ANIMACT: WHAT BREAKS WHEN CHARACTERS MEET? A CAPABILITY TAXONOMY FOR MULTI-REFERENCE ANIMATED VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-conditioned animated video generation requires models to preserve visually specified characters and props while executing a short story with the correct actions, interactions, and bindings. Animation makes this evaluation particularly challenging: stylized deformation and exaggerated motion may be intentional, while identity, contact, weight, and who-acted-on-whom must remain visually clear. Existing benchmarks capture parts of this problem, but do not jointly evaluate reference-grounded story execution within a single generated clip. We introduce **ANIMACT**, a holistic human evaluation benchmark that combines 14 overall criteria with eight targeted capability probes, covering reference fidelity, action execution, role assignment, interaction, identity preservation, and temporal, spatial, and physical stability. The benchmark contains 317 human-verified stories spanning 64 characters and 13 props; 95% are multi-character, and each conditions on 1–4 reference assets together with evaluation-only visual and action annotations. Across 1,902 clips from six reference-to-video systems, Seedance-2.5 is the clear overall leader, yet substantial failures remain: 40% of scripted character-to-character interaction edges do not occur, appearance consistency degrades as reference load increases while entity presence remains stable, and only 6.4% of clips pass all applicable evaluation axes. Finally, we show that AI judges recover aggregate model trends much better than individual human judgments, making them useful for relative model comparisons but not yet a replacement for fine-grained human evaluation.

1 INTRODUCTION

Recent advances in video generation have pushed the field beyond producing short, visually plausible clips Ho et al. (2022); Girdhar et al. (2024); Chen et al. (2024) toward coherent narratives with persistent characters and increasingly plausible motion, as seen in OpenAI's Sora Liu et al. (2024b), Google's Veo 3 Wiedemer et al. (2025), Kuaishou's Kling Team et al. (2025), Alibaba's Wan Wan et al. (2025), Tencent's HunyuanVideo Wu et al. (2025a), and ByteDance's Seedance in Seedance et al. (2026). This progress is particularly relevant to animation, which spans film and television, games, advertising, and educational media, yet has received substantially less attention than natural-video generation. Recent animation-specific efforts such as AniSora Jiang et al. (2025) and AnimationBench Wu et al. (2026) highlight this gap: models and evaluation protocols developed primarily for natural or realistic video do not necessarily capture the stylized appearance, exaggerated motion, and character consistency required by animation. Reference-conditioned video is especially relevant in this setting, where creators work with fixed characters and props rather than unconstrained text prompts. In such settings, however, knowing which model looks best on average is not enough. Different shots demand different capabilities: a close-up requires appearance fidelity, an action sequence requires identity to survive motion, a hand-off requires correct character–prop binding, and a multi-beat scene requires the intended events to occur in order. Practitioners therefore need a capability profile that reveals not only which models perform well, but *what they can be trusted to do* and where failures remain likely.

This requires evaluating more than reference fidelity. Given specific characters and props fixed by reference images, a model must preserve who appears and how they look, execute the requested

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.

actions with the correct participants, and maintain coherent motion, contact, objects, and scene geometry throughout the clip. These requirements are tightly coupled: a model may preserve every character but omit an action, complete the action with the wrong look-alike, or follow the story while introducing unstable objects or physically implausible interactions. In animation, this evaluation is further complicated by the fact that deformation, exaggeration, and departures from real-world motion may be intentional, while contact, weight, object binding, and who-acted-on-whom must still remain visually clear and physically sensible under the logic of the scene and the principles of animation Thomas & Johnston (1981); Jiang et al. (2025); Wu et al. (2026). The challenge is therefore not simply whether individual capabilities work in isolation, but whether they hold together while the model executes a story.

Existing evaluations capture important pieces of this problem, but not their conjunction. General-purpose suites measure visual quality, temporal consistency, and text alignment Huang et al. (2024; 2025); Zheng et al. (2025); Liu et al. (2024a); Han et al. (2025). Benchmarks for compositionality Sun et al. (2025), physical commonsense Meng et al. (2025); Bansal et al. (2025), and story completion Wang et al. (2025) probe actions and events, but their entities are specified primarily through text and therefore do not test whether the *correct referenced characters* perform them. Reference-conditioned benchmarks instead emphasize subject fidelity Yuan et al. (2026), character-referenced image sequences Zhuang et al. (2026), or multi-element conditioning Fei et al. (2025). Animation-specific evaluation has also begun to emerge: AniSora evaluates visual quality, motion, image–video consistency, and character consistency Jiang et al. (2025); AnimeReward studies human preferences over anime appearance and temporal consistency Zhu et al. (2025); and, more recently, AnimationBench evaluates character-centric image-to-video generation through animation principles, IP preservation, and broader video quality Wu et al. (2026). Multi-shot benchmarks follow characters across cuts Wu et al. (2025b); Shi et al. (2026); He et al. (2026), but distribute the narrative over separately generated shots, while multi-subject methods report identity entanglement and degraded interactions as the cast grows Liao et al. (2025); Meng et al. (2026); Deng et al. (2026), typically on method-specific evaluations. What remains missing is a holistic evaluation of *reference-grounded story execution*: whether multiple visually specified characters and props can preserve their identities while carrying out the intended actions and interactions within the same generated clip.

We introduce **ANIMACT (ANIMated-video Assessment by Capability Taxonomy)**, a holistic human evaluation study of multi-reference, story-driven animated video generation. Each task provides fixed visual references for its characters and props together with a short narrative, and asks a model to realize that story in a single generated clip. ANIMACT evaluates the resulting videos along two complementary views: broad criteria that characterize the quality and completeness of every generation, and targeted probes that isolate specific capabilities such as interaction, role assignment, identity preservation, and physical stability. This distinction lets us ask both *which model performs better overall* and *why a model succeeds or fails on a particular kind of production demand*. We rely on human judgments because many such failures require jointly tracking reference identity and temporal events; existing automated evaluators are generally validated on coarser dimensions or aggregate rankings Wang et al. (2025); Zhuang et al. (2026); Wu et al. (2026); Shi et al. (2026). To reduce subjective preference, evaluation is based on factual judgments about what appears and happens in the clip rather than generic visual appeal Liu et al. (2024a); Han et al. (2025); Zhang et al. (2026).

Our evaluation of six reference-to-video systems reveals complementary failures across capabilities. Seedance-2.5 ranks first among the evaluated models on most criteria, yet strong aggregate performance rarely translates into a fully correct clip. Models realize only 71.8% of scripted actions overall, and character-to-character interaction is particularly difficult: 40% of scripted interaction edges never occur. Increasing reference load leaves entity presence largely intact while progressively degrading appearance consistency. Failures also accumulate across otherwise strong capabilities, and only 6.4% of clips pass all applicable axes. These results suggest that the central challenge is not mastering any single aspect of reference-conditioned generation, but satisfying identity, narrative, interaction, and physical constraints *jointly*. Finally, we ask whether multimodal AI judges can replace human evaluation. We find that they recover aggregate model comparisons reasonably well, but do not reliably reproduce the individual human judgments needed for fine-grained capability analysis. Thus, AI judges are useful for relative model comparison, but human evaluation remains necessary for detailed ANIMACT assessment.

Our main contributions are:

- We formulate **reference-grounded animated video generation** as a holistic evaluation problem and introduce **ANIMACT**, a human evaluation benchmark with two complementary families: *overall rubric* that assess every clip for reference fidelity, action execution, and temporal, spatial, and physical stability, and *targeted probes* that isolate specific capabilities such as interaction, role assignment, occlusion recovery, and identity preservation.
- We construct a **human-verified corpus of 317 reference-grounded stories** spanning 64 characters and 13 props, with 95% multi-character scenes. Each story pairs 1–4 reference assets with a generation prompt and *evaluation-only annotations* of visual details and scripted key actions.
- We evaluate six reference-to-video systems and find **Seedance-2.5 to be the clear overall leader across the evaluated suite**, while uncovering complementary failures in action completion, character interaction, reference fidelity, and physical stability; only 6.4% of clips satisfy all applicable overall criteria.
- We show that AI judges recover aggregate model trends substantially better than they reproduce human judgments on individual clips, supporting their use for relative model comparison but not as a replacement for fine-grained human evaluation.

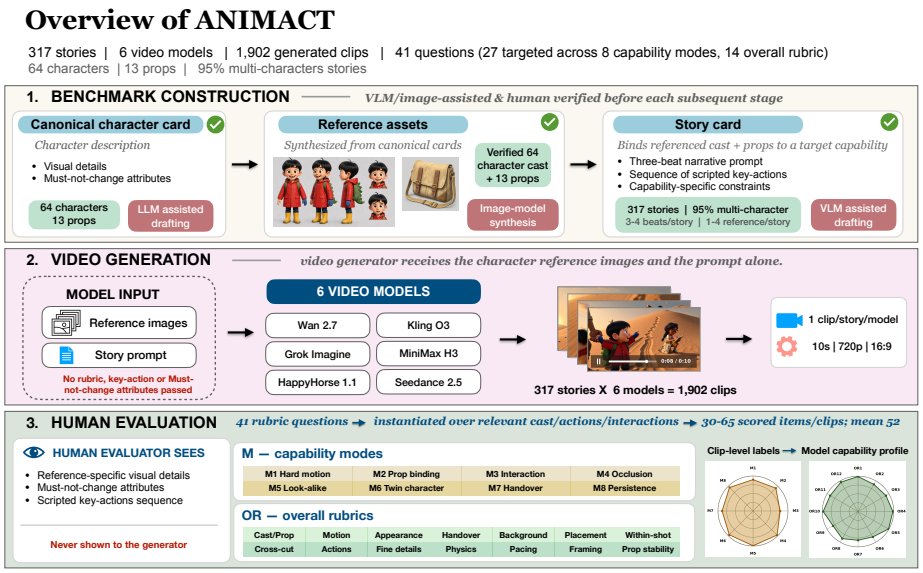

Figure 1: **Overview of ANIMACT.** ANIMACT comprises 317 stories spanning 64 characters and 13 props, each targeting one of eight capability modes (35–41 stories/mode). Six video-generation models produce 1,902 clips from reference images and story prompts alone; human evaluators additionally access hidden reference-specific constraints and scripted key actions. Evaluation uses 27 targeted questions across the eight modes and 14 overall rubric criteria to produce fine-grained clip- and model-level capability profiles.

2 ANIMACT Corpus and Evaluation

ANIMACT (ANIMated-video Assessment by Capability Taxonomy) evaluates whether a reference-to-video model can realize a short, multi-character story from fixed visual references specifying the characters and any relevant props. Each task pairs one to four such references with a single narrative prompt constructed from 3–4 ordered story beats, each denoting a major action or event in the intended sequence. The *story* is the unit of benchmark design, and the generated *clip* is the unit of measurement. We organize the study in three stages: (i) benchmark construction, comprising verified reference assets, capability-targeted stories, and their evaluation criteria; (ii) video generation under a common protocol; and (iii) human evaluation. The frozen benchmark contains 317 stories, yielding 1,902 clips across six video-generation models. We additionally study whether multimodal AI judges can reproduce the resulting human evaluations.

2.1 CORPUS CONSTRUCTION

Reference assets. We first construct a frozen bank of 64 characters and 13 props to control the visual identity of characters and props used across stories, as shown in Stage-1 in Fig.1. For each character, Gemini 3.1-Pro Google DeepMind (2026) is used to draft a canonical card, which we then manually refine to specify the character's visual appearance, distinguishing details, and `must_not_change` attributes. We generate character reference images from the canonical-card specifications and prop reference images from their visual descriptions using GPT-Image-2 [1]. Three annotators independently verify character references against their canonical-card descriptors and check consistency across rendered views; prop references are similarly checked against their visual specifications. Images with baked-in backgrounds, cropped body parts, missing or mismatched details, or cross-view inconsistencies are revised based on annotator feedback and re-verified before entering the frozen benchmark. Further details on the verification protocol and agreement statistics are provided in Appendix A.2.

Capability-targeted story construction. Each story is constructed around one of eight "M" capability modes, covering difficult motion, character–prop binding, interaction, identity preservation under occlusion or visually similar casts, handovers, and temporal persistence, as shown in Fig. 1. The target capability is built into the narrative rather than assigned post hoc: for example, an occlusion story requires a character to disappear and re-emerge, while a handover story requires an object to physically change hands. We use GPT-5.5 OpenAI (2026) to draft the story cards, and use the same model during story card revision. The modes are approximately balanced, with 35–41 stories per mode. Stories contain one to four characters and zero to two props, with at most four reference images per task; 95% involve multiple characters. Each story also stores evaluation-only metadata—reference-specific visual details, `must_not_change` attributes, and scripted key actions—used solely for evaluation. Interaction-focused stories include an ordered interaction graph for grading contact events individually. Before generation, each story card is independently human-verified for capability alignment, prompt–asset consistency, and feasibility; flagged cards are revised and re-verified. Full mode definitions, story schema, verification protocol, and corpus distributions are provided in Appendix A.3.

