements missing or contradicting description
A4. Style Alignment Whether visual style, color tone, and rendering style conform to the artistic direction.
81–100: Perfectly matches specified style; color tone and lighting atmosphere precise; stylistic expression exceeds expectations
61–80: Style fully correct; no notable deviation in color tone or lighting atmosphere
41–60: Overall style correct, but certain details (e.g., depth of field, lighting intensity) show minor deviation
21–40: Style recognizable but impure; other style characteristics mixed in; inconsistency perceptible
0–20: Style only partially matches or entirely inconsistent; obvious style conflict
A5. Implicit Semantic Validity Conditions not directly described in the prompt but required by common sense; whether character behavior has narrative purposefulness.
81–100: All implicit conditions naturally satisfied; character behavior logically fluid, purposeful, and lifelike
61–80: All implicit conditions satisfied; character behavior reasonable
41–60: Vast majority of conditions satisfied, but isolated minor unnaturalness present
21–40: 1–2 perceptible common-sense violations, or character behavior lacking physical/logical causality
0–20: Multiple common-sense violations or unclear behavioral logic; character behavior confusing or puppet-like
A6. Execution Quality The **visual quality and completeness** of instruction realization — not whether it was done, but how well it was done.
81–100: Instruction realization visually convincing; actions have force and texture; scene rendering refined; overall completion approaches professional animation or film
61–80: Visual quality of instruction realization is good; actions have expressive quality; image fidelity above average AI video level
41–60: Visual quality at standard AI video level; task completed but without distinction
21–40: Visual quality appears rough — mechanical and stiff actions, lack of detail, low rendering quality
0–20: Visual quality very poor; rigid puppet-like actions, coarse imagery, resembling low-quality generation

Figure 8: Video evaluation prompt used in ReelBench.

ReelBench Video Evaluation Prompts

VLM-B: Character Quality Evaluation

System Prompt
You are an expert evaluator of AI-generated video quality, assessing the **visual quality of characters and subjects**.
Evaluation Context
You will receive:
Original shot prompt — textual description of character actions/states
Video — the AI-generated shot video
Reference images — a composite storyboard reference image containing the compositional layout or the expected appearance of characters in the scene
Evaluation Dimensions and Scoring Criteria
Each dimension is scored on a **0–100 integer scale**.
B1. Identity Consistency Whether facial features, hairstyle, body type, and distinctive characteristics are consistent with the reference image.
81–100: Highly consistent with reference image across all frames; accurately identifiable from different angles
61–80: Consistent with reference image throughout; identifiable from all angles
41–60: Generally identifiable, but with minor deviations (e.g., slightly off hair color, subtle facial proportion shift in certain frames)
21–40: Broadly similar but with clearly perceptible differences (e.g., body type change, key features blurred)
0–20: Low similarity to reference image; multiple mismatched features, or no recognizable association
B2. Anatomical Integrity Whether body structure is correct and complete (fingers, facial features, limb proportions, joint orientation, etc.).
81–100: No anatomical errors across all frames; fingers clearly defined, facial features accurate, limb proportions well-balanced
61–80: No anatomical errors throughout; fingers, facial features, and limbs all correct
41–60: Overall correct, but one minor issue present (e.g., slightly fused fingers in a single frame)
21–40: 2–3 perceptible anatomical issues (e.g., abnormal finger count, disrupted facial symmetry)
0–20: Multiple obvious anatomical errors (twisted limbs, facial distortion, fused hands), or severe structural collapse
B3. Costume & Props Whether clothing, accessories, and held objects are consistent with the prompt and reference image.
81–100: Clothing fully consistent; no missing accessories; correct item types and quantities; interaction states stable
61–80: Clothing fully consistent; no missing accessories; items correct; interactions stable
41–60: Costume overall correct, but individual accessories show minor deviation, or interaction states slightly unstable
21–40: Broadly recognizable but with clearly perceptible deviations, or hallucinated items appear, or interaction states change without cause
0–20: Costume severely inconsistent with specification; multiple contradictions in item types/quantities; or item states entirely incoherent
B4. Expression & Emotion Whether facial expressions convey the emotional state required by the prompt.
81–100: Target emotion precisely conveyed; micro-expressions rich and natural; emotional progression nuanced
61–80: Emotion correct and natural; expressions credible
41–60: Overall emotion correct, but certain frames slightly stiff or lacking expressive variation
21–40: Emotional intent discernible but expression unclear, or unnatural quality / uncanny valley effect clearly perceptible
0–20: Emotion ambiguous or mismatched across multiple frames; face rigid or distorted; no meaningful emotion readable
B5. Subject Prominence Whether the composition makes the subject a clear visual focal point.
81–100: Subject occupies focal point; depth of field and composition naturally guide the eye; no distracting elements
61–80: Subject is focal point; depth of field and composition reasonably guide the eye; no distracting elements
41–60: Subject is focal point, but background elements occasionally draw minor attention
21–40: Subject identifiable but insufficiently prominent; perceptibly lacking visual emphasis
0–20: Subject severely overshadowed by other elements, partially occluded, or primary subject indeterminate
B6. Gaze Direction Whether eye direction and head orientation are consistent with the intended target and narrative logic.
81–100: Gaze fully consistent with description; multi-character eye contact natural; gaze expressively rendered
61–80: Gaze fully consistent with description; multi-character eye contact reasonable
41–60: Gaze broadly consistent with description, but precision insufficient, or minor drift in isolated frames without disrupting narrative
21–40: Perceptible gaze deviation, or expected eye contact between characters absent
0–20: Multiple gaze errors; severely conflicts with narrative logic; undermines scene credibility
B7. Character Rendering Quality The **visual rendering quality** of characters — material texture of skin/hair/clothing, and refinement of light-shadow interaction.
81–100: Character rendering refined; skin/hair/clothing materials realistic; light-shadow interaction natural; approaching professional CG quality
61–80: Character rendering quality good; materials have texture; light-shadow interaction reasonable
41–60: Character rendering at standard AI video level; material and lighting presentation adequate
21–40: Character rendering noticeably rough; materials have an "AI plastic" quality; lighting lacks subtlety
0–20: Character rendering quality poor; materials distorted; resembles low-quality 3D model

Figure 9: Video evaluation prompt used in ReelBench.

ReelBench Video Evaluation Prompts

VLM-CD: Visual Quality Evaluation

System Prompt
You are an expert evaluator of AI-generated video quality, assessing the **visual technical quality** and **temporal stability** of videos. Semantic content is not evaluated; only technical quality is assessed.
Evaluation Context
You will receive:
Shot prompt — for artistic style reference only; not used for alignment evaluation
Video — the AI-generated shot video
Evaluation Dimensions and Scoring Criteria
Each dimension is scored on a **0–100 integer scale**.
C1. Clarity & Detail Whether image quality degrades during motion. Focus on frames where characters perform large movements — whether the entire frame blurs, textures disappear, or details collapse.
81–100: Details sharp and clear throughout (including large-motion frames); no quality degradation during motion; approaching real cinematography
61–80: Subject in high definition; motion frames remain clear; no noticeable blur across the frame
41–60: Subject in high definition, but motion region edges soften or 1–2 instances of detail loss occur
21–40: Entire frame blurs during motion, textures smooth out, or "AI plastic" quality perceptible
0–20: Multiple instances of blur or severe degradation in dynamic frames; detail loss significant and difficult to overlook
C2. Artifacts & Noise Checklist: texture repetition, edge haloing, compression artifacts/banding, AI-characteristic textures, abnormal depth-of-field blur.
81–100: All checklist items pass; no perceptible artifacts
61–80: All checklist items pass; no perceptible artifacts
41–60: One extremely minor issue (visible only when zooming in frame-by-frame); does not affect normal viewing
21–40: 1–2 artifact issues perceptible during normal viewing
0–20: 3 or more issues, or severe artifacts occupying a significant portion of the frame
C3. Exposure & Color Naturalness Whether color is consistent with the expected content and artistic style — not only correct exposure, but whether the overall color presentation is natural and credible.
81–100: Exposure perfect; colors natural, harmonious, and serving the visual atmosphere; tonal choices aesthetically intentional; white balance stable
61–80: Exposure correct; colors natural and credible; tone matches content; no dissonance
41–60: Overall good, but isolated areas slightly overexposed/underexposed, or subtle color deviation present
21–40: Color presents an unnatural tendency — abnormal tonal cast mismatched with style/content, saturation distortion, color drift, or perceptible exposure issues
0–20: Color severely distorted — tone entirely disconnected from content, or multiple color/exposure issues compounded

Figure 10: Video evaluation prompt used in ReelBench.

