MSVBench: Towards Human-Level Evaluation of Multi-Shot Video Generation

Haoyuan Shi[1], Yunxin Li[1], Nanhao Deng[1], Zhenran Xu[1], Xinyu Chen[1], Longyue Wang[2], Baotian Hu[1]*, Min Zhang[1],
[1]Harbin Institute of Technology, Shenzhen, China,
[2]Alibaba Group, Hangzhou, China,
{shihaoyuanhitsz, liyunxin987}@163.com
{hubaotian, zhangmin2021}@hit.edu.cn

Abstract

The evolution of video generation toward complex, multi-shot narratives has exposed a critical deficit in current evaluation methods. Existing benchmarks remain anchored to single-shot paradigms, lacking the comprehensive story assets and cross-shot metrics required to assess long-form coherence and appeal. To bridge this gap, we introduce **MSVBench**, the first comprehensive benchmark featuring hierarchical scripts and reference images tailored for **M**ulti-**S**hot **V**ideo generation. We propose a hybrid evaluation framework that synergizes the high-level semantic reasoning of Large Multimodal Models (LMMs) with the fine-grained perceptual rigor of domain-specific expert models. Evaluating 20 video generation methods across diverse paradigms, we find that current models—despite strong visual fidelity—primarily behave as visual interpolators rather than true world models. We further validate the reliability of our benchmark by demonstrating a state-of-the-art Spearman's rank correlation of **0.944** with human judgments. Finally, MSVBench extends beyond evaluation by providing a scalable supervisory signal. Fine-tuning a lightweight model on its pipeline-refined reasoning traces yields human-aligned performance comparable to commercial models like Gemini-2.5-Flash.

1 Introduction

The field of video generation is transitioning from isolated, short clips to complex, multi-shot narratives, poised to revolutionize industries ranging from film production to interactive gaming. Yet, while commercial pioneers like Sora2 and Veo3.1 propel the community toward world models, evaluation methods lag significantly behind these generative breakthroughs, creating a critical bottleneck.

Existing benchmarks exhibit fundamental limitations when applied to complex multi-shot video evaluation. Early efforts such as VBench (Huang et al., 2024) and EvalCrafter (Liu et al., 2024b) depend on lightweight expert models with limited video-understanding ability, leading to unreliable assessments for intricate actions or narrative-rich content. More recent attempts, including VideoBench (Han et al., 2025), adopt LMMs to approximate human reasoning; however, their exclusive reliance on LMMs introduces a lack of objective, standardized criteria and insufficient domain-specific perceptual grounding. More importantly, these benchmarks remain anchored to single-shot prompt–video pairs and are therefore misaligned with the emerging paradigm of multi-shot, story-driven video generation.

Although pioneers like OpenS2V-Nexus (Yuan et al., 2025) and ViStoryBench (Zhuang et al., 2025) take initial steps toward story-level video evaluation, they still face two structural limitations that are particularly critical for multi-shot video evaluation. First, their story assets remain incomplete: the absence of fully detailed scripts and per-shot reference images limits the diversity of generation paradigms that can be evaluated. Second, their metric suites do not yet capture essential shot-level and cross-shot properties—such as temporal logic across consecutive shots—resulting in insufficient coverage of the core challenges unique to multi-shot narratives. Consequently, despite meaningful progress, these benchmarks cannot yet replace the costly and non-scalable practice of human evaluation in multi-shot video generation.

To address this, we introduce **MSVBench**, a comprehensive framework specially designed for multi-shot video evaluation. We formulate the task hierarchically, decomposing stories into global priors, scene-level segments, and shot-level conditions. To achieve human-level precision, our framework synergizes the semantic reasoning of LMMs with the perceptual fidelity of domain-specific expert models. This hybrid design enables a unified