Evaluation taxonomy. As summarized in Fig. 1, we evaluate each clip with two complementary question families: 27 targeted **M** questions, each applied only to stories belonging to its corresponding capability mode, and 14 **OR** (*overall rubric*) criteria, each applied to every clip regardless of capability mode, covering reference fidelity, action execution, and temporal, spatial, and physical stability. The 41 rubric questions are instantiated over relevant characters, props, actions, and interactions, yielding 30–65 scored items per clip (52 on average). Questions are binary, ternary (*Yes/Partial/No*), or checklist-based, with conditional items applied only when relevant; scores are normalized so that higher values indicate better performance. Appendix A.5 provide further details on the taxonomy and evaluation protocol.

2.2 REFERENCE-TO-VIDEO GENERATION

We generate one clip per story with six reference-to-video models: Wan 2.7 (Wan et al. (2025)), Kling O3[2], Grok Imagine[3], MiniMax-H3 (MiniMax (2026)), Happy Horse 1.1[4], and Seedance 2.5 (ByteDance Seed (2026)). Each model receives the same narrative prompt and frozen reference images under a common protocol targeting 10 s, 16:9, and 720p, except Kling O3, whose endpoint fixes the output at 3840×2160. All models receive the same style instruction, "stylized 3D CGI, animated-film video.". Differences are otherwise limited to vendor-imposed constraints such as reference caps and supported output settings. We disable prompt expansion where available and keep safety filters at their default settings; refusals and API errors are recorded rather than regenerated under altered settings. Wan 2.7 also receives a fixed negative prompt suppressing speech and lip-sync because it generates audio unconditionally. Full endpoint configurations and payloads are provided in Appendix A.6.

[1] `https://developers.openai.com/api/docs/models/gpt-image-2`
[2] `https://app.klingai.com`
[3] `https://x.ai/news/grok-imagine-video-1-5`
[4] `https://fal.ai/models/alibaba/happy-horse/v1.1/reference-to-video`

2.3 HUMAN EVALUATION

Human evaluation is organized by story: evaluators see the reference assets, narrative prompt, evaluation-only visual constraints, and scripted key actions, and score each of the six clips using the applicable M and OR questions. Evaluating all six models on the same story controls for differences in story difficulty. Model identity is hidden, and clip order is shuffled within each story. Each clip requires scoring 52 items on average and takes approximately 20–30 minutes to evaluate. Given this annotation burden, we use an annotate-then-review protocol rather than collecting three independent annotations per clip. Labels are collected from 13 annotators and checked by 9 reviewers ; accordingly, we do not estimate a human performance ceiling. Further details on the annotation pool and review protocol are provided in Appendix A.7.

3 MAIN FINDINGS

We now report the human-evaluation results across the six models, beginning with overall performance under the OR criteria. We then use the targeted M probes to examine capability-specific failure patterns.

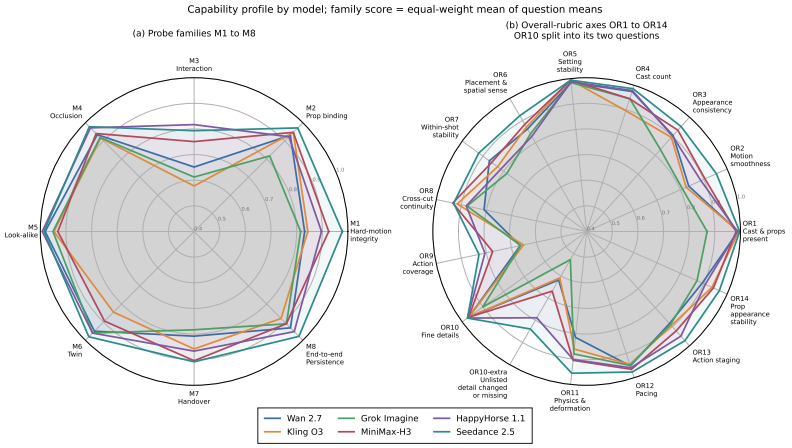

Figure 2: **ANIMACT capability profile across models.** A family or axis score is the equal-weight mean of its question means. (a) The eight probe families M1–M8, each scored only on stories designed to test that capability. (b) The fourteen overall-rubric axes, scored on every clip regardless of what its story was designed to test.

3.1 ONE CLEAR LEADER; INTERACTION IS THE WALL

We summarize the performance profiles of all six models across the two evaluation families in Fig. 2. On the overall criteria reported in Fig. 2(b), Seedance-2.5 leads on 13 of the 15 reported OR scores, with the two OR10 components reported separately because they use different response formats. Across all 23 rubric rows, Seedance-2.5 leads on 19, while MiniMax-H3 and HappyHorse-1.1 each lead on two. By mean video-clip-level score, the ordering is Seedance-2.5 $>$ MiniMax-H3 $>$ HappyHorse-1.1 $>$ Wan-2.7 $>$ Kling-O3 $>$ Grok-Imagine. However, story-level bootstrapping with Holm correction does not reliably separate HappyHorse, Wan and Kling (Appendix Table 7).

Importantly, this ordering is not an artifact of aggregating questions into rubric families. At the individual-question level, Seedance-2.5 leads on 32 of 58 questions, and the models retain the same ordering by mean rank. The complete question-level results are given in Appendix Table 8 and 9. While the Fig. 9 shows the within-family distributions underlying the aggregate scores, both in the Appendix. Despite these performance differences, Fig. 2(a) reveals a shared weakness: interaction (M3) is the lowest-scoring probe family for all six models. We examine this failure next.

Takeaway. Seedance-2.5 leads overall and at the question level, followed by MiniMax-H3, although several middle-ranked models are not reliably separated.

3.2 INTERACTION IS CONSISTENTLY UNDER-PRODUCED

Interaction (M3) is the weakest probe family for all six models. Across 39 stories, 339 of 840 scripted character-to-character interaction edges (40%) are absent overall. This omission occurs across every model, ranging from 23.6% for HappyHorse to 56% for Kling (Fig. 3a). Breaking failures down by contact type (Fig. 3b), braiding or weaving is absent in 67% of instances (no. of interaction edges, $n = 96$), compared with 31% for object hand-off ($n = 228$). This contrast suggests that sustained, coordinated interaction is harder to realize than a brief transfer between characters. The broader action metrics show that omission is not limited to interaction. Across all 317 stories, models realize only 71.8% of all the scripted actions (OR9); among actions that do occur, 87.2% are clearly staged and a further 10.1% partially staged (OR13), indicating that missing actions are a larger problem than poor staging (Fig. 10).

Action coverage declines as story demands increase, from a pooled mean of 0.93 at 1–3 scripted actions to 0.69 at ≥ 12, with a similar decline across prompt-length bins (see Fig.4). The two load variables are correlated ($\rho = +0.60$ over 317 stories); after controlling for each other, the association remains for prompt length ($\rho = -0.296$) but is weak for action count ($\rho = -0.086$). We therefore interpret the trend as an association between greater story demand and action omission, driven primarily by prompt length rather than by the number of scripted actions alone. Importantly, interaction remains the least-covered action category overall: contact actions achieve 63.3% coverage, compared with 74.4% for locomotion and 72.3% for prop manipulation (Fig. 11). Thus, the tendency to omit character-to-character interaction persists whether it is measured through the targeted M3 probe or through corpus-wide action coverage. M7 then examines a complementary question for handover stories: when contact does occur, how physically convincing is it? Here, mean scores for contact realism, weight, and staging are comparatively higher (0.738–0.963), suggesting that occurrence is the more persistent weakness.

Takeaway. Every model drops scripted character-to-character contact: 40% of scripted interaction edges do not occur, and contact is the least-covered action type overall.

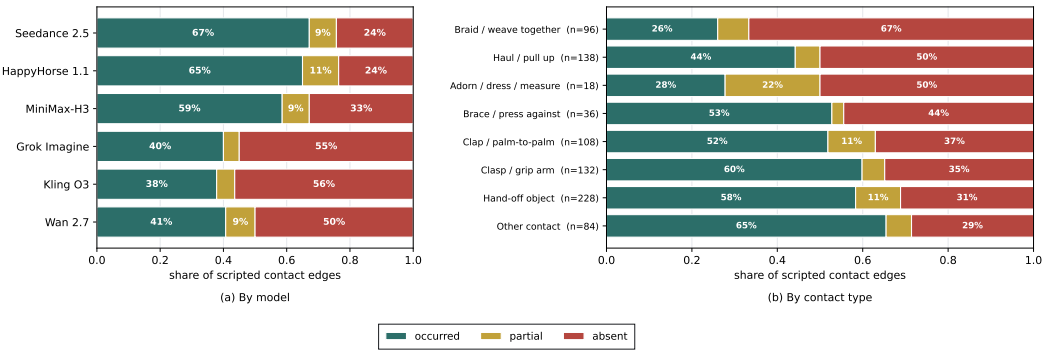

Figure 3: **Scripted interaction occurrence.** Interaction edges are labeled *Occurred*, *Partial*, or *Absent*. Across 39 stories and 6 models, 40% of 840 scripted interactions are absent. (a) By model. (b) By interaction type, pooled across models; n denotes the number of interaction edges, not stories.

3.3 REFERENCE LOAD STRAINS FIDELITY, NOT PRESENCE

Reference load is the number of reference images a clip must honor—cast plus props—and ranges from one to four in our corpus. As reference load increases, models remain good at preserving *which* entities appear, but become less reliable at preserving *how* they appear (Fig. 5). Cast and prop presence (OR1) remains essentially flat (0.958 to 0.966), whereas appearance consistency (OR3) falls from 0.983 at one reference to 0.858 at four. This decline persists after controlling for prompt length (Spearman $\rho = -0.25$; Appendix Fig. 8), while presence remains unaffected ($\rho = +0.04$). Fine-detail and prop fidelity also decline in the pooled view, but show much weaker independent associations with reference load after controlling for prompt length.

Takeaway. Entity presence stays essentially flat as reference load increases, while appearance consistency declines; this relationship persists after controlling for prompt length.

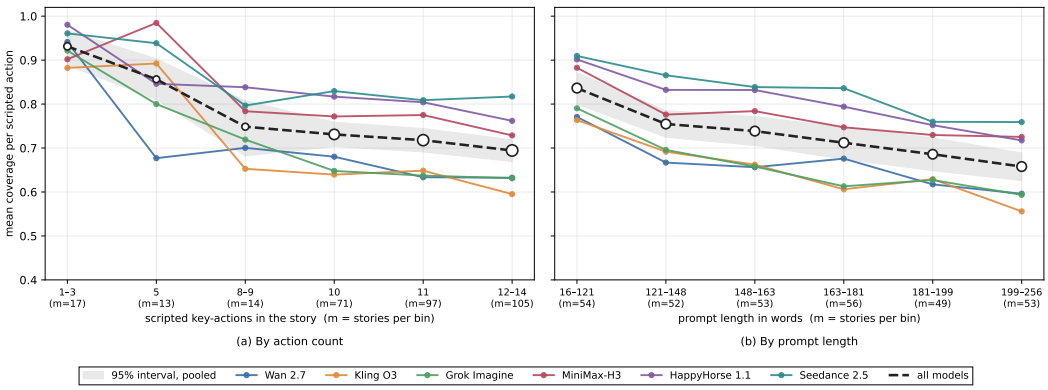

Figure 4: Action coverage (OR9) against how much a story demands, plotted for per model with the pooled mean dashed. (a) against the number of scripted key-actions, with counts merged (1-3, 8-9, 12-14) so no bin rests on fewer than ten stories. (b) against prompt length in equal-frequency bins. Marker area is proportional to the stories behind each point and the shaded band is a 95% bootstrap interval over stories.

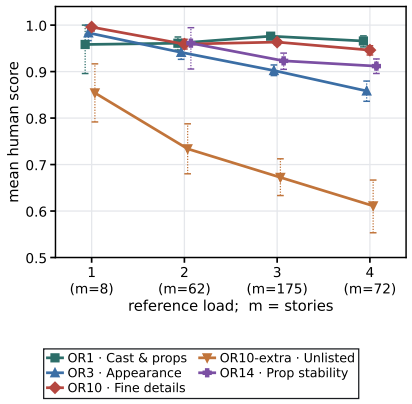

Figure 5: **Reference fidelity across reference load.** Load is the number of cast and prop images a clip must honor. Fidelity measures entity presence, character appearance, fine-detail preservation, and prop appearance. Error bars show 95% bootstrap CI over stories (2k resamples).