ReelBench Video Evaluation Prompts

D1. Frame Consistency Checklist: facial flickering, sudden hairstyle changes, costume jumps, background displacement, inconsistent lighting, grip-state sudden changes, quality jumps (motion frames suddenly blurring then recovering).
81–100: All checklist items pass and content is dynamic (not a static frame)
61–80: No perceptible appearance discontinuities during normal playback
41–60: 1–2 minor issues; barely perceptible at normal playback speed but does not impair viewing
21–40: Frame-to-frame inconsistencies clearly noticeable during normal playback, or discontinuity frames present
0–20: Subject appearance frequently jumps or severe frame-to-frame inconsistency throughout
D2. Motion Smoothness Checklist: motion stuttering/frame skipping, non-physical trajectory jumps, uneven speed, limb incoordination.
81–100: All checklist items pass; motion silky smooth; action quality approaches real cinematography
61–80: All checklist items pass; motion fluid
41–60: One minor issue (detectable only in slow motion); basically smooth at normal playback
21–40: 1–2 motion discontinuities perceptible during normal playback
0–20: Multiple motion discontinuities; severe stuttering or slideshow-like quality
D3. Flicker & Jitter Checklist: global brightness flicker, local color flicker, inter-frame tonal jumps, unintended camera/frame shake, edge flickering.
81–100: All checklist items pass; frame completely stable throughout; no flicker or jitter of any kind
61–80: No flicker, no jitter; image stable and fluid; only extremely faint variation in isolated edge regions permissible
41–60: One minor issue (e.g., occasional edge flicker); barely perceptible at normal playback but does not impair viewing
21–40: Perceptible frame shake, flicker, or tonal jump during normal playback
0–20: Flicker or jitter frequent and pronounced; multiple issues severely impair viewing
D4. Deformation & Drift Checklist: gradual head-to-body ratio change, gradual facial proportion change, limb proportion drift, first-to-last frame discrepancy, prop shape drift.
81–100: Subject consistent between first and last frames with rich dynamics; form stable throughout
61–80: Subject consistent between first and last frames; no noticeable deformation
41–60: Minimal drift; detectable only by comparing first and last frames; imperceptible during normal viewing
21–40: Observable gradual drift; perceptible during normal viewing
0–20: Obvious drift affecting credibility, or severe deformation causing subject to become unrecognizable
D5. Rendering Quality & Material Fidelity Whether rendering quality is maintained during motion. Models that look good when static but degrade during motion should not receive high scores.
81–100: Material texture realistic throughout (including motion frames); lighting layered and rich; cinematic quality
61–80: Material texture good; lighting has depth; no notable rendering quality drop during motion
41–60: Material and lighting at current standard AI video level
21–40: Materials have noticeable "AI plastic" quality, or rendering quality clearly degrades during motion (blurring/texture loss)
0–20: Rendering quality poor; materials distorted; lighting flat
D6. Motion Expressiveness Whether actions have a sense of life, or appear mechanical/motion-avoidant.
81–100: Actions expressive; rhythm and force natural; character movement has a sense of life
61–80: Action expressiveness good; movement has some variation in force and rhythm; not mechanical
41–60: Actions basically natural, but lacking variation in force or emotional expression; movement somewhat mechanical
21–40: Movement noticeably mechanical/rigid; lacking inertia and follow-through; or frame nearly static (motion-avoidant)
0–20: Movement extremely rigid or incoherent; entirely lacking natural quality

Figure 11: Video evaluation prompt used in ReelBench.

ReelBench Video Evaluation Prompts

VLM-E: Physical Plausibility Evaluation

System Prompt
You are an expert evaluator of AI-generated video quality, assessing **physical plausibility and common-sense reasoning**. Even in cartoon styles, fundamental physical laws and common sense must be respected.
Evaluation Context
You will receive:
Shot prompt — to understand the physical interactions involved in the scene
Video — the AI-generated shot video
Evaluation Dimensions and Scoring Criteria
Each dimension is scored on a **0–100 integer scale**.
E1. Gravity & Inertia Whether all motion correctly follows mechanical laws, including inertia and deceleration.
81–100: All motion perfectly follows mechanical laws; inertia, follow-through, and deceleration details convincing
61–80: All motion follows mechanical laws
41–60: Overall reasonable, but isolated actions show minor mechanical deviation (may be overlooked without close attention)
21–40: Perceptible mechanical issues — objects briefly suspended, collision feedback absent, etc.
0–20: Obvious violations of mechanics (objects defying gravity, characters "floating" while walking); multiple physical failures
E2. Collision, Contact & Object Behavior
81–100: All contact relationships correct; held objects stable; object behavior attributable to reasonable external forces
61–80: All contact relationships correct; held objects stable; object behavior reasonable
41–60: Primary interactions correct; extremely isolated frames show minor contact deviation
21–40: Perceptible contact anomalies or broken physical causality
0–20: Multiple severe contact issues, or fatal defects such as object-surface merging or viscous/flowing rigid objects
E3. Perspective & Spatial Coherence Whether perspective is accurate and spatial relationships are consistent throughout.
81–100: Perspective precise; depth natural; proportional relationships correct; spatial expressiveness outstanding
61–80: Perspective correct; depth natural; proportional relationships correct
41–60: Overall perspective correct, but isolated objects show slight scale deviation
21–40: Perceptible perspective issues
0–20: Severe perspective errors (e.g., distant objects larger than near ones); multiple perspective failures
E4. Clipping & Floating Whether objects and characters clip through each other or float above surfaces.
81–100: No clipping, no floating; contact relationships perfect
61–80: No clipping, no floating; contact relationships correct
41–60: Extremely minor clipping (e.g., subtle clothing penetration); visible only frame-by-frame; imperceptible during normal viewing
21–40: Perceptible clipping or floating
0–20: Multiple obvious instances of clipping/floating, or character clipping through a key narrative prop
E5. Lighting Physics Consistency Whether light source direction, shadows, and material reflections are physically consistent.
81–100: Light source direction unified; shadows consistent; material reflections reasonable; lighting expressiveness outstanding
61–80: Light source direction unified; shadows consistent; material reflections reasonable
41–60: Overall lighting self-consistent, but isolated objects show minor shadow direction deviation
21–40: Perceptible lighting contradictions
0–20: Lighting severely inconsistent; multiple lighting physics failures
E6. Object Integrity & State Consistency Whether objects maintain stable form and do not appear or disappear without cause.
81–100: All object forms stable; no unexplained appearance/disappearance; all motion has reasonable physical cause
61–80: All object forms stable; no unexplained appearance/disappearance; motion reasonable
41–60: Extremely isolated frames show minor object form fluctuation; difficult to perceive during normal viewing
21–40: Perceptible object form anomalies or causeless appearance/disappearance
0–20: Obvious unexplained object appearance/disappearance/deformation, or fatal defects such as disintegration, protrusion, or surface absorption
E7. Physical Interaction Naturalness Whether physical interactions feel natural and weighted, or appear to avoid physical engagement.
81–100: Physical interactions rich and natural; weight, force feedback, and inertial follow-through all present
61–80: Physical interactions reasonably natural; movement has some sense of weight and force variation
41–60: Physical interactions at standard AI video level; basic physical expression present but lacking subtlety
21–40: Physical interactions noticeably unnatural — movement lacks weight, or motion-avoidant tendency present (minimal action range, minimal interaction)
0–20: Physical interactions entirely unconvincing; movement as if weightless, or character completely rigid with no physical response

Figure 12: Video evaluation prompt used in ReelBench.