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 24034–24058

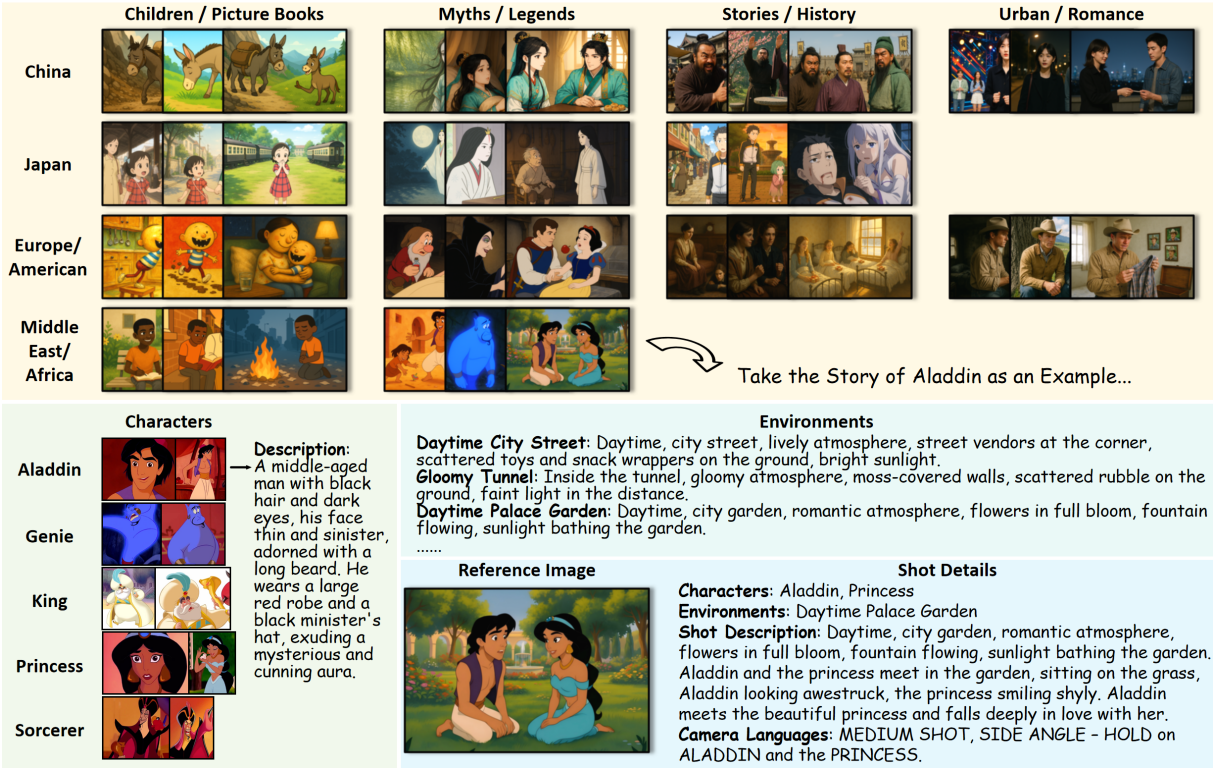

Figure 1: The Hierarchical Data Organization of MSVBench.

evaluation that jointly captures low-level visual fidelity and high-level narrative coherence, achieving a record-high Spearman's rank correlation of 0.944 with human judgments.

We conduct an in-depth evaluation of 20 diverse generation methods, ranging from commercial leaders to open-source and agent-based paradigms. Results reveal that while top-tier commercial models excel in dimensions like motion quality, open-source contenders are rapidly narrowing the gap. Furthermore, we expose a fundamental limitation in current video generation frameworks: Despite high prompt alignment capacity, current systems struggle with long-form stability and consistency across shots, functioning more as visual interpolators than true world models.

Finally, we construct a data pipeline that converts evaluation traces into thinking process data for high-quality supervision. By fine-tuning a lightweight Qwen3-VL-4B model on this data using Group Relative Policy Optimization (GRPO), we achieve human-aligned performance that surpasses commercial models like Gemini-2.5-Flash.

Our contributions are threefold:

- We propose **MSVBench**, the first comprehensive benchmark tailored for multi-shot video generation. MSVBench features a hierarchical data formulation and a hybrid evaluation framework, harmonizing the semantic reasoning capabilities of LMMs with the perceptual precision of domain-specific expert models.
- We evaluate 20 diverse systems, validating our benchmark's reliability with a state-of-the-art 0.944 human correlation. The results expose that current models operate primarily as visual interpolators rather than true world models.
- We construct a systematic pipeline transforming evaluation traces into supervision data. Training a specialized lightweight model on this data yields human-aligned performance surpassing even commercial models.