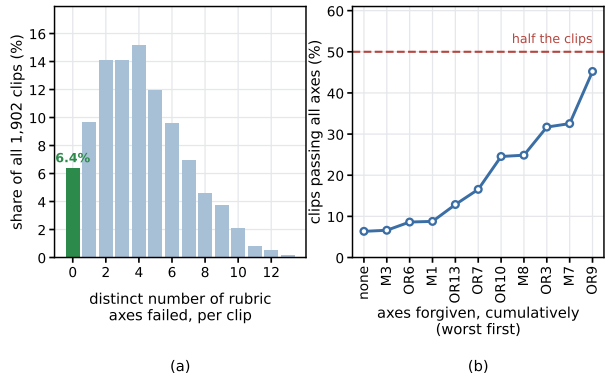

Figure 6: **Why a saturated rubric still fails almost every clip.** (a) Each bar counts clips by the number of distinct rubric axes they fail, combining all applicable OR axes with the single M axis probed by the story. Thus, 0 means no failed axes, 1 means exactly one failed axis, and so on. (b) Clean rate after cumulatively forgiving the k most frequently failed axes ($k = 10$); the axis removed at each step is shown on the x-axis.

3.4 Failures accumulate across rubric axes

We examine the rubric at the question level (Tables 8 and 9). Several capabilities look close to solved on one check but expose weaknesses on another. In the Twin probe (M6), both twins are almost always present and active (M6.1: 0.949–1.000), yet assigning the action to the correct twin (M6.3) drops to 0.800–1.000. The contrast is sharper in the overall rubric: Placement & spatial sense (OR6) is near ceiling for lighting and shadows (OR6: *light-shadow-shadow*, 0.975–0.994) but falls to 0.562–0.757 for *passes-through-solid*; Within-shot stability (OR7) preserves backgrounds well (0.956–0.987) but struggles with worn or held objects (0.574–0.811). These within-family gaps motivate asking whether strong checks coincide within the same clip.

We next ask whether strong question-level performance translates into success at the clip level. Only 121 of 1,902 clips (6.4%) fail no applicable rubric axis—i.e., pass all applicable OR criteria together with the targeted M criterion—and the median clip fails four axes (Fig. 6a). The clean-clip rate also rises only gradually as the most frequently failed axes are cumulatively forgiven (Fig. 6b), showing that failures are distributed across multiple rubric dimensions rather than concentrated in a few globally weak checks.

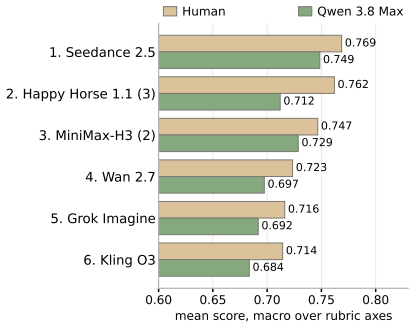

Figure 7: **The ANIMACT leaderboard.** Human annotation and Qwen 3.8 Max scores on the same rubric axes, shown in human rank order. Scores are rubric pass rates, not agreement.

Judge	ρ	ρ_c	Concord.	Minority R / P
Gemini 3.1 Pro	0.771	0.566	0.738	**0.330** / 0.173
Muse Spark 1.2	0.429	0.466	0.681	0.169 / 0.259
Qwen 3.8 Max	**0.886**	**0.693**	0.783	0.128 / 0.300
Majority-of-3	**0.886**	0.688	**0.796**	0.171 / **0.330**

Table 1: **AI-judge agreement with human.** ρ: leaderboard rank correlation; ρ_c: question-centred rank correlation; Concord.: question-level ordering agreement; Minority Recall/Prec.: performance on non-majority human labels.

Takeaway. Per-axis success overstates practical video quality: clean clips require many rubric constraints to hold jointly.

4 AI JUDGES FOR ANIMACT EVALUATION

We further investigate whether AI judges can reproduce human evaluation at the level of model comparison and individual judgments. Figure 7 compares the human ANIMACT leaderboard with the same rubric scored by Qwen 3.8 Max; for both, Y/P/N is mapped to 1/0.5/0, averaged over questions, and then over rubric axes. Qwen systematically assigns lower absolute scores, but largely preserves the relative ordering of the six models, yielding a leaderboard Spearman correlation of $\rho = 0.886$. The only rank change is a swap of HappyHorse 1.1 and MiniMax-H3. This agreement persists at a finer granularity. Qwen obtains a question-centred rank correlation of $\rho_c = 0.693$ and reproduces 78.3% of human question-level model orderings (95% CI: 0.74–0.82). Among the individual judges, Qwen shows the strongest overall agreement with human model comparisons (Table 1); majority voting improves ordering concordance only marginally, to 0.796.

However, this aggregate agreement does not translate to reliable individual judgments. On items where the human label departs from the question's majority label, Qwen recovers only 12.8% at 0.30 precision. Thus, aggregate agreement substantially overstates reliability on the fine-grained cases needed for capability analysis. We therefore use Qwen for relative model comparison, while human annotations remain the basis for absolute ANIMACT scores and fine-grained analysis. Further results across rubric families and aggregation levels are reported in Section C.

Takeaway. AI judges can compare models, but cannot yet replace the fine-grained human evaluation.

5 RELATED WORK

5.1 REFERENCE-TO-VIDEO GENERATION AND CAPABILITY EVALUATION

Video-generation benchmarks increasingly move beyond perceptual quality toward fine-grained capability evaluation. General-purpose suites such as VBench Huang et al. (2024), VBench++ Huang et al. (2025), EvalCrafter Liu et al. (2024a), and Video-Bench Han et al. (2025) evaluate dimensions including visual quality, temporal consistency, motion, and conditioning alignment, while VBench-2.0 Zheng et al. (2025) further evaluates intrinsic faithfulness. These benchmarks provide broad capability profiles, but are not designed around stories in which multiple referenced entities must preserve their identities while taking distinct roles in the same generation. Other benchmarks isolate specific failure modes. T2V-CompBench Sun et al. (2025) evaluates compositional generation; PhyGenBench Meng et al. (2025) and VideoPhy Bansal et al. (2025) probe physical commonsense and physical-law adherence; VMBench Ling et al. (2025) focuses on motion quality; and StoryEval

Wang et al. (2025) tests completion of sequential events. ANIMACT similarly uses capability-targeted tasks, but under reference grounding: the intended character must perform the intended action while its identity, bindings, interactions, and surrounding scene remain coherent.

Closer to ANIMACT are reference-conditioned benchmarks that evaluate generation from one or more visual references. OpenS2V-Eval Yuan et al. (2026) includes both single- and multi-subject settings and evaluates subject consistency, naturalness, and text relevance. A2-Bench Fei et al. (2025) is closer in conditioning setup, combining multiple referenced elements—including characters, objects, and backgrounds—and evaluating composition consistency, visual quality, and prompt following. More recently, AnimationBench Wu et al. (2026) evaluates character-centric image-to-video generation, primarily assessing a single focal character's IP preservation and animation quality, rather than multi-character interaction. ANIMACT differs in what the references are required to do: each story conditions on up to four character and prop references, and evaluates whether the correct entities preserve their appearance while taking the correct actions, bindings, and interactions throughout the story.

A separate line of work evaluates consistency across multi-shot or multi-scene narratives. MovieBench Wu et al. (2025b) provides hierarchical movie-, scene-, and shot-level data for long-video generation. ViStoryBench Zhuang et al. (2026) studies story visualization over generated image sequences, emphasizing character consistency and narrative coherence. Recently, MSVBench Shi et al. (2026) extends this setting to multi-shot video using hierarchical scripts, reference images, and cross-shot evaluation. A concurrent preprint, EntityBench He et al. (2026), focuses on long-range consistency of characters, objects, and locations across generated shots. ANIMACT instead evaluates a single short clip in which multiple referenced entities must remain correctly bound while executing the intended actions and interactions. This enables joint evaluation of reference fidelity, action assignment, contact, motion, and physical and temporal consistency.

5.2 Human-centric Evaluation for AI Video Generation

Recent work increasingly automates video evaluation using learned metrics or multimodal judges. VideoScore learns an evaluator from human ratings He et al. (2024), while Video-Bench, StoryEval, AnimationBench, and MSVBench use MLLM/VLM-based evaluation Han et al. (2025); Wang et al. (2025); Wu et al. (2026); Shi et al. (2026). Several report strong aggregate agreement with human judgments, but this agreement can weaken on harder cases: StoryEval relies on repeated judging and unanimous voting for its strongest alignment, with performance dropping on its harder subset Wang et al. (2025). More broadly, JudgeAnything Pu et al. (2025) finds MLLM judging to be roughly 20–30% less reliable on multimodal generation than on multimodal understanding. This gap matters for ANIMACT, where evaluation requires fine-grained, reference-dependent decisions; for example, whether the correct look-alike performs an action, whether the same character returns after occlusion, or whether a specified contact occurs. Such errors can be hidden by aggregate correlation. We therefore use human judgments as benchmark labels and evaluate MLLMs separately as judges, measuring both aggregate agreement and reliability on individual rubric decisions.

6 Conclusion

We introduced **ANIMACT**, a human-evaluated capability framework for multi-reference animated video generation, designed to test whether referenced characters and props can preserve their identities while carrying out the intended story. Across current systems, character-to-character interaction remains a persistent weakness, increasing reference load strains appearance consistency more than entity presence, and failures accumulate across capabilities even when individual criteria appear strong. We further find that AI judges can support relative model comparison, but cannot yet replace fine-grained human evaluation. More importantly, ANIMACT provides a capability profile rather than only a leaderboard: its overall criteria and targeted probes expose where a system breaks and which production demands remain unreliable. This can serve as a practical guide for model developers and practitioners to identify which axes require better data, targeted post-training, or generation-time intervention, and to select models according to the capabilities required by a particular story. The benchmark itself provides one instantiation of this framework, while its broader utility lies in enabling systematic diagnosis and improvement of reference-grounded animated video generation.

AI USE STATEMENT

We used generative AI only for specific data-generation, artifact-generation, and minimal writing-assistance. Benchmark story prompts and related itemizations were generated from author-written templates and a fixed asset inventory, as described in Appendix A. Character, prop, and environment reference images were generated using the image-generation model specified as part of our benchmark construction methodology, and all evaluated clips were produced by the video generation models under study. Generated prompts and reference images were manually inspected against the source assets, and defective items were regenerated or removed before inclusion in the benchmark. For manuscript preparation, we used large language models only to rephrase author-written text, reduce redundancy, improve clarity and flow, and condense prose. They were not used to formulate research questions, develop the capability taxonomy or evaluation rubric, design experiments, conduct literature review, produce human annotations, interpret results, or determine scientific claims. These components were developed and validated by the authors. Separately, three AI systems were evaluated as experimental conditions in our AI-as-judge analysis; their outputs constitute experimental results rather than authoring assistance. All human evaluation labels were produced by trained annotators under our quality-control process. We take full responsibility for the final manuscript and for all benchmark assets produced with generative AI assistance.

ETHICS STATEMENT

Our data corpus uses synthetic characters, props, and generated videos; the reference assets are not based on real individuals and are not intended to depict identifiable people (detailed in section A.2). Human annotators were recruited through a third-party annotation provider and compensated for their work; further details on the annotation process and compensation are provided in the section A.7. We did not collect any personal or demographic information from annotators. Annotators participated solely as evaluators rather than creative contributors or professional animators, and were asked to make structured judgments using our predefined rubric. To reduce the influence of individual subjective judgments, our pipeline includes explicit quality-control and review stages during both data verification and video evaluation, and the rubric comprises observable, factual questions rather than holistic aesthetic preference (definitions in Table 5 and 6). We also perform reliability checks on a subset of annotations through independent review in section A.7. We discuss the limitations and scope of our evaluation study in Section D. The released materials contain no personally identifying information about annotators.

REPRODUCIBILITY STATEMENT

We document the construction and evaluation of ANIMACT in detail to support replication of both the benchmark protocol and the reported analyses. Section 2 summarizes the data-construction pipeline, while Appendix A provides its full implementation details, including the schemas for story cards and reference assets, the evaluation-only fields supplied to annotators, and the verification procedures used throughout corpus construction. We report the exact video-generation endpoints, generation settings, and prompt templates used for all studied models (see section A.6), as well as the prompts and configurations used for the AI judges see section A.10). Because several evaluated systems are proprietary and non-deterministic, exact regeneration of individual clips cannot always be guaranteed; we therefore disclose the generation protocol needed to reproduce the experiments.