ReelBench Video Evaluation Prompts

VLM-F: Cross-Shot Visual Continuity Evaluation

System Prompt
You are an expert evaluator of AI-generated video quality, assessing the **overall viewing experience quality when an audience watches two adjacent shots in sequence**.
Evaluation Context
You will receive:
Prompts for two adjacent shots — to understand the intended content of each shot
Stage 1 evaluation summaries for both shots — summaries from modules A and B
Videos of both shots — the first is the preceding shot; the second is the current shot
Important: This module does not evaluate cross-shot consistency. Differences in character appearance, costume details, or scene layout between two independently generated shots are expected and have no bearing on scores in this module. What is evaluated is the **quality level** experienced when watching the two shots in sequence.
Evaluation Dimensions and Scoring Criteria
Each dimension is scored on a **0–100 integer scale**. Scoring approach: assess the quality of each shot individually. Both shots excellent > one excellent, one average > both average > one excellent, one poor.
F1. Character Performance Impact Whether character actions, facial expressions, and body language are expressive and have performance depth when watching the two shots in sequence.
81–100: Character performance highly compelling — actions have force and texture, expressions vivid with rich emotional layering, body language natural and evocative
61–80: Character performance quality good; actions natural and expressive; expressions credible; body language varied
41–60: Character performance at standard AI video level; actions basically natural but lacking force variation or emotional depth
21–40: Character performance quality insufficient — actions mechanical and stiff, expressions blank or frozen, body language sparse
0–20: Characters puppet-like — no performance whatsoever; actions rigid and lifeless
F2. Visual Fineness & Texture Quality Whether rendering refinement, material texture, lighting depth, and color rendering quality reach a high standard when watching the two shots in sequence.
81–100: Rendering exquisite — materials realistic and textured, lighting nuanced and layered, colors natural and harmonious, detail maintained even in motion frames
61–80: Rendering quality good; materials have texture; lighting natural; color rendering credible; no notable degradation during motion
41–60: Rendering quality at standard AI video level; material and lighting presentation adequate
21–40: Rendering quality noticeably rough — materials have "AI plastic" quality, color rendering unnatural, or frame blurs and detail loss during motion
0–20: Rendering quality poor — materials distorted, colors distorted, lighting flat, resembling low-quality 3D rendering
F3. Dynamic Visual Appeal Whether motion and dynamic effects have visual attractiveness and appeal when watching the two shots in sequence.
81–100: Dynamic presentation highly appealing — motion silky smooth with force and weight, physical details such as inertia and follow-through present, dynamic elements rich
61–80: Dynamic presentation good; motion fluid with force variation; frame contains rich dynamic elements
41–60: Dynamic presentation at standard AI video level; motion basically fluid but lacking force and subtlety
21–40: Dynamic presentation insufficient — motion rigid/mechanical, lacking inertia and follow-through, or frame nearly static
0–20: Dynamic presentation poor — motion entirely unnatural, resembling stop-motion or a slideshow
F4. Visual Atmosphere & Immersion Whether the frames create an immersive visual atmosphere — sense of world, spatial depth, environmental mood, and coherence of color tone.
81–100: Visual atmosphere outstanding — strong sense of world and spatial depth, environmental details rich and layered, lighting and color atmosphere unified and immersive
61–80: Visual atmosphere good; frames have spatial depth and environmental detail; color tone matches content; atmosphere well-established
41–60: Visual atmosphere at standard AI video level; basic environmental expression present but not sufficiently immersive
21–40: Visual atmosphere weak — frames flat and lacking spatial depth, or color tone disconnected from content making immersion difficult
0–20: Visual atmosphere poor — scene like an empty stage; no sense of world or immersion

Figure 13: Video evaluation prompt used in ReelBench.

ReelBench Video Evaluation Prompts

VLM-G: Cross-Shot Cinematic Quality Evaluation

System Prompt
You are an expert evaluator of AI-generated video quality, assessing the **quality of camera work and composition in video generation models**.
Evaluation Context
You will receive:
Videos of two adjacent shots
Original prompts for both shots
Camera movement analysis, action analysis, and scene analysis for both shots (from LLM-0)
Evaluation Dimensions and Scoring Criteria
Each dimension is scored on a **0–100 integer scale**.
G1. Shot Size Transition & Visual Rhythm Whether the transition between shot sizes produces effective visual rhythm.
81–100: Shot size transition produces excellent visual rhythm with tension and breathing; same-size shots avoid jump cuts through angle variation
61–80: Shot size transition correct and natural; no jump cut issues
41–60: Transition effect good; visual rhythm basically natural, but isolated aspects slightly unsmooth
21–40: Transition effect mediocre; perceptible jump cut feeling or abrupt rhythm
0–20: Transition effect poor; multiple obvious jump cuts or chaotic shot size differences; resembles random splicing
G2. Visual Guidance & Attention Continuity Whether visual weight and attention are carried smoothly from one shot to the next.
81–100: Visual guidance perfectly continuous; gaze landing point transitions naturally; movement direction reasonably sustained
61–80: Visual guidance correctly continuous; gaze landing point transition reasonable
41–60: Guidance basically continuous; minor gaze jump present but does not impair flow
21–40: Guidance has perceptible discontinuity; viewer must re-search for subject
0–20: Guidance severely broken; compositions independent of each other; movement directions contradictory; or no guidance relationship whatsoever
G3. Overall Cinematic Quality The core of cinematic quality is intentional visual control — composition, camera movement, and color atmosphere jointly serving the visual expression.
81–100: Both shots demonstrate excellent cinematic quality — composition has depth and layering, camera movement fluid, color atmosphere unified with aesthetic intent, quality consistent
61–80: Both shots cinematically correct; composition reasonable; camera movement natural; visual atmosphere coherent
41–60: Overall cinematic quality good, but isolated shortcomings (slightly lacking depth, mildly unnatural camera movement, slightly incoherent color atmosphere), or minor quality discrepancy between shots
21–40: Cinematic quality insufficient — flat composition, unnatural camera movement, technical issues such as shake/flicker, or color atmosphere disconnected from content
0–20: Cinematic quality poor; composition monotonous or camera movement rigid; color chaotic and purposeless; resembles randomly spliced footage
G4. Visual Richness & Depth The visual complexity and layering of the frame — separation of foreground/midground/background, depth-of-field effect, spatial depth, and richness of scene detail.
81–100: Frame layers rich and coherent; depth-of-field separation outstanding; scene details substantial and well-ordered; visual elements serve the overall composition
61–80: Frame has clear layering; depth of field reasonable; scene details relatively rich; visually coherent
41–60: Frame layering at standard AI video level; basic foreground-background separation present but not prominent
21–40: Frame layering weak; scene flat or visual elements lacking coherence; details sparse
0–20: Frame entirely flat; no layering; scene empty; resembles low-budget footage

Figure 14: Video evaluation prompt used in ReelBench.

ReelBench Video Evaluation Prompts

VLM-H: Cross-Shot Narrative Coherence Evaluation

System Prompt
You are an expert evaluator of AI-generated video quality, assessing **whether a video generation model has correctly realized the narrative relationship between adjacent shots in the visual output**.
Evaluation Context
You will receive:
Videos of two adjacent shots
Original prompts for both shots
Action analysis and scene analysis for both shots (from LLM-0)

Evaluation Dimensions and Scoring Criteria
Each dimension is scored on a **0–100 integer scale**.
H1. Narrative Intent Realization First identify the narrative relationship type (direct continuation / temporal progression / parallel narrative / causal progression / tonal transition / other), then assess whether the video successfully expresses it.
81–100: Video fully and clearly realizes the narrative relationship described in the prompt; expressive approach exceeds expectations
61–80: Video fully realizes all required narrative relationships
41–60: Narrative relationship largely realized; isolated details imprecisely expressed but comprehension unaffected
21–40: Narrative relationship partially realized; key details missing; audience can infer but will be confused
0–20: Narrative relationship largely unrealized; prompt must be read to understand; or video contradicts the narrative design
H2. Visual Narrative Clarity Whether an audience watching the video alone (without reading the prompt) can understand the narrative relationship.
81–100: Narrative relationship very clear in the frame; no additional explanation needed; visual storytelling technique outstanding
61–80: Narrative relationship fully clear and comprehensible in the frame; all human viewers can understand without additional explanation
41–60: Basically comprehensible; some cues insufficiently explicit; human viewers need brief reflection to understand
21–40: Narrative relationship ambiguous; human viewers need textual explanation to understand
0–20: Narrative relationship very ambiguous; difficult to understand even after repeated viewing; or narrative entirely unreadable from the video
H3. Information Progression & Non-Redundancy Whether the current shot provides new visual information and advances the story.
81–100: Current shot clearly advances the narrative; presents new content; removing any frame would result in information loss
61–80: Current shot advances the narrative; provides new information; removal would result in content loss
41–60: Basically provides new information; minor overlap with preceding shot; narrative still progresses
21–40: New information limited; visual content noticeably overlapping; narrative progression perceptibly weak
0–20: Almost no new visual information; frames highly similar or entirely redundant; narrative stalled

Figure 15: Video evaluation prompt used in ReelBench.

ReelBench Video Evaluation Output Format

Unified Output Format

```
{
  "{DIM_CODE}_{sub_name}": {
    "score": 0,
    "detail": "Scoring rationale with specific visual evidence..."
  },
  "score": 0,
  "summary": "50-100 word overall evaluation"
}
```

Figure 16: Video evaluation prompt used in ReelBench.