2 Related Work

2.1 Multi-shot Video Generation

Existing approaches fall into four streams: (1) **Storyboard-Driven Synthesis** decouples narrative from dynamics via a two-stage pipeline. Methods like StoryDiffusion (Zhou et al., 2024) and StoryAdapter (Mao et al., 2024) first generate coherent keyframes, subsequently animated by advanced image-to-video models (Wan et al., 2025; Yang et al., 2024b; Kong et al., 2024; HaCohen et al., 2024; Jiang et al., 2024; Huang et al., 2025b). (2)

Table 1: Comparison of video benchmarks. **Ref(C/S)**: Character/Shot reference images. **Aspects**: Visual Quality (VQ), Story/Prompt Video Alignment (SVA/PVA), Video Consistency (VC), Motion Quality (MQ). **Evaluator**: Evaluation backbone (Spec.: Specialized Models). **H-Align**: Correlation analysis with human evaluation.

Benchmark	Ref(C/S)	Aspects	Evaluator	H-Align	Metrics	Methods
VBench(Huang et al., 2024)	-/-	VQ/PVA/VC/MQ	Spec.	**Yes**	16	9
EvalCrafter(Liu et al., 2024b)	-/-	VQ/PVA/VC/MQ	Spec.	**Yes**	16	8
Video-Bench(Han et al., 2025)	-/-	VQ/PVA/VC/MQ	LLM	**Yes**	9	7
OpenS2V-Nexus(Yuan et al., 2025)	479/-	VQ/PVA/VC/MQ	LMM/Spec.	No	19	8
ViStoryBench(Zhuang et al., 2025)	**509**/-	VQ/SVA/VC	LMM/Spec.	Partial	12	18
MSVBench	136/**276**	**VQ/SVA/VC/MQ**	**LMM+Spec.**	**Yes**	**20**	**20**

Agent-Based Frameworks leverage LLM orchestration for planning. MovieAgent (Wu et al., 2025), MM-StoryAgent (Xu et al., 2025a), and VideoGen-of-Thought (Zheng et al., 2024) focus on structural reasoning, while AnimDirector (Li et al., 2024) and AniMaker (Shi et al., 2025) employ a "generate-and-select" mechanism for quality assurance. Recent works like FilmAgent (Xu et al., 2025b), FilMaster (Huang et al., 2025a), and HoloCine (Meng et al., 2025) further incorporate 3D environments or cinematic principles. (3) **Continuous Generation Techniques** maintain long-term coherence via autoregressive mechanisms (e.g., LongLive (Yang et al., 2025), RollingForcing (Liu et al., 2025a)) or test-time optimization (TTT-Video (Dalal et al., 2025)). (4) **Commercial Solutions** such as Sora2 (OpenAI, 2024) and Veo3.1 (Google, 2025) set industry standards, demonstrating exceptional consistency across complex narratives.

2.2 Video Evaluation

The landscape of video evaluation has evolved rapidly to keep pace with generative advancements. Prior reference-free benchmarks, such as VBench (Huang et al., 2024) and EvalCrafter (Liu et al., 2024b), utilized specialized small models to assess basic quality dimensions. Recognizing the need for semantic reasoning, subsequent works like Video-Bench (Han et al., 2025) shifted towards LMMs to better align with human perception. More recently, the focus has expanded to story-oriented generation. Benchmarks like OpenS2V-Nexus (Yuan et al., 2025) introduced reference image integration, while ViStoryBench (Zhuang et al., 2025) pioneered story-level assessment. However, a critical gap remains in evaluating complex, multi-shot temporal logic and consistent character identities across extended narratives. Table 1 presents a detailed comparison between these baselines and our MSVBench.

3 MSVBench

3.1 Hierarchical Dataset Schema

To evaluate multi-shot videos, MSVBench organizes data into a structured hierarchy comprising global priors, scene-level segments, and shot-level conditions. This schema accommodates diverse generation paradigms (Sec. 2.1) through the following components:

Global Context. We define global assets as a set of n characters $\mathcal{C} = \{C_1, \ldots, C_n\}$ and k environments $\mathcal{E} = \{E_1, \ldots, E_k\}$. Each character C_i is represented by a tuple $(T_{name}, T_{desc}, \mathcal{I}_{ref})$, ensuring identity consistency via reference images. Similarly, each environment E_j consists of a unique designation and a textual setting description.