For the evaluation part, we provide the complete rubric, including every overall and capability-specific question presented to human and AI judges in Table 5 and Table 6, together with the response formats and scoring procedure. We also document the quality-control, verification, and review stages used throughout data construction (section A.1) and human evaluation (section A.5), including additional review used to assess annotation reliability (detailed in section A.7). The supplementary ZIP contains four representative story cards, their character reference images, and the corresponding generations from the six studied models, allowing readers to inspect the evaluation inputs and outputs directly.

REFERENCES

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fce2d8a485746f76aac7b5650db2679d-Paper-Conference.pdf.

ByteDance Seed. One-take creation, flexible referencing: Introducing seedance 2.5. https://seed.bytedance.com/en/blog/one-take-creation-flexible-referencing-introducing-seedance-2-5, July 2026. Published July 31, 2026.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320, June 2024.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. Magref: Masked guidance for any-reference video generation with subject disentanglement. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 1–32, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0021c2cb1b9b6a71ac478ea52a93b25a-Paper-Conference.pdf.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning, 2024. URL https://arxiv.org/abs/2311.10709.

Google DeepMind. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026. Published February 19, 2026.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, and Yongxin Ni. Video-bench: Human-aligned video generation benchmark. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18858–18868, 2025. doi: 10.1109/CVPR52734.2025.01757.

Ruozhen He, Wei Meng, Ziyan Yang, and Vicente Ordonez. Entitybench: Towards entity-consistent long-range multi-shot video generation. *Preprint*, 2026.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.127. URL https://aclanthology.org/2024.emnlp-main.127/.

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=f3zNgKga_ep.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, June 2024.

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

Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Ying, Jing Liu, Chao Xu, Siqi Wang, Yidi Wu, Bingwen Zhu, Yue Zhang, Jinlong Hou, and Huyang Sun. Exploring the frontiers of animation video generation in the sora era: Method, dataset and benchmark. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 1260–1268. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/141. URL https://doi.org/10.24963/ijcai.2025/141. Main Track.

Tingting Liao, Chongjian Ge, Guangyi Liu, Hao Li, and Yi Zhou. Character mixing for video generation, 2025. URL https://arxiv.org/abs/2510.05093.

Xinran Ling, Chen Zhu, Meiqi Wu, Hangyu Li, Xiaokun Feng, Cundian Yang, Aiming Hao, Jiashu Zhu, Jiahong Wu, and Xiangxiang Chu. Vmbench: A benchmark for perception-aligned video motion generation, 2025. URL https://arxiv.org/abs/2503.10076.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149, June 2024a.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models, 2024b.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dIjMswSzgF.

Xiangyu Meng, Zixian Zhang, Zhenghao Zhang, Junchao Liao, Long Qin, and Weizhi Wang. Identity-GRPO: Optimizing multi-human identity-preserving video generation via reinforcement learning. In *ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling*, 2026. URL https://openreview.net/forum?id=MfFAEJjZmm.

MiniMax. MiniMax-H3. https://huggingface.co/MiniMaxAI/MiniMax-H3, 2026. Open-weight model release.

OpenAI. GPT-5.5 System Card. https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf, April 2026. Published April 23, 2026.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, and Philip S. Yu. Judge anything: Mllm as a judge across any modality. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 5742–5753, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737409. URL https://doi.org/10.1145/3711896.3737409.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang,

Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

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. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24034–24058, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1203. URL https://aclanthology.org/2026.findings-acl.1203/.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416, June 2025.

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

F Thomas and Ollie Johnston. The illusion of life : Disney animation. 1981. URL https://api.semanticscholar.org/CorpusID:191253497.

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

Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13629–13638, June 2025.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, Linus, Patrol, Peizhen Zhang, Peng Chen, Penghao Zhao, Qi Tian, Songtao Liu, Weijie Kong, Weiyan Wang, Xiao He, Xin Li, Xinchi Deng, Xuefei Zhe, Yang Li, Yanxin Long, Yuanbo Peng, Yue Wu, Yuhong Liu, Zhenyu Wang, Zuozhuo Dai, Bo Peng, Coopers Li, Gu Gong, Guojian Xiao, Jiahe Tian, Jiaxin Lin, Jie Liu, Jihong Zhang, Jiesong Lian, Kaihang Pan, Lei Wang, Lin Niu, Mingtao Chen, Mingyang Chen, Mingzhe Zheng, Miles Yang, Qiangqiang Hu, Qi Yang, Qiuyong Xiao, Runzhou Wu, Ryan Xu, Rui Yuan, Shanshan Sang, Shisheng Huang, Siruis Gong, Shuo Huang, Weiting Guo, Xiang Yuan, Xiaojia Chen, Xiawei Hu, Wenzhi Sun, Xiele Wu, Xianshun Ren, Xiaoyan Yuan, Xiaoyue Mi, Yepeng Zhang, Yifu Sun, Yiting Lu, Yitong Li, You Huang, Yu Tang, Yixuan Li, Yuhang Deng, Yuan Zhou, Zhichao Hu, Zhiguang Liu, Zhihe Yang, Zilin Yang, Zhenzhi Lu, Zixiang Zhou, and Zhao Zhong. Hunyuanvideo 1.5 technical report, 2025a. URL https://arxiv.org/abs/2511.18870.

Leyi Wu, Pengjun Fang, Kai Sun, Yazhou Xing, Yinwei Wu, Songsong Wang, Ziqi Huang, Dan Zhou, Yingqing He, Ying-Cong Chen, and Qifeng Chen. Animationbench: Are video models good at character-centric animation?, 2026. URL https://arxiv.org/abs/2604.15299.

Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie level dataset for long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28984–28994, June 2025b.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Chongyang Ma, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=XKhLsRPMsw.

Diandian Zhang, Tingyu Song, Lin Fu, Zheyuan Yang, and Yilun Zhao. Sci-VBench: Evaluating knowledge- and reasoning- intensive video generation in science domains. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=lyz7edDsY1.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, 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.

Bingwen Zhu, Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Yin, Yidi Wu, Huyang Sun, and Zuxuan Wu. Aligning anime video generation with human feedback, 2025. URL https://arxiv.org/abs/2504.10044.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, Xuanyang Zhang, Xianfang Zeng, Zhewei Huang, Gang Yu, and Chi Zhang. Vistorybench: Comprehensive benchmark suite for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9455–9467, June 2026.

A Additional Details on the ANIMACT Corpus and Evaluation

This appendix provides additional technical details for Section 2. We describe the corpus curation and filtering here, which includes reference-asset construction and verification, capability-targeted

story construction, story-card schema, evaluation protocol, video-generation configuration, and human annotation pool and the review procedure. The complete question-level definitions of the M (capability modes) and OR (overall rubric) families are provided in Tables 5 and 6 respectively.

A.1 Corpus Construction

The final ANIMACT corpus contains 317 story cards, yielding 1,902 evaluated clips across the six studied video-generation models. The supplementary ZIP includes four representative story cards, the corresponding character reference images, and generated clips from all six models.

The reference bank contains 64 characters and 13 props in the final corpus. Stories contain one to four characters and zero to two props, with a maximum of four reference images per task so that every story can be submitted to the most restrictive generation endpoint. Approximately 95% of stories contain multiple characters. Stories are silent by design and therefore contain no required dialogue or speech-generation component (see Table 2).

Quantity	Value
Final complete stories	317
Video-generation systems	6
Final evaluated clips	1,902
Characters in reference bank / used	64 / 64
Props in reference bank / used	13 / 13
Appearance variants	19
Capability modes (M family)	8
Targeted M questions	27
Overall OR criteria	14
Rubric questions	41
Scored items per clip	30–65 (mean 52)
Characters per story	1–4
Props per story	0–2
References per story	1–4
Multi-character stories	95%
Stories requiring dialogue	0

Table 2: ANIMACT corpus statistics. The final analysis uses only stories with outputs from all six evaluated video-generation models.

A.2 Reference Asset Bank and Verification

Canonical character specifications. We construct each character around a canonical card before generating its reference images. Gemini 3.1-Pro Google DeepMind (2026) is used to draft the initial card, which is then manually authored. The card records the character's visible appearance, costume, distinguishing features, accessories, and a `must_not_change` list containing details that should remain stable whenever that character is rendered. These attributes subsequently serve two purposes: they condition reference-image generation and provide reference-specific visual criteria for downstream human evaluation. Props are represented separately through visual descriptions rather than character-style canonical cards. This avoids assigning character-specific fields to objects while still providing a fixed visual specification for prop generation and verification.

Reference-image generation. Character reference images are generated from their canonical-card specifications using GPT-Image-2[5]. Prop references are generated from their corresponding visual descriptions. Character sheets are rendered at 1536×1024 and contain multiple views of the same identity, including full-body front, three-quarter, and side views together with an expression row. Generation uses a common visual domain so that the principal variation across the bank is the identity, costume, and character-specific attributes rather than an uncontrolled change in render style. The `must_not_change` attributes attached to each accepted character are carried into every story

[5] `https://developers.openai.com/api/docs/models/gpt-image-2`

that references that character. They are never supplied to the video-generation system as evaluation instructions; instead, they provide ground truth for evaluating whether reference-specific visual details survive in the generated clip.

Story settings. The asset bank contains 64 original character designs. Eighteen are principals: twelve in traditional attire drawn from twelve cultural settings: Andean (Quechua), Balinese, Japanese, Korean, Kuban Cossack, Maasai, Mexican, Palestinian, Punjabi, S'ami, Scottish Highland, and Yoruba, and six in contemporary dress. Nine of the traditional settings additionally carry a within-setting supporting cast of 17 characters in total, so that a story can be populated entirely from one setting. The remaining designs support the identity-discrimination probes: 14 counterfactual twins, each differing from its base character in exactly one named attribute, and 12 look-alikes, separate individuals designed to be confusable with a principal. Settings broaden the visual diversity of characters, garments, and accessories; we treat them as descriptive composition rather than as evaluation strata and do not compare performance across settings. We want to emphasize that all characters are original fictional designs rather than depictions of real individuals.

Human verification and repair. Generated assets are verified before they are frozen for story construction. For characters, annotators compare each rendered reference against its canonical-card descriptors and separately assess whether the identity remains consistent across the rendered views. The latter check covers clothing, accessories, body details, colours, and other identity-defining attributes. Prop references are similarly checked against their visual specifications. Annotators are also given an *optional* field to flag any visible glitch in the generated reference sheet such as baked-in backgrounds or clipped body parts, and are explicitly asked to identify inconsistencies across different views of the same character. Flagged images are revised from the annotator feedback and re-checked before they can enter the frozen reference bank.

For the 70 assets included in the detailed anchor verification study, annotators judged between 94.8% and 96.9% of the specified visual details to be present in the rendered references, depending on the annotator. Cross-view consistency was substantially more difficult: 59.1% of assets were flagged by a majority of annotators and 63.7% were flagged by at least one annotator. In response to these checks, 46 assets were revised and re-checked before their accepted versions were frozen.

Each anchor asset is independently evaluated by three annotators. We report mean pairwise agreement as the average proportion of matching labels across the three annotator pairs. Mean pairwise agreement is 96.1% for the presence of canonical-card visual details and 78.3% for cross-view consistency judgments. We treat these measurements as quality-control checks on the benchmark inputs rather than as benchmark results: their purpose is to ensure that an apparent identity failure in a generated clip can reasonably be attributed to the video model rather than to an ambiguous or internally inconsistent reference. Assets revised following this audit are manually re-verified by the authors before inclusion in the benchmark.

A.3 Capability-Targeted Story Construction

Story structure. Each story is constructed against a fixed asset bank of human-verified character and prop references. A story contains one to four characters (median three), zero to two props, and at most four reference images. Conditioned on these references and an initial free-form authoring prompt, GPT-5.5 OpenAI (2026) produces an ordered beat-level story that is rendered as a single continuous narrative for video generation. Characters are referred to using natural visible descriptors rather than bank identifiers or screen positions (e.g., "a boy in a raincoat" or "an old man wearing a hat"). Appearance already conveyed by the reference is generally not restated, except when a distinguishing trait is needed to disambiguate look-alike characters. Stories are designed to fit within a single 10-second generation and require no spoken dialogue.

Evaluation-only metadata. Alongside the narrative, each story stores structured metadata used only for evaluation and never shown to the video model. This metadata captures the target capability, reference-specific identity constraints, required story events, and, where applicable, ordered interactions. The individual fields and their visibility are detailed in Section A.4.