A.3 Auxiliary Storyboard Quality Evaluation Dimensions

Storyboard quality evaluation is organized into two categories, shot-level quality and sequence-level coherence, and serves as an auxiliary module for storyboard optimization and static reference quality analysis.

At the shot level, Subject Alignment (SA) evaluates whether generated narrative subjects are accurate and checks for omissions, misidentifications, or hallucinated entities. Attribute Alignment (AA) evaluates entity-level details such as character species, gender, clothing, accessories, and poses. Relation Alignment (RA) evaluates spatial relationships, including character interactions and relative positioning. Style Alignment (STA) evaluates consistency with the intended artistic style. Scene Alignment (SCA) evaluates environmental settings and key background landmarks. Composition & Visual Quality (CVQ) evaluates production quality, including distortions, commonsense violations, shot composition, and framing.

At the sequence level, Character Consistency (CC) evaluates the stability of character visual identity across an episode. Style Consistency (STC) evaluates consistency of artistic style, tone, and texture details throughout the sequence. Narrative Continuity (NCT) evaluates causal progression, state inheritance, and narrative logic across storyboard frames. Scene Environment Stability (ES) evaluates coherent evolution of background structures, architectural characteristics, and environmental lighting across the sequence.

B Human Evaluation Details

To validate the consistency between ReelBench automatic evaluation and human perception, we design a dimension-wise pairwise preference human evaluation protocol. Annotators are asked to make relative preference judgments between two candidate videos generated from the same story. This pairwise setup reduces noise caused by inconsistent scoring scales across annotators and is more suitable for long-form multi-shot narrative videos.

B.1 Evaluation Interface and Displayed Content

The evaluation interface simultaneously displays two complete story videos side by side, along with shot-level action descriptions beneath the videos (see Figure 17). The left-right order is randomly shuffled for each annotator. The interface hides model names, generation conditions, optimization labels, and automatic evaluation scores.

Storyboard images are not shown during formal evaluation because they may directly reveal the source first frame in I2V settings, potentially compromising blind evaluation of generation conditions. Asset images are also hidden by default unless future experiments explicitly focus on character identity consistency.

Each task contains eight judgment questions: one overall preference question and seven questions corresponding to the ReelBench video evaluation dimensions. Each question uses a three-way choice format: Left is better, Right is better, or Similar / Cannot judge. Annotators may also select

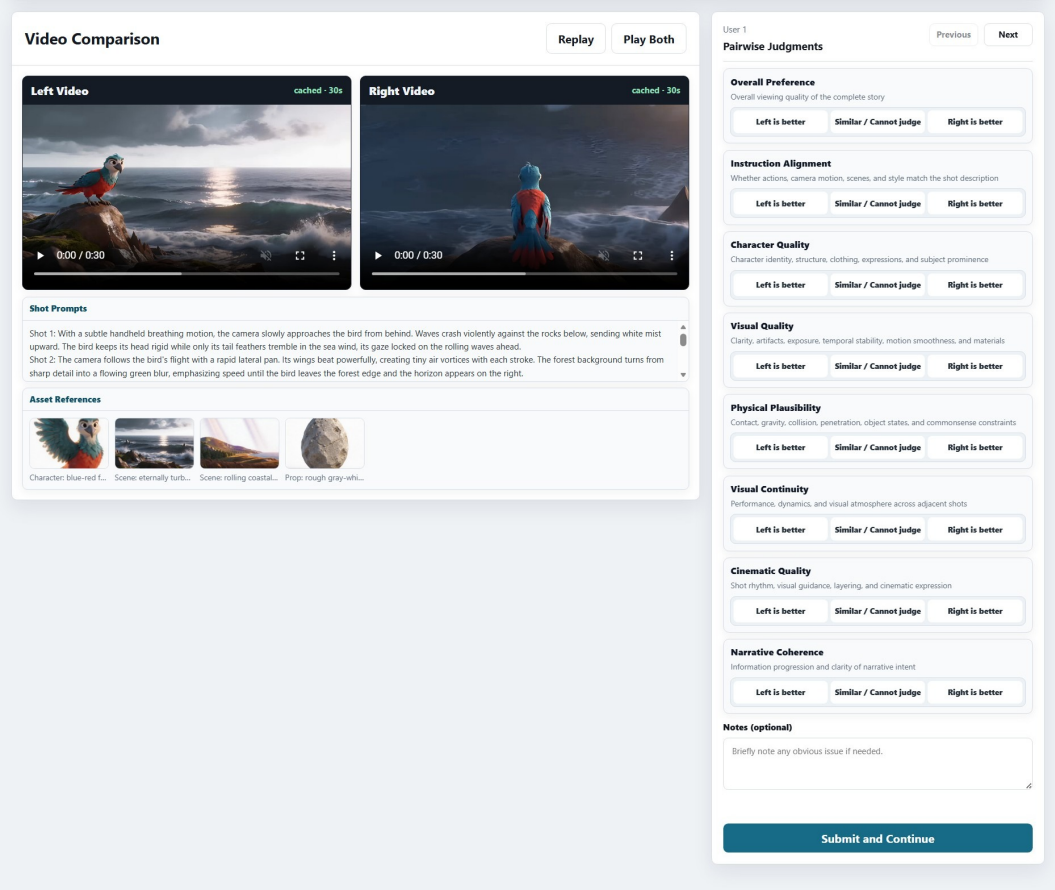

Figure 17: Screenshot of the human evaluation interface. Annotators watch two candidate videos generated from the same story, along with the shot prompts and asset references, and provide pairwise preferences for overall quality and each of the seven evaluation dimensions.

failure reasons, including instruction mismatch, subject errors, technical quality issues, physical implausibility, cross-shot inconsistency, limited cinematic quality, and unclear narrative.

Table 5: Correspondence between human evaluation questions and ReelBench dimensions.

Human Evaluation Question	Corresponding Dimension	Evaluation Focus
Overall Preference	Overall	Overall viewing quality of the complete story
Instruction Alignment	Instruction Alignment	Whether actions, camera motion, scenes, and style match the shot description
Character Quality	Character Quality	Character identity, structure, clothing, expressions, and subject prominence
Visual Quality	Visual Quality	Clarity, artifacts, exposure, temporal stability, motion smoothness, and materials
Physical Plausibility	Physical Plausibility	Contact, gravity, collision, penetration, object states, and commonsense constraints
Visual Continuity	Visual Continuity	Performance, dynamics, and visual atmosphere across adjacent shots
Cinematic Quality	Cinematic Quality	Shot rhythm, visual guidance, layering, and cinematic expression
Narrative Coherence	Narrative Coherence	Information progression and clarity of narrative intent

B.2 Sampling Strategy

Human evaluation samples are drawn through stratified sampling from the controlled image-reference and cross-model experiments. Each pair contains videos generated from the same story to avoid confounding effects from narrative content.

Table 6: Composition of human evaluation samples.

Source	Pair Type	Video Pairs
Controlled Reference Ablation	I2V vs. A-R2V	13
Controlled Reference Ablation	I2V vs. SA-R2V	13
Controlled Reference Ablation	SA-R2V vs. A-R2V	14
Cross-model / Multi-style / Complexity Evaluation	Cross-model Pairwise Comparison	120
Total	–	160

Cross-model samples are further stratified by visual style and complexity level, covering Pixar-style animation, 2D anime, and live-action realism, as well as low, medium, and high complexity settings. For cross-model comparisons, we include samples with large, medium, and small automatic score differences to avoid selecting only obviously distinguishable cases.

B.3 Annotators and Quality Control

Each video pair is assessed by three independent annotators, with 480 annotations collected across 160 video pairs. Each annotation covers overall preference and seven evaluation dimensions, yielding 3,840 preference judgments. Before formal annotation, annotators read concise instructions and complete practice tasks to familiarize themselves with the evaluation dimensions.

The system records each annotator's username, submission time, task ID, randomized left-right mapping, dimension-wise selections, failure reasons, and optional comments.

To ensure annotation quality, we hide all model and condition names to reduce bias from brand perception, open-source versus closed-source assumptions, and expectations tied to experimental hypotheses. We randomize left-right ordering across annotators to reduce positional bias. We also report majority agreement rates and conduct sensitivity analyses for annotators who repeatedly select the same side, finish tasks unusually quickly, or consistently deviate from majority judgments.