Hierarchical Script. The narrative is structured as a sequence of *scenes*, each anchored to a specific environment $E_i \in \mathcal{E}$. Each scene represents a continuous narrative segment, which is further decomposed into a sequence of atomic *shots*.

Shot Annotations. Each shot S_t contains comprehensive multimodal annotations: (i) *Visual Context*, specifying the subset of on-screen characters $\mathcal{C}_{sub} \subseteq \mathcal{C}$ alongside a reference frame; (ii) *Shot Description*, detailing visual states and dynamic actions; and (iii) *Cinematography*, providing instructions for camera movements.

3.2 Dataset Construction

Derived from 20 diverse stories in ViStoryBench (Zhuang et al., 2025), we restructure raw scripts into our hierarchical schema (see Fig. 1). The data curation pipeline comprises three stages:

Visual Grounding. We synthesize high-fidelity reference frames via `GPT-Image-1` and `Nano Banana`, establishing consistent visual narratives.

Prompt Refinement. Text prompts are rewritten to jointly describe static states and dynamic actions, ensuring strict alignment with reference visuals.

Cinematography Enrichment. Leveraging `Gemini-2.5-Flash`, we translate static shot specifications (e.g., scale, angle) into explicit, dynamic camera motion instructions.

Table 2: Overview of MSVBench Metrics.

Dimension	Metric	Underlying Models
Visual Quality	Dover Score	DOVER (VQA_A, VQA_T)
	MusIQ Score	MusIQ
	Visual Attr. Consist.	-
	Style Consist.	CSD-ViT-L
Story Video Alignment	VQAScore	CLIP-FlanT5-XXL (VQAScore)
	Detect & Count Score	Gemini-2.5-Flash
	Shot Perspective Align.	Gemini-2.5-Flash
	State Shift & Persistence	Gemini-2.5-Flash
	Story Video Consist.	ShareCaptioner + KaLM-Embedding-V2
Video Consistency	Face Consist.	SAM-Track + DeepFace / InceptionNeXt + Gemini-2.5-Flash
	Character Consist.	SAM-Track + InceptionNeXt + Gemini-2.5-Flash
	Background Consist.	Step1X-Edit + DreamSim
	Clothes & Color Consist.	Gemini-2.5-Flash
	Relative Size Consist.	Gemini-2.5-Flash
Motion Quality	Action Recognition	SAM2 Tracker + VideoMAE V2 + CLIP + Gemini-2.5-Flash
	Action Strength	RAFT
	Camera Control	MonST3R + Gemini-2.5-Flash
	Phys. Plausibility	Gemini-2.5-Flash
	Phys. Interaction	Gemini-2.5-Flash

3.3 Metrics

Benchmarks relying solely on specialized models objectively measure specific low-level features but fundamentally struggle with tasks requiring complex reasoning, such as long-term narrative alignment and context-aware character consistency. Conversely, exclusive reliance on LMMs yields strong semantic comprehension but often sacrifices fine-grained perceptual precision. More critically, LMM-only evaluations are highly susceptible to metric entanglement (confounding bias). For instance, an LMM (Li et al., 2025a,b) might erroneously penalize motion quality or temporal consistency simply because a video sample suffers from low visual resolution or static artifacts. Consequently, relying exclusively on either small specialized models or large multimodal models to construct metrics inevitably yields biased, incomplete assessments.

To effectively disentangle these evaluation dimensions and resolve the inherent trade-off between semantic reasoning and perceptual precision, we propose the hybrid evaluation framework of MSVBench, as illustrated in Figure 2. This approach synergizes domain-specific expert models (e.g., DOVER, RAFT) for assessing low-level perceptual fidelity with `Gemini-2.5-Flash` for handling high-level semantic reasoning. As summarized in Table 2, we construct a comprehensive evaluation system consisting of **20 sub-metrics** across **four core dimensions**—Visual Quality, Story Video Alignment, Video Consistency, and Motion Quality. This ensures that the capabilities of multi-shot long video generation are assessed systematically, where complex semantic reasoning and precise perceptual accuracy are evaluated independently and robustly, mitigating the confounding biases. The detailed prompts for the LMM are provided in Appendix A.