Story-card verification. Before generation, three annotators independently review each story card using 14 checks covering capability alignment, temporal and interaction consistency, 10-second feasibility, prompt quality, agreement with the reference images, and, for look-alike or twin-focused tasks, whether the characters are confusable yet distinguishable by the intended visual cues. Reviewers also verify that the prompt includes the listed cast and props, remains physically and causally consistent, contains no required dialogue, and can be realized as a continuous 3–4-beat story within a 10-second clip. Our card-check study covers 317 cards with three independent annotations per card. Eighty-nine cards were flagged for redesign, including 86 with prompt–asset contradictions in which a character was required to manipulate a story prop despite already holding a personal prop in the frozen reference. Flagged cards were revised using GPT-5.5, and human-verified before video generation.

Inter-annotator agreement in story-card verification. For the story-card checks, we report pairwise percent agreement: for each question, we compute the proportion of cards on which each pair of reviewers assigns the same label, then average it across the three reviewer pairs and across questions, weighted by the number of cards to which each question applies. Across all 14 checks, the mean IAA is 0.936 and Gwet's AC1 is 0.911.

A.4 Story-Card Representation

Every story is stored as a structured record that separates generator-visible information from evaluation-only metadata. Table 2 summarizes the main fields and their roles. Identity fields specify the target character, secondary characters, props, and frozen reference locations. Capability fields identify the target M mode and any capability-specific structure, such as the interaction graph. Prompt fields contain both the internal beat-level representation and the continuous narrative submitted to the generator. Evaluation fields contain the key actions and reference-specific visual constraints. Presentation fields record the visual style, setting, and ordering of the supplied references, while provenance fields record how generated text fields were created or revised.

Field	Role	Visibility
`task_id`	Unique story identifier	bookkeeping
`target_character_id`	Character primarily associated with the target capability	internal
`secondary_character_ids`	Remaining referenced cast	internal
`cast`	Character records and frozen reference paths	reference images passed to model
`props`	Story props and their frozen references	reference images passed to model
`modes.major`	Target M capability mode	evaluation only
`story_prompt`	Continuous narrative prompt	generator input
`key_actions`	Itemized events required for grading	evaluation only
`must_not_change`	Reference-specific identity constraints	evaluation only
`interaction_graph`	Ordered entity–entity contact edges where applicable	evaluation only
`setting`	Intended scene/environment	represented in prompt
`style`	Common visual-domain instruction	generator input
`stitch_layout`	Stable ordering of supplied references	generation configuration

Table 3: Main fields in an ANIMACT story card. Generator-visible and evaluation-only information are stored separately.

The distinction between the continuous prompt and the key-action representation is deliberate. A narrative may describe a sequence of interactions in ordinary prose, while the hidden metadata decomposes the same sequence into independently gradeable actions and contact events. Exposing this representation to the generator would change the task from story realization toward explicit step-by-step instruction following.

A.5 Evaluation Protocol

The complete M and OR question glossaries are provided in Tables 8 and 9. ANIMACT includes 27 M questions across eight targeted capability modes and 14 OR criteria applied to every clip. M questions are evaluated only for the capability mode used to construct a story, whereas OR criteria capture general failures such as reference fidelity, action execution, temporal and spatial stability, and physical coherence.

These 41 rubric questions expand into story-specific annotation items. Questions involving repeated entities or events are instantiated over the relevant characters, props, scripted actions, or interac-

tion edges using the story-card metadata. Each clip therefore contains 30–65 scored items (52 on average), with evaluators seeing only questions applicable to that story.

Items use binary, *Yes/Partial/No*, or checklist responses. Checklist items are evaluated separately for each applicable character, prop, action, or interaction, while inapplicable items are omitted (e.g., prop-dependent questions for stories without props).

The evidence provided depends on the criterion: identity and fine-detail checks use the reference images and sampled video frames, while motion, interaction, action coverage, pacing, and physical behavior are judged from the full clip. Response polarity is normalized before aggregation so that higher scores consistently indicate better performance.

A.6 Reference-to-Video Generation Protocol

Common generation settings. We generate one clip per story using Wan 2.7, Kling O3, Grok Imagine, MiniMax-H3, Happy Horse 1.1, and Seedance 2.5. Every model receives the same frozen references and continuous narrative prompt. The story-card `style` field is also passed as generator input and is fixed across all models to "*Generate a stylized 3D CGI animated-film video*".

The target generation configuration is 10 s, 16:9, and 720p. Kling O3 is the exception: the reference-to-video endpoint used in our experiments fixes its output at 3840×2160. We do not otherwise tune resolution or duration separately for individual models. Seed 42 is passed whenever an endpoint exposes a seed parameter. Kling O3 and Grok Imagine do not expose a seed in the endpoint configuration used here and therefore cannot be reproduced from a user-specified random seed alone.

Model	Reference argument	Max refs	Seed	Additional arguments passed
Wan 2.7	`reference_image_urls`	8	yes	`multi_shots=True`, `negative_prompt=` ***no-dialogue***, `seed=42`
Kling O3	`image_urls`	4	no	none (resolution fixed by endpoint; `duration` as string)
Grok Imagine	`reference_image_urls`	7	no	none
MiniMax-H3	`reference_image_urls`	9	yes	`seed=42`, `enable_safety_checker=True`, `enable_prompt_expansion=False`
Happy Horse 1.1	`image_urls`	9	yes	`seed=42`
Seedance 2.5	`image_urls`	30	yes	`generate_audio=False`

Table 4: Reference-input constraints and generation arguments. Max refs denotes the endpoint's supported reference cap; our stories use at most four references. Every model receives the identical narrative prompt and the same `aspect_ratio=16:9` and `duration=10`; `resolution` is set to 720p wherever the endpoint exposes it (Kling O3 fixes resolution at the endpoint and MiniMax-H3 uses its own enum, where 720p maps to `768P`). Kling O3 and Happy Horse 1.1 additionally require reference images of at least 300 px and 400 px respectively. Seed is a property of the endpoint, not of our harness: Kling O3, Grok Imagine publish no seed parameter.

Model input. For each story, the continuous narrative prompt is passed together with the frozen reference images for its cast and props, in the ordering stored by the story card. The generator does not receive the scripted key actions, interaction graph, `must_not_change` attributes, capability label, or other evaluation metadata. Where an endpoint exposes automatic prompt expansion, we disable it so that the intended story is not rewritten differently across systems. Content safety filters remain at their published defaults. The API errors are recorded as missing rather than regenerated with altered prompts, references, or safety settings. Wan 2.7 produces audio unconditionally in the endpoint used here. Because ANIMACT stories are intentionally silent, Wan 2.7 additionally receives a fixed negative prompt suppressing speech and lip synchronization. This is treated as an endpoint-specific control needed to preserve the common silent-story protocol rather than as story-specific prompt engineering.

Reference format. References are supplied as individual asset images rather than as a single stitched contact sheet. This choice preserves one reference slot per entity and avoids aspect-ratio restrictions encountered by at least one evaluated endpoint when processing unusually wide stitched sheets. The maximum story load is four references, matching Kling O3, the most restrictive endpoint in the evaluated set.

A.7 Human Annotation and Review Protocol

Annotation unit and interface. Human evaluation is organized by story. For each story, annotators see the frozen reference images, continuous narrative prompt, evaluation-only visual constraints, scripted key actions, and all six generated clips, and score each clip using the applicable M questions and OR rubric. Grouping outputs by story keeps the references, narrative, and evaluation criteria fixed across models. Model identities are hidden and clip order is independently shuffled within each story.

Annotator training and workload. Before evaluation, each annotator completes a 16-clip calibration pilot covering all eight capability modes, with annotations checked against author-provided reference labels before proceeding. Each clip contains 52 scored items on average and typically requires 20–30 minutes to annotate; annotators are compensated at $20 per hour. We do not collect personal or demographic information from annotators. Annotators are trained specifically for our video-evaluation task and rubric and serve as task-trained evaluators, rather than professional animators or other creative-production practitioners.

Reliability study and full-scale protocol. We assess the reproducibility of rubric judgments using a separately triple-annotated subset of 176 clips spanning four models (excludes HappyHorse 1.1 and Kling-o3), all eight capability modes, and 4,202 scored items. The three independent annotators achieve 92.4% raw agreement and Gwet's AC1 of 0.917, corresponding to 7.6% raw disagreement; question-centered Spearman rank agreement is $\rho_c = 0.724$. These results show high item-level reproducibility, while some variability remains in finer-grained relative assessments.

At the scale of approximately 1,900 clips, obtaining three independent annotations per clip would require about 296,000 individual item judgments ($1{,}900 \times 52 \times 3$). Given the sufficient reliability observed in the subset and the annotation burden of triplicate evaluation at full scale, we use an annotate-then-review protocol for the complete evaluation. Labels are produced by 13 annotators and subsequently checked by 9 reviewers against the same story evidence, with disagreements or questionable panels escalated for resolution. This provides a second human quality-control stage for every clip without requiring three independent annotations for all 1,902 outputs.

A.8 Prompt Templates

A.9 Reference-to-Video Generation

No additional generation-time prompt template is used. For each sample, the generator receives the finalized `story_prompt` from its story card, together with the fixed `style` instruction ("*Generate a stylized 3D CGI animated-film video*") from the same card. These fields are passed unchanged, along with the corresponding frozen reference sheets in story order. This input protocol is fixed across all samples, videos, and six model arms, apart from endpoint-specific argument names and unavailable controls.

```
args = {
  "prompt":        {style}+{story_prompt},          # both from story card
  {refs_key}:      [{ref_sheet_url}, ...],      # frozen sheets, story order
  "duration":      10,                          # seconds
  "resolution":    "720p",
  "aspect_ratio":  "16:9",
  "seed":          42,                          # where endpoint exposes one
}
```

A.10 AI Judge for Video Evaluation

The AI judge receives the same reference images and generated video clip used in human evaluation, together with a story-specific prompt instantiated from a fixed template (following the rubric defined in Tables 5 and 6). Human annotators are shown the corresponding annotation form with the same grading instructions and response semantics; story-specific placeholders are instantiated from the relevant metadata for each story. The template passed to the AI judge mirrors this structure: it first defines the grading rules, then provides the generator-visible story and reference descriptions, followed by the story-specific evaluation questions expanded from the same metadata used for human

annotation. The judge is instructed to score only visible evidence in the clip and return a structured JSON response with a short rationale for each item. For a given story, the instantiated prompt is identical across model outputs and does not reveal the generating model.

```
AI-JUDGE PROMPT TEMPLATE

% You are grading one short generated video clip, exactly as a trained
% human reviewer would.

% ========================
% GRADING INSTRUCTIONS
% ========================

% The clip was generated from a short written story and a set of reference
% images. Record WHAT ACTUALLY APPEARS AND HAPPENS IN THE VIDEO, not what
% the story intended.

% Do not grade visual attractiveness or polish. A visually appealing clip
% can still be incorrect, and a rough-looking clip can still satisfy the
% story.

% Judge only visible evidence:
% - Never give credit because the story says something should happen.
% - If it cannot be seen, treat it as not having happened.
% - If an unintended error is visible, count it even if the story does
%   not mention it.

% Response options:
%   Yes      = clearly true throughout the relevant evidence.
%   Partial  = only partly true, inconsistent, or visually ambiguous.
%   No       = clearly not true.

% If a question provides only Yes/No, choose only from those options.
% For checklist questions, select every applicable failure. If nothing is
% wrong, select only the designated "nothing wrong" option. Never return
% an empty checklist.

% Important distinctions:

% 1. Action occurrence vs. visibility
%    "Did it happen?" asks whether the intended event occurs.
%    "Could you see it clearly?" asks whether the event is sufficiently
%    visible and unobstructed to verify.
%    If the action did not happen, return null for its visibility field.

% 2. Within-shot glitches vs. invented content
%    Popping, vanishing, or swapping refers to an expected entity
%    flickering, disappearing, reappearing, or morphing unexpectedly
%    within a shot.
%    Invented content refers to a new action, object, or event not
%    requested by the story.

% 3. Intended state changes
%    Do not penalize an appearance or object-state change explicitly
%    required by the story.

% 4. Shot boundaries
%    A cut is a transition to a new shot. Questions about continuity
%    across cuts are distinct from within-shot stability questions.

% ========================
% STORY
% ========================

% {story_prompt}

% ========================
% REFERENCE CHARACTERS
% ========================

% The reference images are attached in the same order as below.

% {character_1}: {visible_description_1}
% {character_2}: {visible_description_2}
% ...
% {character_n}: {visible_description_n}

% ========================
% EVALUATION QUESTIONS
% ========================

% The following questions are instantiated from this story's evaluation
% metadata. Answer every item shown.

% [CAST AND ENTITY PRESENCE]
% {cast_presence_questions}

% [IDENTITY AND APPEARANCE]
% {identity_and_appearance_questions}

% [BACKGROUND AND SPATIAL CONSISTENCY]
% {spatial_questions}
```

```
% [PROP STABILITY -- if applicable]
% {prop_questions}

% [STORY ACTIONS]
% For each required action:
%   {key_action_i}
%   - did_it_happen: Yes / No
%   - could_you_see_it: Yes / Partial / No
%     Return null if did_it_happen = No.

% [TARGET CAPABILITY]
% This story probes:
%   {capability_description}

% {capability_specific_questions}

% [MOTION AND PHYSICAL BEHAVIOR]
% {motion_physics_questions}

% [WITHIN-SHOT POPPING / VANISHING / SWAPPING]
% {within_shot_questions}

% [CONTINUITY ACROSS CUTS]
% {across_cut_questions}

% [UNREQUESTED CONTENT]
% {extra_content_questions}

% ========================
% OUTPUT FORMAT
% ========================

% Return ONLY one JSON object. Do not include commentary outside the JSON.

% Every option in the instantiated question schedule is numbered.
% Return the NUMBER of the selected option rather than repeating its text.

% Provide one short rationale (maximum 25 words) for every scored entry,
% stating the visible evidence that determined the answer.

% Use the following entry formats:

% Single-choice item:
% {"block": "<id>", "item": <n>, "choice": <option_number>,
%  "rationale": "<=25 words"}

% Multi-select item:
% {"block": "<id>", "item": <n>, "choices": [<option_numbers>],
%  "rationale": "<=25 words"}

% Per-object item:
% {"block": "prop_steady", "item": <n>, "object": "<object_name>",
%  "choices": [<option_numbers>], "rationale": "<=25 words"}

% Action item:
% {"block": "actions", "item": <n>,
%  "did_it_happen": <option_number>,
%  "could_you_see_it": <option_number_or_null>,
%  "rationale": "<=25 words"}

% Free-text item:
% {"block": "<id>", "item": <n>, "text": "<text or 'none'>"}

% Wrap all entries as:

% {"answers": [ ... ]}

% Return exactly {n_items} entries, one for every instantiated evaluation
% item.
```