B.4 Ethics Review and Participant Risk

The human evaluation only involves pairwise preference judgments on generated videos and does not collect sensitive personal information from annotators. Participants are informed of the task purpose, the types of content to be evaluated, the expected annotation procedure, and their right to stop participation at any time. The evaluation interface records only information needed for quality control and aggregation, such as anonymized annotator identifiers, task IDs, submission time, dimension-wise selections, failure reasons, and optional comments.

C Evaluation Statistics and Reliability Analysis

C.1 Pairwise Agreement Rate

For each sample pair and evaluation dimension, we compare whether the side with the higher ReelBench automatic score matches human preference.

For sample pair i and dimension d, let the automatic evaluation scores for the left and right videos be $s^L_{i,d}$ and $s^R_{i,d}$, respectively. If $s^L_{i,d} > s^R_{i,d}$, the automatic preference is assigned to the left side; otherwise, to the right side. If the scores are equal or differ by less than threshold ϵ, the sample pair is treated as an automatic tie and excluded from the denominator.

Human preference is determined via majority voting. If the majority of annotators choose the same side preferred by the automatic evaluator, the sample is counted as an agreement; otherwise, it is counted as a disagreement. If the majority selects "Similar / Cannot judge," the sample is counted separately as a human tie and excluded from agreement statistics.

The agreement rate for dimension d is defined as:

$$Acc_d = \frac{N_d^{agree}}{N_d^{valid}},$$

where N_d^{valid} denotes the number of valid sample pairs after excluding automatic ties and human-majority ties, and N_d^{agree} denotes the number of samples where automatic and human preferences agree.

Evaluator Replacement. Using the human evaluation samples from the original overall-preference comparison, we replace the default evaluator Qwen3.5-Plus with three alternative VLMs, keeping the original prompts, scoring rules, and aggregation procedure unchanged. Table 7 reports the agreement with human preferences for all four evaluators. The agreement remains above 90% under every replacement, indicating that ReelBench's agreement with human preferences is robust to the choice of evaluator.

Table 7: Agreement with human preferences under evaluator replacement on the original overall-preference samples.

Evaluator	Agreement (%)
Qwen3.5-Plus (default)	95.60
Qwen3.6-27B (local)	93.33
GPT-5.6 Luna	95.00
Gemini 3.8 Flash	95.12

C.2 Majority Agreement Rate

The majority agreement rate measures whether annotators form stable majority opinions. For each sample pair, we compute the proportion of votes received by the most selected option and average this proportion across all samples within the same dimension.

This metric does not require annotators to share a consistent absolute scoring scale and is therefore more suitable for the three-choice pairwise preference setting adopted in this work.

Complete updated results are reported in Table 4 of the main paper. Pairwise agreement rate serves as the primary reliability metric for human evaluation, while majority agreement rate reflects the stability of consensus among annotators. When the quality difference between two candidate videos is small, annotators may choose "Similar / Cannot judge." Therefore, instead of using continuous correlation coefficients as the primary evidence, we directly report the two statistics most aligned with the pairwise preference protocol.

C.3 Significance Tests

Controlled Reference Comparison. We pair overall scores by story in the controlled Pixar-style experiment with Vidu Q2 and perform paired t-tests with Benjamini–Hochberg false discovery rate

correction across the three comparisons. Table 8 reports the paired sample size, mean difference, adjusted p-value, and paired effect size (Cohen's d_z); 95% confidence intervals are given below the table. Differences are computed from the story pairs included in each test.

Table 8: Paired significance tests for controlled reference conditions. Differences are condition A minus condition B.

Condition A	Condition B	n	$\Delta(A-B)$	Adjusted p	d_z
SA-R2V	A-R2V	100	+2.89	< 0.001	1.07
SA-R2V	I2V	99	+3.43	< 0.001	1.16
A-R2V	I2V	99	+0.54	0.054	0.20

The pairing unit is a story. The 95% confidence intervals for the three differences are [2.36, 3.43], [2.84, 4.02], and $[-0.01, 1.09]$, respectively.

SA-R2V significantly outperforms both A-R2V and I2V with large effect sizes, supporting the complementary roles of storyboard structure and asset identity references. The mean difference between A-R2V and I2V is small and its confidence interval includes zero, providing no evidence of a stable overall advantage. This is consistent with the finding that first-frame anchoring does not necessarily improve overall quality: the combined references have statistical support under this model and style, whereas the two individual reference forms do not differ significantly. A nonsignificant result does not establish equivalence.

We further compute the minimum and maximum shot score within each story and then average these statistics across stories. The mean minimum-shot scores are 64.64, 60.78, and 60.03 for SA-R2V, A-R2V, and I2V, respectively, whereas the corresponding mean maximum-shot scores are 75.70, 73.43, and 73.37. Their mean within-story shot-score standard deviations are 4.15, 4.62, and 4.91. Thus, SA-R2V's larger improvement occurs among low-scoring shots and is accompanied by lower within-story quality variation.

Cross-Model Comparison. We perform paired t-tests across the 11 model and reference-condition combinations, pairing the same story within the same style. Each comparison includes 60 story-style pairs. Table 9 reports all 55 comparisons, with mean paired differences, Benjamini–Hochberg-adjusted p-values, and Cohen's d_z.

The three reference conditions within Seedance 2.0 do not differ significantly (adjusted $p = 0.570$, 0.195, and 0.340; $d_z = 0.08$, 0.17, and 0.13). For Vidu Q2, SA-R2V exceeds A-R2V and I2V by 1.46 and 2.09 points, respectively (both adjusted $p < 0.001$; $d_z = 0.55$ and 0.74). These different patterns are consistent with reference effectiveness depending on how a model uses visual constraints.

Table 9: Complete cross-model paired significance tests. Differences are condition A minus condition B.

Condition A	Condition B	n	$\Delta(A-B)$	Adjusted p	d_z
Seedance 2.0 (A-R2V)	Seedance 2.0 (I2V)	60	+0.21	0.570	0.08
Seedance 2.0 (A-R2V)	Seedance 2.0 (SA-R2V)	60	+0.59	0.195	0.17
Seedance 2.0 (A-R2V)	Vidu Q2 (SA-R2V)	60	+2.43	< 0.001	0.65
Seedance 2.0 (A-R2V)	Vidu Q2 (A-R2V)	60	+3.90	< 0.001	1.22
Seedance 2.0 (A-R2V)	Vidu Q3 (I2V)	60	+4.08	< 0.001	1.31
Seedance 2.0 (A-R2V)	Vidu Q2 (I2V)	60	+4.53	< 0.001	1.53
Seedance 2.0 (A-R2V)	Wan2.2-TI2V-5B (I2V)	60	+7.66	< 0.001	2.59
Seedance 2.0 (A-R2V)	LTX-Video 2.3 (I2V)	60	+9.51	< 0.001	2.92
Seedance 2.0 (A-R2V)	SkyReels-V3 (SA-R2V)	60	+11.44	< 0.001	2.87
Seedance 2.0 (A-R2V)	LTX-Video (I2V)	60	+13.91	< 0.001	2.86
Seedance 2.0 (I2V)	Seedance 2.0 (SA-R2V)	60	+0.38	0.340	0.13
Seedance 2.0 (I2V)	Vidu Q2 (SA-R2V)	60	+2.22	< 0.001	0.81
Seedance 2.0 (I2V)	Vidu Q2 (A-R2V)	60	+3.69	< 0.001	1.55
Seedance 2.0 (I2V)	Vidu Q3 (I2V)	60	+3.87	< 0.001	1.56
Seedance 2.0 (I2V)	Vidu Q2 (I2V)	60	+4.32	< 0.001	1.81
Seedance 2.0 (I2V)	Wan2.2-TI2V-5B (I2V)	60	+7.45	< 0.001	3.06
Seedance 2.0 (I2V)	LTX-Video 2.3 (I2V)	60	+9.30	< 0.001	3.36

Continued on next page

Table 9 (continued)