3.3.1 Visual Quality

Dover Score DOVER (Wu et al., 2023) is employed to assess video quality via VQA_A (aesthetic appeal, e.g., composition) and VQA_T (technical distortions, e.g., noise).

MusIQ Score MusIQ (Ke et al., 2021) provides a unified perceptual index by averaging frame-level composition, sharpness, and artifact assessments.

Visual Attribute Consistency This metric evaluates the stability of brightness (mean intensity), contrast (intensity standard deviation), and saturation (mean HSV S-channel) by calculating their absolute differences across targeted frame pairs (e.g., adjacent frames within a clip or boundary frames between clips).

Style Consistency CSD-ViT-L (Somepalli et al., 2024) is employed to extract style-disentangled embeddings at 1 fps, and the consistency score is defined as their mean cosine similarity across inter- and intra-shot frames.

3.3.2 Story Video Alignment

VQAScore To assess semantic consistency, we utilize VQAScore (Lin et al., 2024) with the CLIP-FlanT5-XXL backbone. Unlike standard embeddings matching, this metric employs a visual question-answering formulation to quantify text-frame alignment, averaged across frames sampled at 0.5s intervals.

Detection & Count Score This metric verifies the generation of script-required objects and characters. Leveraging `Gemini-2.5-Flash`, we first extract target entities from the input prompt and then detect their actual presence in the video. The final score quantifies the inclusion of these mandated elements, accounting for their visibility duration.

Shot Perspective Alignment We leverage `Gemini-2.5-Flash` to align intended shot

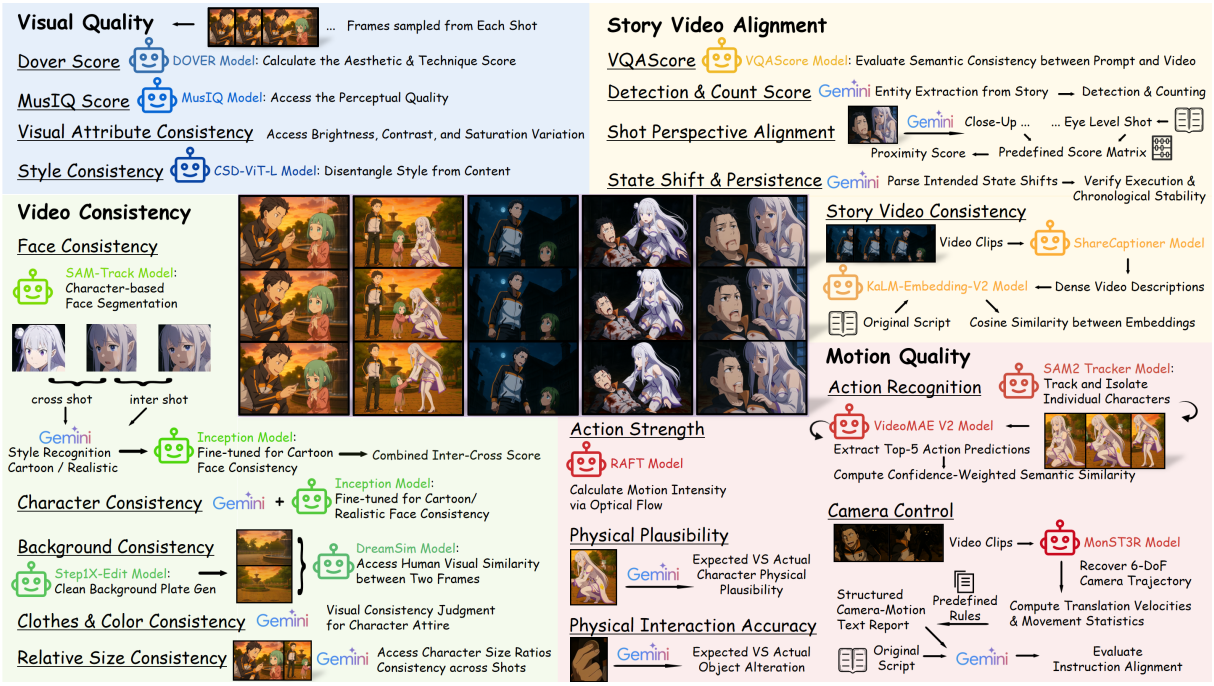

Figure 2: The MSVBench Evaluation Framework.

attributes with video realizations. Using a compatibility matrix over binary matching, the score reflects semantic proximity, penalizing severe deviations (e.g., Close-Up vs. Long Shot) more than minor shifts (e.g., Full Shot vs. Long Shot).