B SUPPORTING EVIDENCE FOR MAIN FINDINGS

How well is the model ordering resolved? The mean-score ranking gives an ordering of the six models, but not every adjacent difference is statistically resolved. We bootstrap mean clip scores over the 317 stories and apply Holm correction across all 15 pairwise comparisons. Thirteen of the 15 pairs remain significant at $\alpha = 0.05$. Among the five adjacent pairs in the reported ranking, three separate reliably, while HappyHorse-1.1 vs. Wan-2.7 and Wan-2.7 vs. Kling-O3 do not (Table 7). Thus, the top two and bottom model are clearly separated from their immediate neighbors, whereas the middle of the ranking is only partially resolved: Wan-2.7 cannot be distinguished from either adjacent model on this corpus.

Separating reference load from prompt length. Reference load and prompt length co-vary, so we ask whether the reference fidelity decline in the section 3.3 is explained by longer prompts rather

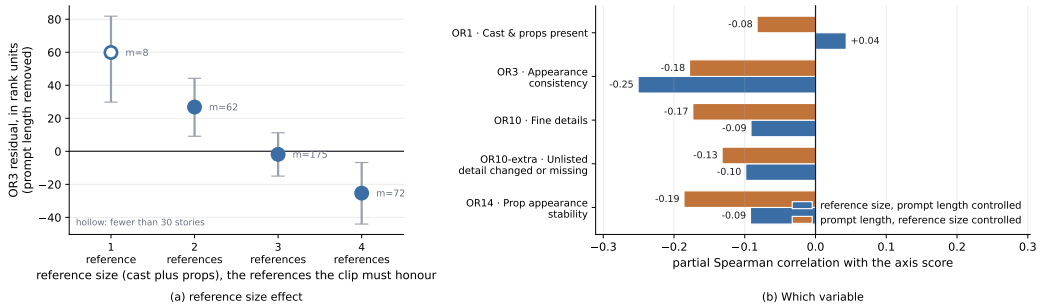

Figure 8: **Separating reference size from prompt length.** Reference size (cast plus props) and prompt length are moderately correlated across the 317 stories (Spearman $\rho = 0.474$), appreciably more than the 0.264 of a characters-only definition, so residualising one out of the other matters here. (a) Appearance consistency (OR3) by reference size after removing the rank effect of prompt length; 0 is the corpus average at that prompt length. Error bars show 95% bootstrap intervals over stories, and hollow markers denote groups with fewer than 30 stories. (b) Partial Spearman correlations ρ of each fidelity axis with reference size (blue) and prompt length (orange), each controlling for the other. OR14 is evaluated only on stories containing props.

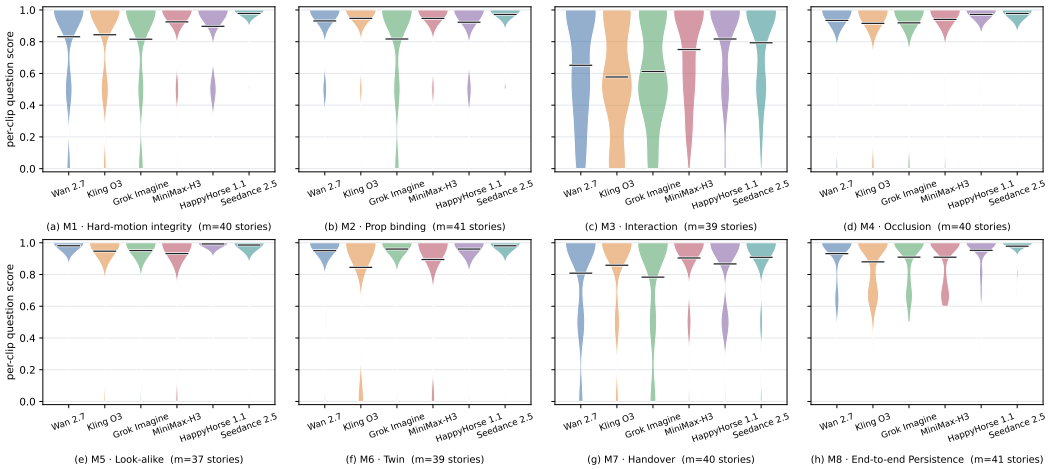

Figure 9: Per-clip score distributions for each probe family and model; ticks mark the equal-weight family means. Story counts (m) are shown separately in the sub-caption of each plot.

than by the number of references itself. Figure 8 separates the two effects. After removing the rank effect of prompt length, appearance consistency (OR3) still declines steadily as reference load increases (Fig. 8a); the one-reference estimate is less certain because it contains only eight stories. The partial correlations show the same pattern (Fig. 8b): OR3 retains the strongest independent association with reference load ($\rho = -0.25$), while cast and prop presence remains essentially flat ($\rho = +0.04$). In contrast, fine details, unlisted details, and prop appearance have weaker associations with reference load ($\rho = -0.09$, -0.10, and -0.09) than with prompt length ($\rho = -0.17$, -0.13, and -0.19). Thus, the drop in character appearance consistency cannot be explained by prompt length alone, whereas the other fidelity declines are more closely tied to how much the prompt specifies.

C Further Analysis of AI Video Judges

We further examine the distinction established in Section 4: *AI judges recover aggregate differences between video-generation models substantially better than the fine-grained clip-level judgments*,

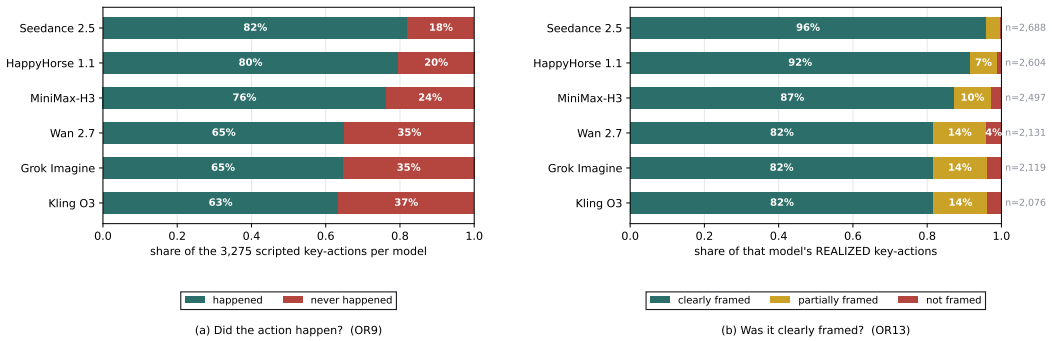

Figure 10: **Action coverage and staging by model.** (a) OR9 measures the share (percentage) of scripted key actions that occur. (b) OR13 evaluates framing only for actions that occurred, classifying them as clearly staged, partially staged, or not staged; therefore, the two panels use different denominators. Bars are normalized within each model.

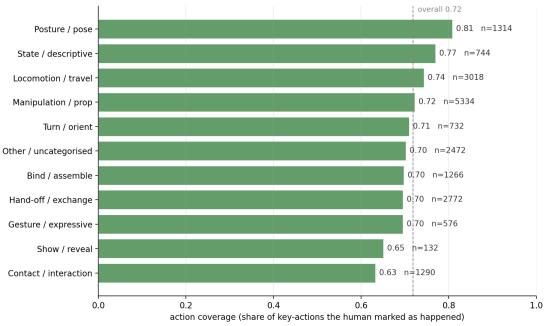

Figure 11: **Action coverage by action type.** Here, n denotes the number of scripted actions belonging to each action type. Contact, directly probed in M3, also has the lowest action coverage under OR9.

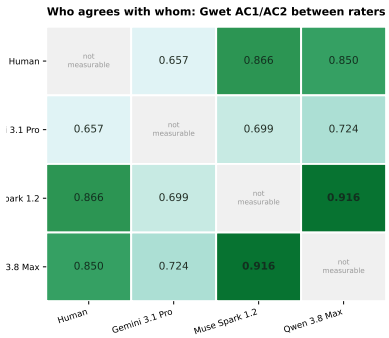

Figure 12: **Pairwise judge agreement.** Gwet's AC1/AC2 measures how often two judges assign the same rubric label to the same clip, correcting for agreement expected by chance. Higher values indicate more consistent clip-level judgments.

underlying those differences. Unless otherwise noted, these analyses use the 1,876 clips from 314 stories for which both human and AI-judge labels are available.

Aggregate model comparison. We first ask whether an AI judge preserves the relative performance of the six video-generation models. Qwen 3.8 Max obtains a leaderboard Spearman correlation of $\rho = 0.886$ with human evaluation and reproduces 78.3% of question-level human model orderings (95% CI: 0.74–0.82). Its question-centred rank correlation is $\rho_c = 0.693$, and it retains $0.87\times$ of the score separation observed under human evaluation (Table 10). Among the individual judges, Qwen therefore provides the strongest overall agreement with human relative model comparisons. Majority-of-3 increases ordering concordance only slightly, from 0.783 to 0.796, while requiring predictions from all three judges and retaining less of the human model separation ($0.66\times$). We therefore use Qwen when an automated judge is needed for relative model comparison.

Aggregation masks weak clip-level agreement. Preserving model rankings does not imply reproducing the clip-level judgments from which those rankings are constructed. Table 11 makes this distinction explicit by computing human–judge rank correlation at two aggregation levels. For Qwen, the median correlation across rubric families is 0.829 when computed over the six model-level means, but only 0.238 when computed over individual clips. The gap appears across many capabilities. For example, Qwen reaches model-level correlations of 0.943 for Motion smoothness (OR2), Appearance consistency (OR3), and Action coverage (OR9), while the corresponding clip-

level correlations are only 0.169, 0.173, and 0.499. Interaction (M3) similarly rises from 0.466 over clips to 0.886 over model means. Aggregation therefore recovers a stable comparative signal even when agreement on individual generations is much weaker. Consistent with this gap, the AI judges can agree more strongly with one another than with humans: Qwen and Muse reach Gwet agreement of 0.916, exceeding Qwen–human (0.850) and Muse–human (0.866) agreement (Fig. 12).

Diagnostic failures are particularly difficult. We further isolate the clip-level cases that depart from the dominant outcome for a question. These minority-label items are especially relevant for failure diagnosis because they correspond to clips on which the human judgment differs from what would be obtained by simply predicting the question's majority label. Qwen recovers only 12.8% of these cases at 0.300 precision. Gemini recovers more (33.0%), but at substantially lower precision (0.173); Muse lies between them at 16.9% recall and 0.259 precision. Majority voting does not resolve the problem, reaching 17.1% recall at 0.330 precision. Thus, no evaluated judge provides both high recall and high precision on the atypical clip-level outcomes that matter for fine-grained diagnosis.