Condition A	Condition B	n	$\Delta(A-B)$	Adjusted p	d_z
Seedance 2.0 (I2V)	SkyReels-V3 (SA-R2V)	60	+11.22	< 0.001	3.36
Seedance 2.0 (I2V)	LTX-Video (I2V)	60	+13.70	< 0.001	3.07
Seedance 2.0 (SA-R2V)	Vidu Q2 (SA-R2V)	60	+1.85	< 0.001	0.51
Seedance 2.0 (SA-R2V)	Vidu Q2 (A-R2V)	60	+3.31	< 0.001	1.05
Seedance 2.0 (SA-R2V)	Vidu Q3 (I2V)	60	+3.49	< 0.001	1.25
Seedance 2.0 (SA-R2V)	Vidu Q2 (I2V)	60	+3.94	< 0.001	1.33
Seedance 2.0 (SA-R2V)	Wan2.2-TI2V-5B (I2V)	60	+7.07	< 0.001	2.23
Seedance 2.0 (SA-R2V)	LTX-Video 2.3 (I2V)	60	+8.93	< 0.001	2.34
Seedance 2.0 (SA-R2V)	SkyReels-V3 (SA-R2V)	60	+10.85	< 0.001	2.98
Seedance 2.0 (SA-R2V)	LTX-Video (I2V)	60	+13.32	< 0.001	2.87
Vidu Q2 (SA-R2V)	Vidu Q2 (A-R2V)	60	+1.46	< 0.001	0.55
Vidu Q2 (SA-R2V)	Vidu Q3 (I2V)	60	+1.64	< 0.001	0.54
Vidu Q2 (SA-R2V)	Vidu Q2 (I2V)	60	+2.09	< 0.001	0.74
Vidu Q2 (SA-R2V)	Wan2.2-TI2V-5B (I2V)	60	+5.23	< 0.001	1.62
Vidu Q2 (SA-R2V)	LTX-Video 2.3 (I2V)	60	+7.08	< 0.001	1.85
Vidu Q2 (SA-R2V)	SkyReels-V3 (SA-R2V)	60	+9.00	< 0.001	2.55
Vidu Q2 (SA-R2V)	LTX-Video (I2V)	60	+11.47	< 0.001	2.35
Vidu Q2 (A-R2V)	Vidu Q3 (I2V)	60	+0.18	0.571	0.07
Vidu Q2 (A-R2V)	Vidu Q2 (I2V)	60	+0.63	0.045	0.27
Vidu Q2 (A-R2V)	Wan2.2-TI2V-5B (I2V)	60	+3.76	< 0.001	1.25
Vidu Q2 (A-R2V)	LTX-Video 2.3 (I2V)	60	+5.62	< 0.001	1.56
Vidu Q2 (A-R2V)	SkyReels-V3 (SA-R2V)	60	+7.54	< 0.001	2.19
Vidu Q2 (A-R2V)	LTX-Video (I2V)	60	+10.01	< 0.001	2.34
Vidu Q3 (I2V)	Vidu Q2 (I2V)	60	+0.45	0.099	0.22
Vidu Q3 (I2V)	Wan2.2-TI2V-5B (I2V)	60	+3.58	< 0.001	1.35
Vidu Q3 (I2V)	LTX-Video 2.3 (I2V)	60	+5.44	< 0.001	1.76
Vidu Q3 (I2V)	SkyReels-V3 (SA-R2V)	60	+7.36	< 0.001	2.27
Vidu Q3 (I2V)	LTX-Video (I2V)	60	+9.83	< 0.001	2.30
Vidu Q2 (I2V)	Wan2.2-TI2V-5B (I2V)	60	+3.13	< 0.001	1.21
Vidu Q2 (I2V)	LTX-Video 2.3 (I2V)	60	+4.99	< 0.001	1.73
Vidu Q2 (I2V)	SkyReels-V3 (SA-R2V)	60	+6.91	< 0.001	2.03
Vidu Q2 (I2V)	LTX-Video (I2V)	60	+9.38	< 0.001	2.44
Wan2.2-TI2V-5B (I2V)	LTX-Video 2.3 (I2V)	60	+1.85	< 0.001	0.69
Wan2.2-TI2V-5B (I2V)	SkyReels-V3 (SA-R2V)	60	+3.78	< 0.001	1.01
Wan2.2-TI2V-5B (I2V)	LTX-Video (I2V)	60	+6.25	< 0.001	1.53
LTX-Video 2.3 (I2V)	SkyReels-V3 (SA-R2V)	60	+1.92	< 0.001	0.48
LTX-Video 2.3 (I2V)	LTX-Video (I2V)	60	+4.39	< 0.001	1.10
SkyReels-V3 (SA-R2V)	LTX-Video (I2V)	60	+2.47	< 0.001	0.50

The pairing unit is the same story in the same style. Both significance tables use BH-adjusted p-values and paired effect sizes d_z.

C.4 Repeated Evaluation

We repeat evaluation on fixed half-sized subsets, keeping generated videos, evaluator, prompts, scoring rules, and aggregation unchanged. Run-1 denotes the original evaluation on the same subset; Run-2 and Run-3 are additional evaluations. The controlled experiment compares three runs with 148 matched condition–story units. The cross-model experiment compares Run-1 and Run-2 with 319 matched model–style–story units.

Table 10: Agreement between repeated and original evaluations. n counts matched units. Mean shifts are repeated minus original scores; all statistics use the same subset.

Experiment	Comparison	n	Mean shift	MAE	RMSE	Pearson	Spearman
Controlled	Run-2 vs. Run-1	148	+0.71	1.38	1.80	0.847	0.843
Controlled	Run-3 vs. Run-1	148	+1.40	1.79	2.28	0.817	0.812
Cross-model	Run-2 vs. Run-1	319	+1.34	1.92	2.41	0.935	0.933

SA-R2V has the highest condition mean across all three controlled-experiment runs. In the cross-model experiment, Pearson and Spearman correlations are 0.935 and 0.933, respectively. Despite upward shifts in absolute scores, the main reference-condition ordering and consistency of cross-model scores are maintained.

D Evaluation Visualizations and Case Studies

D.1 Physical Bottleneck Case Study Across Multiple Reference Interfaces

This case supports the physical bottleneck finding in Section 4.3. The visual style is 2D anime. The shot comes from "Yu Gong Moves the Mountains" and includes mountain excavation, tool swinging, multi-person collaboration, and moving rubble. We compare multiple Vidu Q2 reference interfaces for the same episode and focus on shot_3 to show that physical interaction errors appear across interfaces.

Table 11: Comparison of Vidu Q2 under different reference interfaces.

Model	Reference	Overall	PP	Visualization Focus
Vidu Q2	SA-R2V	62.27	49.3	Weak tool interaction and rubble feedback
Vidu Q2	I2V	64.00	50.9	Unstable force and collision dynamics
Vidu Q2	A-R2V	62.73	49.5	Misaligned excavation feedback

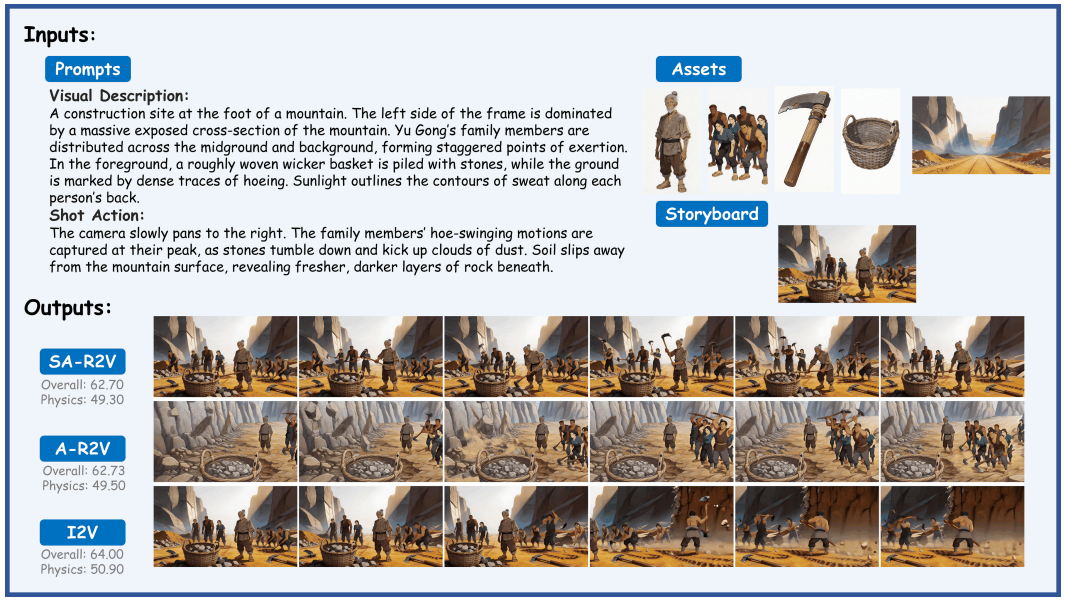

Figure 18: Qualitative example of the shared physics bottleneck across reference conditions. In the "Yu Gong Moves Mountains" case, the same shot requires tool use, multi-person coordination, stone movement, and contact feedback. Although different reference conditions provide different visual anchors, all generated results struggle with force transfer, tool–object contact, debris motion, and coordinated physical interaction. This supports our observation that static image references can constrain appearance and layout, but cannot fully compensate for weak temporal and physical modeling.