State Shift & Persistence To assess chronological continuity, we employ a two-stage pipeline using `Gemini-2.5-Flash`. The model first identifies script-driven state changes (e.g., character appearance or environment) and localizes their occurrence in the video. Then it examines the **persistence** of the new state, aggregating scores for both the initial shift's execution and its subsequent stability.

Story Video Consistency To assess script-video consistency, we generate dense video descriptions via ShareCaptioner (Chen et al., 2024). These descriptions are compared against the input script within the KaLM-Embedding-V2 (Zhao et al., 2025) semantic space, with the final score defined as the cosine similarity of their embeddings.

3.3.3 Video Consistency

Face Consistency We quantify facial consistency using a style-adaptive pipeline. Following localization via SAM-Track (Cheng et al., 2023), `Gemini-2.5-Flash` routes inputs to domain-specific extractors: DeepFace (Taigman et al., 2014) for realistic content, and our custom InceptionNeXt (Yu et al., 2024) variant fine-tuned on a composite of anime datasets (splcher, 2019; Naftali et al., 2022; Zheng et al., 2020) for animation. The final score is calculated as the feature distance between extracted embeddings.

Character Consistency Adopting a similar pipeline, we evaluate character consistency by processing SAM-Track outputs via our specialized InceptionNeXt extractors. Guided by style identification from `Gemini-2.5-Flash`, the system selects between two variants we specifically fine-tune: a realistic model adapted from MVHumanNet (Xiong et al., 2024), and an animated model trained on Hunyuan3D-1.0 (Yang et al., 2024a) synthetic data. The final score is derived from the embedding feature distance.

Background Consistency To assess environmental stability, we employ an occlusion-aware pipeline. Utilizing Step1X-Edit (Liu et al., 2025b) to remove foreground entities (e.g., characters) and yield clean background plates, we quantify their perceptual consistency via the DreamSim (Fu et al., 2023) metric based on feature distance.

Clothes & Color Consistency We evaluate attire stability using `Gemini-2.5-Flash` via a two-step protocol: confirming character presence before grading clothing consistency on a 5-point scale. Notably, the metric is robust to view changes, distinguishing true style discrepancies from natural

variations caused by lighting or angles.

Relative Size Consistency Perspective-invariant size ratios between character pairs are estimated by `Gemini-2.5-Flash`. This metric isolates actual body dimensions from viewing angles, quantifying consistency via the temporal variance of these pairwise proportions.

3.3.4 Motion Quality

Action Recognition We adopt a character-centric approach for complex scenarios. Following isolation via `Gemini-2.5-Flash` and SAM2 Tracker (Ravi et al., 2024), VideoMAE V2 (Wang et al., 2023) generates top-5 action predictions from the respective spatiotemporal regions. The consistency score is derived from the confidence-weighted sum of CLIP cosine similarities between these predictions and character-specific targets.

Action Strength To quantify motion intensity, we utilize RAFT (Teed and Deng, 2020) to extract dense optical flows between consecutive frames. The final score is defined as the average flow magnitude across the sequence, explicitly penalizing overly static content.

Camera Control To assess camera control fidelity, we employ MonST3R (Zhang et al., 2024) to recover frame-wise 6-DoF trajectories. These physical parameters are then translated into semantic motion descriptors (e.g., "pan right"), which `Gemini-2.5-Flash` compares against textual instructions to quantify execution accuracy.

Physical Plausibility We employ a character-centric pipeline powered by `Gemini-2.5-Flash` to verify adherence to Newtonian mechanics (e.g., gravity, momentum). The process involves extracting expected characters and dynamics from the prompt, followed by grading their motion realism in the video on a 5-point scale. The final metric averages these scores, explicitly penalizing violations like floating, interpenetration, or unnatural balance.