Implication for ANIMACT. These results reveal an aggregate–diagnostic gap. AI judges can preserve enough signal to compare video-generation models and recover many question-level model orderings, while remaining substantially less reliable on the individual clip decisions that explain *why* a model succeeds or fails. We therefore use Qwen only for relative automated model comparison. Human annotations remain the basis for absolute ANIMACT scores, capability-level evaluation, and fine-grained failure analysis.

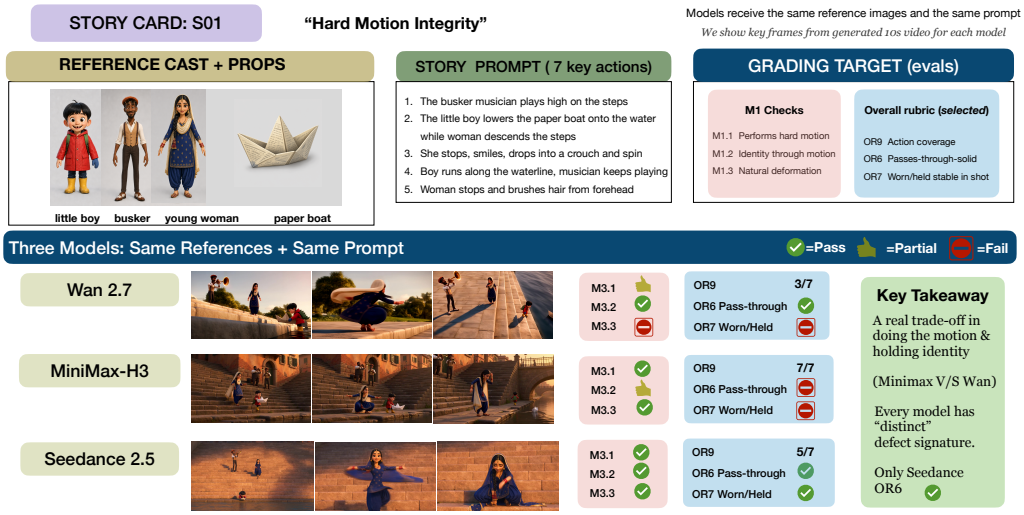

Figure 13: **Visual Case Study of "Hard Motion Integrity" story card.** Results shown for three models.

D Limitations and Future Scope

ANIMACT keeps generation settings as consistent across models as their public interfaces allow, although endpoint capabilities are not identical. In particular, Kling O3 and Grok Imagine do not expose a user-specified seed in the configurations used here. Because we generate one clip per story–model pair, our results characterize performance under a standardized single generation rather than the full output distribution of each model. Repeated generations would enable estimation of per-model failure probabilities, but would substantially increase both generation and human-evaluation cost. Our evaluation also focuses on approximately 10 s clips, a common visual domain, a fixed aspect ratio, and no required spoken dialogue. We chose 10 s because shorter settings often leave insufficient time to realize three ordered story beats and their interactions. Extending the evaluation

to shorter and longer durations, dialogue-bearing content, and longer multi-shot narratives would broaden the operating regimes covered by the benchmark.

ANIMACT's controlled asset construction enables systematic coverage of difficult cases such as look-alike casts, identity confusions, handovers, and rare interactions that are difficult to curate at scale. At the same time, parts of the corpus are model-assisted: reference images are generated with an image-generation model and story-card text is generated from author-specified constraints. Although all generated assets are manually verified before use, some visual, representational, or linguistic regularities of these models may remain in the benchmark distribution. Future work can broaden the asset pool with independently sourced, artist-created, multi-model-generated, or appropriately licensed production characters and environments, while retaining the same controlled capability design.

A small number of criteria retain some annotation subjectivity despite the structured rubric. Shot continuity is one example: annotators currently infer shot boundaries from the generated video and then judge whether characters, props, and surroundings remain consistent across the transition. Providing verified shot boundaries, or localizing failures directly to frames, intervals, contacts, or state transitions, could further reduce annotation noise. More generally, automated evaluation may benefit from specialized verifiers for identity, action completion, contact events, state changes, and continuity rather than relying on a single general-purpose VLM judge.

Beyond evaluation, ANIMACT provides a structured basis for improving multi-reference video generation. Richer beat-level or hierarchical conditioning may improve temporal execution, but because current systems differ in whether they natively support such inputs, this setting should be evaluated separately from the matched-prompt comparison used here. The same taxonomy extends naturally to multi-shot generation by evaluating identity, action, interaction, and state consistency both within and across shot boundaries. A complementary direction is to train a model-agnostic prompt rewriter that translates the same story into generation-ready instructions emphasizing temporal decomposition, actor–action binding, interaction, and persistence. Finally, ANIMACT's structured failure annotations can support targeted SFT, preference optimization, rejection sampling, and reinforcement learning with capability-specific verifiable rewards on trainable public models, using clean synthetic examples and specialized verifiers to connect diagnosis directly to model improvement.

Code	Nickname	Question as shown to the annotator
M1.1	Performs hard motion	Does the target actually perform the difficult whole-body motion (not just stand/walk)?
M1.2	Identity through motion	Is the target clearly the same character before, during AND after the motion?
M1.3	Natural deformation	During the motion do the limbs/body deform naturally (no rubber-hose/broken/wrongly-bent parts)?
M1.4	Others stay separate	Do the other characters keep doing their own thing without any merging with the target?
M2.1	Prop used in action	Is the prop actually used in the action (not ignored/absent)?
M2.2	Prop matches reference	Does the prop look like its reference the whole time (no morph/colour/shape/material change)?
M2.3	Prop use sensible	Is the way the prop is used physically sensible for that body and object?
M3.1	Interaction happened	(per edge) Did this interaction actually happen?
M3.2	Direction correct	(per edge) Is the direction right (correct do-er and receiver)? *Asked only when M3.1 records that the contact happened, and carries no independent score once M3.1 is answered, so it is counted in the question total but has no row in Table 1. Asked only when the contact happened, and it carries no independent score, so it has no row in the per-question table.*
M3.3	Contact-rich, not idle	Were the interactions genuinely contact-rich (touching/exchanging), not idle proximity?
M4.1	Visible part correct	While partly hidden, does the visible part look right (correct parts, clean edge)?
M4.2	Same character returns	After being hidden, does the SAME character come back out?
M4.3	Depth correct	Is depth handled right (scale with distance; front-vs-behind consistent)?
M4.4	Bindings survive occlusion	Across the occlusion, does the target keep the same prop, emotion and behaviour?
M5.1	Both present & active	Are BOTH the target and the look-alike visible and each doing something?
M5.2	Target does the action	Does the TARGET (not the look-alike) perform the key interaction with the partner?
M5.3	Look-alike never interacts	Does the look-alike NEVER interact with the target directly?
M5.4	Never swapped	Throughout, do target and look-alike stay tellable-apart (never swapped)?
M6.1	Both twins present & active	Are BOTH near-identical twins visible and each doing something?
M6.2	Distinguishing detail kept	Does each twin keep its own distinguishing detail (difference never leaks/merges)?
M6.3	Correct twin acts	Does the correct TARGET twin (not the counterfactual) perform the key interaction?
M6.4	Twin never interacts	Does the counterfactual twin NEVER interact with the target directly?
M6.5	Exactly two (no duplicate)	Are there exactly two distinct people (no duplicated third copy)?
M7.1	Contact physically real	Does the contact/handover look physically real (bodies turn to meet, actual touch)?
M7.2	Weight reads right	Does the weight read correctly (heavy item looks heavy)?
M7.3	Contact clearly staged	Is the contact clearly staged/visible (not hidden/cramped/ambiguous)?
M8.1	Identity holds first-to-last	From first shot to last, does each character stay the same recognizable character? (tick each cast member)
M8.2	Prop lifecycle complete	Does the prop go through its whole arc (handed→used→fate) without dropping from the plot?

Table 5: Rubric question glossary, probe families M1 to M8. Every coded name used in this paper is mapped to the exact wording the annotator was shown. Codes are numbered densely within each rubric, so a gap never implies a missing question. A tick-list parent's question is shown once in italics above its options, and each option is scored as its own binary. Conditional questions state in italics what they are conditional on. Shading groups the sub-questions of one rubric and carries no other meaning.

Code	Nickname	Question as shown to the annotator
OR1	Cast & props present	Tick each expected character & prop: present AND correct entity, prop on right character.
OR10	Fine details	Check each must-not-change small detail vs the sheet (correct Yes/No). Naming an unlisted change is OR10 · extra-changed; naming one that never appeared is OR10 · extra-missing.
OR10 · extra-changed-or-missing	Unlisted detail changed or missing	Derived binary. The two OR10 free-text fields, which record a listed detail that changed and an unlisted detail the annotator noticed, are merged into one yes/no: did any detail outside the listed set change or go missing?
OR11	Physics & deformation	Is body deformation/weight believable (joints, cloth lag, no rubber limbs/weightless heavy)?
OR12	Pacing	Is pacing right (progresses vs dwells vs too fast)?
OR14: Per prop: does the object change on its own — shape, colour, material, coming loose from the hand, or morphing into something else? (tick each; ignore change the STORY causes — crumpled, torn, soaked, broken) — PROPS only; matching the reference image is M2.2 Each option below is scored as its own binary.		
OR14 · colour-changes		"Colour changes"
OR14 · comes-loose-floats		"Comes loose — floats free of the hand holding it"
OR14 · material-changes		"Material changes"
OR14 · morphs-into-else		"Morphs into something else"
OR14 · shape-changes		"Shape changes"
OR2	Motion smoothness	How smooth/coherent is the motion (no judder/teleport/stop-start/jumps)?
OR3: Across the cast: does anything about how the characters look change, break up or go missing — colours, surfaces, shape, or a part gained/lost? (tick each kind you see — ONE tick-set for the clip, not per character) — CHARACTERS only, APPEARANCE only; 'same person?' is M8.1, and the NAMED must-not-change details are OR10 Each option below is scored as its own binary.		
OR3 · colour-changes		"Colour changes"
OR3 · melts-smears-loses		"Melts, smears or loses its shape"
OR3 · part-missing-appears		"A part of them is missing, or something appears that should not be there — a limb, a hand, hair, an accessory"
OR3 · shape-proportions-change		"Shape or proportions change"
OR3 · surface-texture-changes		"Surface or texture changes"
OR4	Cast count	Per character: exactly one — not duplicated and not silently missing? (tick each cast member)
OR5	Setting stability	Did the background setting change/morph into a different place? (Yes = it changed — polarity is inverted vs every other binary)
OR6: Does everything sit right in the scene — people AND objects AND the place: positions, sizes, resting on something, nothing clipping through, materials and light behaving? (tick each thing that does not) — PLACEMENT & PHYSICAL PLAUSIBILITY; a character's own appearance is OR3, a prop's is OR14 Each option below is scored as its own binary.		
OR6 · light-shadow-shadow		"Light or shadow wrong — a shadow that does not match what is there, or is a solid shape"
OR6 · materials-behave-wrongly		"Materials behave wrongly — water like a solid rod, cloth like metal, hair like stiff tubes"
OR6 · passes-through-solid		"Something passes through something solid"
OR6 · positions-front-behind		"Positions wrong — who is in front of / behind what"
OR6 · resting-person-hovering		"Not resting on anything — a person hovering or sunk into the ground, an object floating with nothing holding it"
OR6 · sizes-person-object		"Sizes wrong — a person or object the wrong size for what is around it"
OR7: Inside a single scene (no cut): did anything pop, swap, vanish or appear out of nowhere? (tick what it happened to) — the BACKSTOP for whole-person and background events; a character's own appearance is OR3, a prop's is OR14 Each option below is scored as its own binary.		
OR7 · background		"Part of the background or set"
OR7 · object		"An object in the scene"
OR7 · whole-person		"A whole person"
OR7 · worn-or-held		"Something worn or held"
OR8	Cross-cut continuity	Across cuts, does continuity hold (same chars/costume/props/place, sensible time)? (no N/A option exists on the sheet — every pilot clip is multi-shot)
OR9	Action coverage	Of the key actions decomposed from the prompt, which happened? (per action)
OR13	Framing & staging	Of the key actions, which are clearly framed & fully visible (not cut off/cramped/hidden)? (per action) *Conditional on OR9: asked per key action, only for the 14,115 of 19,650 actions OR9 records as having happened. All 5,535 unasked items are actions OR9 scored as not having happened, so the OR13 mean is framing GIVEN occurrence, not framing overall.*

Table 6: Rubric question glossary, overall-rubric axes. Every coded name used in this paper is mapped to the exact wording the annotator was shown. Codes are numbered densely within each rubric, so a gap never implies a missing question. A tick-list parent's question is shown once in italics above its options, and each option is scored as its own binary. Conditional questions state in italics what they are conditional on. Shading groups the sub-questions of one rubric and carries no other meaning.