Figure 18 shows key frames of the same shot_3 under three reference interfaces, with focus on tool contact, rubble feedback, and multi-person collaboration.

The physical plausibility scores of all three results are low, and the low scores are not limited to one reference interface. This case indicates that image references can provide character, mountain, and tool appearance, but tool contact, force feedback, rubble displacement, and multi-person collaboration

still need to be generated by the video model over time. These physical errors are therefore better viewed as a shared capability boundary of current video models, rather than an issue that can be solved by adding more static reference images.

D.2 Case Study on Style Differences

This case examines style differences in Section 4.3. We fix the model to Vidu Q2 A-R2V and vary only the visual style of assets and storyboards. The shot depicts a street music scene under an urban overpass, including walking, guitar playing, a crouching character, gaze interaction, and puddles after rain.

Table 12: Performance of Vidu Q2 A-R2V under different visual styles.

Style	Overall	IA	CQ	VQ	PP	VC	Cine	NC
Pixar	72.40	71.4	76.6	69.4	65.5	78.8	73.7	74.4
2D anime	61.95	63.8	66.4	60.1	56.9	60.2	58.9	66.2
live-action	55.00	50.6	61.1	56.8	54.9	55.4	52.3	52.2

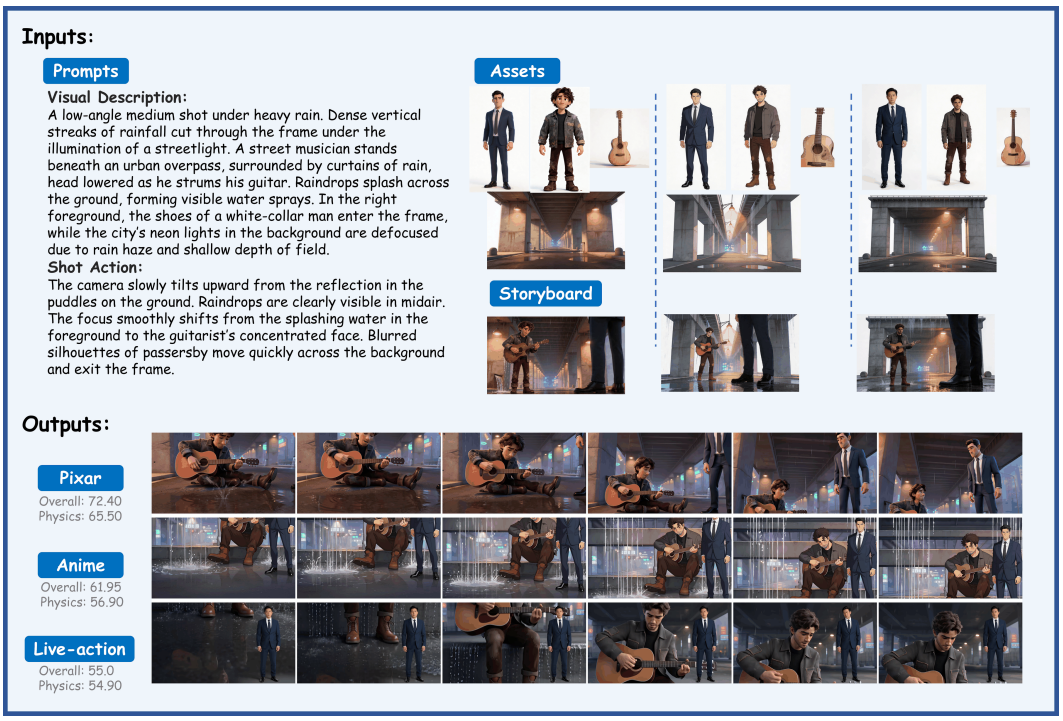

Figure 19: Qualitative comparison across visual styles under the same model and reference setting. The street musician shot is generated with the same A-R2V setting but different visual styles: Pixar-style animation, 2D anime, and live-action realism. Pixar-style generation is more tolerant of stylized motion and appearance variation, whereas live-action realism exposes stricter requirements on body structure, guitar-playing pose, rain interaction, lighting, and perspective consistency. The case shows how visual style changes the effective difficulty of image-reference-conditioned video generation.

Figure 19 keeps the model and reference interface fixed and compares key frames of the same shot_2 under Pixar-style animation, 2D anime, and live-action realism styles.

Under the Pixar-style animation, exaggerated character proportions and unified materials provide a higher tolerance for minor motion errors. In contrast, live-action realism requires stricter consistency in human anatomy, hand poses during guitar playing, skin material, lighting, shadow, and perspective, making visual and physical problems easier to expose. This case supports the conclusion in Section 4.3 that visual style changes the effectiveness of image reference conditions.

D.3 Case Study on Reference Complexity

This case analyzes reference complexity in Section 4.3. We fix the model to Vidu Q2 A-R2V and compare stories with different asset counts under the Pixar-style animation. We count assets as the number of distinct character, scene, and prop reference images used in each story. The low-asset sample mainly involves single-person adventure action, while the high-asset sample includes a full elephant and multiple blind men in group discussion, which more easily exposes spatial relation and multi-subject coordination errors.

Table 13: Comparison between low- and high-asset samples.

Sample Type	Number of Assets	Overall	PP
Low-asset sample	1	68.60	56.2
High-asset sample	6	70.03	50.7

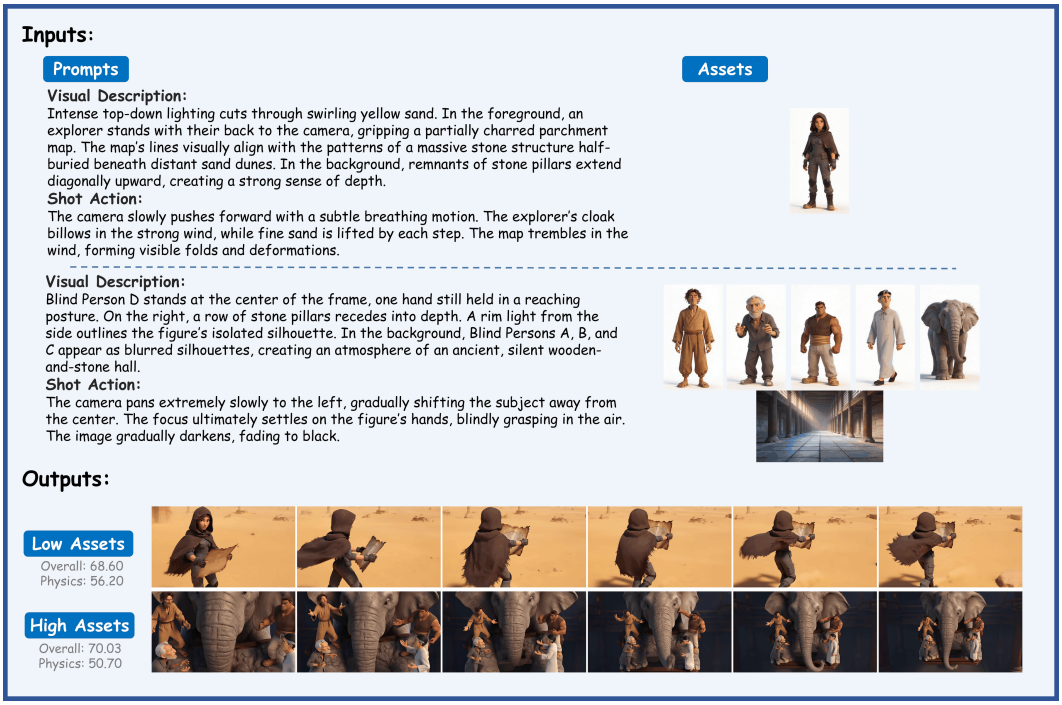

Figure 20: Qualitative comparison between low- and high-reference-complexity examples. The low-asset case mainly involves a single explorer, while the high-asset case contains an elephant, multiple characters, and a shared indoor space. Although the high-asset example does not necessarily receive a lower overall score, it more clearly exposes failures in contact relations, body orientation, multi-subject coordination, and spatial organization. This example shows that reference complexity is not a simple monotonic difficulty measure, but changes the types of errors that become visible.

Figure 20 presents low-asset and high-asset samples and shows how visible error types change as asset complexity increases. In the high-asset sample, the full elephant and multiple blind men appear in one frame. The model must therefore keep consistency in touched body parts, body orientations, and debate positions, which makes contact-relation and spatial-organization errors easier to expose.