Adjacent pair	Δ mean clip score	p (Holm)	Verdict
Seedance 2.5 > MiniMax-H3	0.039	0.001	separated
MiniMax-H3 > HappyHorse 1.1	0.017	0.006	separated
HappyHorse 1.1 > Wan 2.7	0.010	0.163	**not separated**
Wan 2.7 > Kling O3	0.004	0.391	**not separated**
Kling O3 > Grok Imagine	0.033	0.001	separated

Table 7: The 5 adjacent comparisons that the printed leaderboard order asserts. Differences are in mean video-clip score over the 317 stories; p is bootstrap over stories (4,000 resamples) with Holm correction across all 15 pairwise comparisons. 3 of 5 adjacent steps survive, so the leaderboard is a partial order rather than a ranking.

Rubric family	Question	Nickname	m	Wan 2.7	Kling O3	Grok Imagine	MiniMax-H3	HappyHorse 1.1	Seedance 2.5
Hard-motion integrity	M1.1	Performs hard motion	40	0.550	0.637	0.600	0.800	0.738	**0.938**
Hard-motion integrity	M1.2	Identity through motion	40	0.988	0.963	0.875	0.988	**1.000**	0.988
Hard-motion integrity	M1.3	Natural deformation	40	0.787	0.775	0.812	0.912	0.850	**0.988**
Hard-motion integrity	M1.4	Others stay separate	40	1.000	1.000	0.975	**1.000**	1.000	1.000
Prop binding	M2.1	Prop used in action	41	1.000	1.000	1.000	**1.000**	1.000	1.000
Prop binding	M2.2	Prop matches reference	41	0.854	0.890	0.598	0.878	0.817	**0.939**
Prop binding	M2.3	Prop use sensible	41	0.939	0.951	0.854	0.963	0.951	**0.976**
Interaction	M3.1	Did this interaction happen?	39	0.470	0.412	0.443	0.630	**0.724**	0.715
Interaction	M3.3	Contact-rich, not idle	39	0.833	0.744	0.782	0.872	**0.910**	0.872
Occlusion	M4.1	Visible part correct	40	0.900	0.863	0.875	0.900	**0.975**	0.963
Occlusion	M4.2	Same character returns	40	0.912	0.938	0.938	0.975	**1.000**	1.000
Occlusion	M4.3	Depth correct	40	0.975	0.988	0.963	0.963	0.988	**1.000**
Occlusion	M4.4	Bindings survive occlusion	40	**0.950**	0.875	0.900	0.925	0.925	0.950
Look-alike	M5.1	Both present & active	37	1.000	1.000	1.000	**1.000**	1.000	0.973
Look-alike	M5.2	Target does the action	33	0.970	0.970	**1.000**	0.909	1.000	1.000
Look-alike	M5.3	Look-alike never interacts	33	**0.970**	0.848	0.848	0.818	0.970	0.970
Look-alike	M5.4	Never swapped	33	0.985	0.970	0.955	**1.000**	1.000	1.000
Twin	M6.1	Both twins present & active	39	**1.000**	0.949	1.000	0.974	1.000	1.000
Twin	M6.2	Distinguishing detail kept	39	0.910	0.833	**0.987**	0.923	0.910	0.962
Twin	M6.3	Correct twin acts	35	0.971	0.800	**1.000**	0.857	0.971	1.000
Twin	M6.4	Twin never interacts	35	**0.971**	0.771	0.943	0.743	0.971	0.943
Twin	M6.5	Exactly two (no duplicate)	39	0.897	0.872	0.872	0.974	0.949	**1.000**
Handover	M7.1	Contact physically real	40	0.775	0.838	0.738	**0.912**	0.875	0.875
Handover	M7.2	Weight reads right	40	0.838	0.800	0.787	0.838	0.800	**0.912**
Handover	M7.3	Contact clearly staged	40	0.812	0.938	0.825	**0.963**	0.925	0.938
End-to-end Persistence	M8.1	Identity holds first-to-last	41	**1.000**	0.959	0.963	0.982	0.974	1.000
End-to-end Persistence	M8.2	Prop lifecycle complete	39	0.863	0.800	0.855	0.836	0.931	**0.957**

Table 8: Per-question capability profile, probe families M1 to M8. Mean human score for every scored rubric question, by model; the best model in each row is bold. m is the number of story prompts the question was asked on. The rubric family is named as the Figure 2 spokes label it, and a rule separates each family. A blank nickname means the question's nickname is the family name itself. Shading groups the questions of one family and carries no other meaning. Exact wording for every code is in Table 5.

Rubric family	Question	Nickname	m	Wan 2.7	Kling O3	Grok Imagine	MiniMax-H3	HappyHorse 1.1	Seedance 2.5
Cast & props present	OR1	Cast & props present	317	0.991	0.993	0.870	0.983	0.988	**0.997**
Motion smoothness	OR2	Motion smoothness	317	0.836	0.825	0.809	0.905	0.888	**0.954**
Appearance consistency	OR3	colour-changes	317	0.937	0.940	0.852	0.946	0.975	**0.994**
Appearance consistency	OR3	melts-smears-loses	317	0.978	0.956	0.968	0.978	0.981	**0.997**
Appearance consistency	OR3	part-missing-appears	317	0.744	0.713	0.653	0.804	0.779	**0.811**
Appearance consistency	OR3	shape-proportions-change	317	0.874	0.871	0.744	0.937	0.776	**0.953**
Appearance consistency	OR3	surface-texture-changes	317	0.987	0.991	0.934	0.994	0.994	**1.000**
Cast count	OR4	Cast count	317	0.974	0.895	0.945	0.944	0.978	**0.986**
Setting stability	OR5	Setting stability	317	0.984	0.994	0.984	**0.994**	0.987	0.994
Placement & spatial sense	OR6	light-shadow-shadow	317	0.987	0.975	0.991	0.984	0.984	**0.994**
Placement & spatial sense	OR6	materials-behave-wrongly	317	0.946	0.950	0.931	0.956	0.893	**0.978**
Placement & spatial sense	OR6	passes-through-solid	317	0.744	0.669	0.659	0.691	0.562	**0.757**
Placement & spatial sense	OR6	positions-front-behind	317	0.830	0.839	0.858	0.877	0.931	**0.965**
Placement & spatial sense	OR6	resting-person-hovering	317	0.732	0.842	0.700	0.814	0.691	**0.861**
Placement & spatial sense	OR6	sizes-person-object	317	0.931	0.946	0.896	**0.975**	0.909	0.975
Within-shot stability	OR7	background	317	0.956	0.981	0.959	**0.987**	0.978	0.981
Within-shot stability	OR7	object	317	0.798	0.748	0.644	0.760	0.770	**0.931**
Within-shot stability	OR7	whole-person	317	0.943	0.871	0.877	0.946	0.934	**0.959**
Within-shot stability	OR7	worn-or-held	317	0.779	0.719	0.659	0.738	0.574	**0.811**
Cross-cut continuity	OR8	Cross-cut continuity	317	0.809	0.916	0.877	**0.934**	0.880	0.931
Action coverage	OR9	Action coverage	317	0.665	0.652	0.663	0.775	0.806	**0.829**
Fine details	OR10	Fine details	317	0.969	0.970	0.901	0.969	0.974	**0.975**
Fine details	OR10	extra-changed-or-missing	317	0.618	0.609	0.527	0.669	0.789	**0.839**
Physics & deformation	OR11	Physics & deformation	317	0.815	0.861	0.880	0.905	0.902	**0.956**
Pacing	OR12	Pacing	317	0.951	0.945	0.950	0.965	0.959	**0.976**
Action staging	OR13	Action staging	317	0.883	0.894	0.889	0.922	0.949	**0.973**
Prop appearance stability	OR14	colour-changes	133	0.891	0.940	0.902	0.917	0.985	**0.992**
Prop appearance stability	OR14	comes-loose-floats	133	0.917	0.940	0.910	**0.977**	0.812	0.970
Prop appearance stability	OR14	material-changes	133	0.895	0.947	0.891	0.917	0.985	**0.992**
Prop appearance stability	OR14	morphs-into-else	133	0.977	0.992	0.966	**1.000**	0.992	1.000
Prop appearance stability	OR14	shape-changes	133	0.789	0.857	0.688	**0.895**	0.759	0.880

Table 9: Per-question capability profile, overall-rubric axes. Mean human score for every scored rubric question, by model; the best model in each row is bold. m is the number of story prompts the question was asked on. The rubric family is named as the Figure 2 spokes label it, and a rule separates each family. A blank nickname means the question's nickname is the family name itself. Shading groups the questions of one family and carries no other meaning. Exact wording for every code is in Table 6.

Judge	ρ (95% CI)	ρ_c	Concordance (95% CI)	Margin	Gwet	Sep. (×human)	Minority recall / prec.	Lat. (s)
Gemini 3.1 Pro	0.771 (0.54–0.94)	0.566	0.738 (0.70–0.78)	−0.221	0.657	**1.03**	**0.330** / 0.173	50
Muse Spark 1.2	0.429 (-0.14–0.83)	0.466	0.681 (0.63–0.73)	−0.015	0.866	0.45	0.169 / 0.259	**20**
Qwen 3.8 Max	**0.886 (0.71–0.94)**	**0.693**	0.783 (0.74–0.82)	−0.028	0.850	0.87	0.128 / 0.300	256
Majority-of-3	0.886 (0.66–0.94)	0.688	**0.796** (0.76–0.83)	**−0.010**	**0.868**	0.66	0.171 / **0.330**	–

Table 10: **Judge selection.** Full version of Table 1. All metrics are computed on the same video clips and prompt (1,876 clips over 314 stories with judge labels, drawn from the 317-story human panel). ρ is the Spearman correlation between the human and judge orderings of the six models; ρ_c is the question-centred rank correlation over model-ordering cells. *Concordance* is the share of question-level model pairs whose human direction the judge reproduces, with a bootstrap interval over questions. *Margin* is agreement minus the majority-class baseline. *Separation* (Sep.) reports how much of the score spread between models is retained relative to human annotation, where 1.0× denotes the same spread as humans. *Minority recall and precision* are measured on the items where the human departs from the question's majority answer. *Latency* (Lat.) is the median evaluation time per clip in seconds. Bold marks the best value per column; for separation the best is the value closest to 1.

		over clips			over the six models		
Rubric family	#clips	Gemini	Muse	Qwen	Gemini	Muse	Qwen
M1 · Hard-motion integrity	240	0.222	0.058	0.115	0.486	−0.371	0.771
M2 · Prop binding	245	0.127	0.125	0.251	0.657	0.429	0.829
M3 · Interaction	216	0.313	0.394	0.466	0.771	0.886	0.886
M4 · Occlusion	237	0.124	0.151	0.063	0.486	0.600	0.714
M5 · Look-alike	222	0.372	0.600	0.424	0.600	0.580	0.429
M6 · Twin	233	0.469	0.484	0.514	0.714	0.943	0.829
M7 · Handover	239	0.205	0.151	0.245	0.943	0.886	0.943
M8 · End-to-end Persistence	243	0.348	0.451	0.336	0.771	0.429	0.829
OR1 · Cast & props present	1,876	0.248	0.227	0.232	0.314	0.029	0.455
OR2 · Motion smoothness	1,876	0.135	0.077	0.169	0.886	0.314	0.943
OR3 · Appearance consistency	1,869	0.120	0.129	0.173	0.771	0.771	0.943
OR4 · Cast count	1,876	0.570	0.597	0.586	0.943	0.543	0.943
OR5 · Setting stability	1,876	−0.012	−0.006	−0.002	0.725	0.034	0.133
OR6 · Placement & spatial sense	1,875	0.131	0.060	0.118	0.943	0.829	0.886
OR7 · Within-shot stability	1,875	0.188	0.112	0.330	0.543	0.829	0.886
OR8 · Cross-cut continuity	1,873	0.123	0.084	0.205	0.714	0.600	0.886
OR9 · Action coverage	1,876	0.461	0.423	0.499	0.829	0.943	0.943
OR10 · Fine details	1,876	0.280	0.356	0.351	0.829	−0.657	0.486
OR11 · Physics & deformation	1,876	0.113	0.103	0.194	0.829	−0.029	0.600
OR12 · Pacing	1,876	0.067	0.071	0.053	0.829	−0.086	0.657
median over families		**0.196**	**0.140**	**0.238**	**0.771**	**0.561**	**0.829**

Table 11: **Human versus AI-judge rank correlation by rubric family at two aggregation levels.** Left: Spearman correlation between human and judge mean clip scores across individual clips. Right: Spearman correlation computed over the six model-level mean scores within each rubric family. We use rank correlation because judge scores are systematically lower than human scores, while monotonic score shifts do not affect the ordering.