This case shows that asset count is not a simple indicator of overall difficulty. High-asset samples can sometimes maintain comparable or even slightly higher overall scores, but physical plausibility and multi-subject coordination are more likely to degrade. When a shot needs to simultaneously maintain the full body of the elephant, multiple blind men, the touched object, and the spatial relationships of the hall, the model is more likely to produce unclear contact relationships, drifting body states, and unstable spatial motion.

E STATIC IMAGE REFERENCE OPTIMIZATION RESULTS AND CASE STUDY

Section 4.2 reports the main conclusion on storyboard optimization. This appendix provides a detailed analysis of static image evaluation and optimization results and includes a representative case to show why static reference optimization does not reliably transfer to video quality.

For storyboards, we use a static storyboard evaluation system to assess both shot-level and cross-shot dimensions and generate fine-grained reports. Based on these reports, we optimize the storyboard descriptions, regenerate optimized storyboards, and re-evaluate them using the same evaluation protocol. We then adopt a best-of-2 strategy to select the best storyboard, compare storyboard quality before and after optimization, and further compare video quality when using the original and optimized storyboards as references.

Tables 14 and 15 show that optimization improves most storyboard-quality metrics, whereas Table 16 shows lower video scores when the optimized storyboards are used as references. Under SA-R2V, Overall decreases from 70.40 to 68.17 ($\Delta = -2.23$); under I2V, it decreases from 66.95 to 66.38 ($\Delta = -0.57$). This divergence reflects a mismatch between the objectives used to optimize static storyboards and the information needed for video generation. Storyboard optimization emphasizes composition, aesthetics, and semantic alignment within individual frames, whereas video generation also requires clear guidance for action progression, contact relations, object-state continuity, and cross-shot transitions. The effectiveness of a storyboard as a video reference therefore depends on both its static quality and its support for action execution and temporal state transitions.

Table 14: Single-frame storyboard quality before and after optimization.

Evaluation Dimension	Overall	SA	AA	RA	STA	SCA	CVQ
Before Optimization	82.44	92.54	94.43	84.13	92.85	87.92	96.95
After Optimization	86.08	95.11	96.06	90.03	93.14	89.30	97.87
Δ	+3.64	+2.57	+1.63	+5.90	+0.29	+1.38	+0.92

Table 15: Sequence-level storyboard quality before and after optimization.

Evaluation Dimension	Overall	NCT	ES	CC	STC
Before Optimization	79.86	80.46	87.46	86.08	92.62
After Optimization	79.98	81.39	87.89	86.20	91.85
Δ	+0.12	+0.93	+0.43	+0.12	-0.77

Table 16: Video quality changes when using optimized storyboards as references. Differences are computed from unrounded means.

Condition	Overall	IA	CQ	VQ	PP	VC	Cine	NC
SA-R2V Baseline	70.40	69.63	74.81	65.96	58.72	76.54	76.22	75.35
SA-R2V Optimized	68.17	65.94	72.77	64.67	58.09	73.73	74.21	71.97
Δ (SA-R2V)	-2.23	-3.69	-2.04	-1.29	-0.63	-2.81	-2.01	-3.38
I2V Baseline	66.95	65.90	68.08	62.94	57.62	70.83	74.45	73.73
I2V Optimized	66.38	65.89	68.03	62.38	57.60	69.80	73.18	72.15
Δ (I2V)	-0.57	0.00	-0.05	-0.56	-0.02	-1.03	-1.27	-1.57

In the representative case, Overall decreases from 69.75 to 63.43 under SA-R2V and from 67.05 to 65.92 under I2V. The optimized storyboards improve composition and subject presentation, but provide no clearer guidance for the sequence of knocking, waiting, and turning, the associated contact relations, or the corresponding cross-frame state changes. The resulting videos therefore receive lower scores in Narrative Coherence and Cinematic Quality. This case illustrates why a visually stronger storyboard does not necessarily provide more effective guidance for video generation.

F Interaction Analysis Between Style and Reference Complexity

To complement the cross-model analysis in Section 4.3, we further stratify the results in Table 3 by visual style and reference complexity. Reference complexity is categorized as low, medium, or high according to the number of unique assets in each story and the maximum number of references within a single shot. We examine how performance varies with complexity across video models and visual styles, and whether models can effectively use denser reference inputs.

Table 17 and Figure 21 show that the relationship between reference complexity and Overall is nonmonotonic and varies across models and visual styles. For example, Seedance 2.0 A-R2V performs similarly on low- and high-complexity Pixar-style samples, whereas SkyReels-V3 decreases by 4.90 points on live-action samples. To separate model and style effects, we group the 100 stories in the controlled experiment by asset count. As shown in Table 18, from 1–2 to 7–10 assets, mean Overall across the three reference conditions decreases from 70.07 to 68.11, while Physical Plausibility decreases from 60.99 to 55.32. The corresponding Physical Plausibility declines are 6.51, 5.67, and 4.84 points for SA-R2V, A-R2V, and I2V, approximately three times their decreases in Overall. These larger declines indicate that multi-asset scenes place greater demands on multi-subject interaction, contact relations, and prop-state consistency.



Figure 21: Stratified model performance across visual styles and reference complexity levels. The left panel reports average overall scores under Pixar-style animation, 2D anime, and live-action realism, while the right panel reports average overall scores under low, medium, and high reference complexity. The results show that reference complexity is not a simple monotonic difficulty factor: stronger models can maintain competitive performance under dense references, whereas weaker models are more sensitive to style and high-complexity multi-asset settings.

Figure 21 shows the stratified performance of models across visual styles and reference complexity.

Table 17: Cross-stratified overall scores by model, visual style, and reference complexity. Low–High is the low-complexity score minus the high-complexity score, computed from unrounded means; a positive value indicates degradation from low to high complexity.

Model / Interface	Style	Low	Medium	High	Low–High
Seedance 2.0 (A-R2V)	Pixar	70.56	69.81	70.59	-0.03
	2D anime	69.31	66.32	66.54	+2.77
	live-action	67.29	65.66	67.10	+0.19
Seedance 2.0 (I2V)	Pixar	71.04	70.51	70.05	+0.99
	2D anime	66.81	66.41	65.89	+0.92
	live-action	66.18	66.31	64.83	+1.34
Seedance 2.0 (SA-R2V)	Pixar	70.81	70.98	69.07	+1.74
	2D anime	67.11	65.35	66.91	+0.20
	live-action	64.78	65.19	65.06	-0.28
Vidu Q2 (SA-R2V)	Pixar	69.20	70.89	70.19	-0.99
	2D anime	64.03	63.45	63.11	+0.92
	live-action	63.20	62.20	61.34	+1.86
Vidu Q2 (A-R2V)	Pixar	66.44	68.08	67.08	-0.64
	2D anime	64.71	63.57	62.32	+2.40
	live-action	61.78	61.07	59.39	+2.39
Vidu Q3 (I2V)	Pixar	67.32	66.59	66.82	+0.50
	2D anime	63.57	62.18	64.14	-0.57
	live-action	61.42	60.98	61.73	-0.31
Vidu Q2 (I2V)	Pixar	68.56	66.35	66.01	+2.55
	2D anime	62.68	62.07	61.85	+0.82
	live-action	61.24	60.66	60.44	+0.79
Wan2.2-TI2V-5B (I2V)	Pixar	62.45	62.49	62.63	-0.17
	2D anime	59.24	58.88	58.06	+1.18
	live-action	58.70	59.18	59.42	-0.72
LTX-Video 2.3 (I2V)	Pixar	62.14	60.58	60.54	+1.60
	2D anime	54.79	56.43	56.99	-2.19
	live-action	58.00	57.35	58.37	-0.37
SkyReels-V3 (SA-R2V)	Pixar	60.95	61.19	59.49	+1.46
	2D anime	55.55	53.38	54.03	+1.52
	live-action	56.77	53.84	51.87	+4.90
LTX-Video (I2V)	Pixar	56.56	56.32	57.70	-1.14
	2D anime	50.03	51.44	51.36	-1.33
	live-action	55.97	52.67	54.32	+1.64

Table 18: Asset-count stratification in the controlled experiment.

Assets per Story	Stories	Overall	Physical Plausibility
1–2	8	70.07	60.99
3–4	46	68.37	58.46
5–6	34	68.17	57.46
7–10	12	68.11	55.32

Scores are story-equal means averaged across SA-R2V, A-R2V, and I2V. Asset count is the number of distinct character, scene, and prop reference images in each story.