Physical Interaction Accuracy Complementing character-centric metrics, this module targets object-centric interaction fidelity. Utilizing `Gemini-2.5-Flash`, we extract expected reactions (e.g., deformation) from the prompt and verify their adherence to causal logic and material properties in the video. The score reflects the physical accuracy of consequential motions and structural changes.

4 Experiments

4.1 Settings

We evaluate a comprehensive range of methods on MSVBench across four distinct paradigms: (1) Storyboard-Driven Synthesis: We assess StoryGen (Liu et al., 2024a), StoryDiffusion (Zhou et al., 2024), and StoryAdapter (Mao et al., 2024), alongside decoupled I2V pipelines (Wan series (Wan et al., 2025), HunyuanVideo (Kong et al., 2024), AniSora (Jiang et al., 2024)) using `GPT-Image-1` keyframes. (2) Agent-Based Frameworks: We benchmark systems including MovieAgent (Wu et al., 2025), MM-StoryAgent (Xu et al., 2025a), VideoGen-of-Thought (Zheng et al., 2024), and AniMaker (Shi et al., 2025). (3) Continuous Generation Techniques: We evaluate the autoregressive method LongLive (Yang et al., 2025). (4) Commercial Solutions: Models like Sora2 (OpenAI, 2024) and Veo3.1 (Google, 2025) are included.

4.2 Main Results

Table 3 details the performance of 20 multi-shot video generation methods across four dimensions. Commercial models (Sora2, Veo3.1) currently define the state-of-the-art, exhibiting superior robustness in Story Video Alignment and Motion Quality. However, the gap with open-source systems is rapidly narrowing. The Wan2.2 family emerges as the premier open-source contender; remarkably, Wan2.2-I2V achieves parity with commercial models in Video Consistency, while Wan2.2-T2V demonstrates highly competitive Motion Quality. Within Agent-Based frameworks, AniMaker distinguishes itself through its well-balanced performance profile across all evaluation dimensions.

4.3 Critical Insights

Synthesizing the quantitative results of evaluated methods and case studies of evaluation metrics—representative examples of which are visualized in Appendix C and D—we move beyond surface-level rankings to uncover fundamental architectural constraints preventing current models from functioning as true world models. Three structural impediments are identified:

Fragmented Generation rather than Holistic Modeling. While current models excel at single-shot interpretation—evidenced by high Story Video Alignment (S.V.A)—they largely fail to **model** the generated content. This deficiency manifests at two scales: locally, low Physical Interaction Accuracy

Table 3: Comprehensive quantitative evaluation results. The metrics are categorized into four dimensions: **Visual Quality** (Dov.: Dover Score, Mus.: MusIQ Score, V.A.C: Visual Attribute Consistency, S.C.: Style Consistency); **Story Video Alignment** (VQA: VQAScore, Det.: Detection & Count Score, S.P.A: Shot Perspective Alignment, S.S.P: State Shift & Persistence, S.V.C: Story Video Consistency); **Video Consistency** (Fac.: Face Consistency, Cha.: Character Consistency, Bac.: Background Consistency, Clo.: Clothes & Color Consistency, Siz.: Relative Size Consistency); and **Motion Quality** (A.R.: Action Recognition, A.S.: Action Strength, Cam.: Camera Control, Phy.P: Physical Plausibility, Phy.I: Physical Interaction Accuracy). The background colors represent the best to the 8th best models, respectively. In cases of ties, the median color is adopted.

	Visual Quality				Story Video Alignment					Video Consistency					Motion Quality				
Model	**Dov.**	**Mus.**	**V.A.C**	**S.C.**	**VQA**	**Det.**	**S.P.A**	**S.S.P**	**S.V.C**	**Fac.**	**Cha.**	**Bac.**	**Clo.**	**Siz.**	**A.R.**	**A.S.**	**Cam.**	**Phy.P**	**Phy.I**
Storyboard-Driven Synthesis																			
AniSora3.2	0.68	0.21	81.50	0.96	81.72	83.02	4.20	3.93	81.41	87.05	85.38	0.53	3.63	0.98	23.57	17.41	3.57	3.41	2.41
CogVideoX1.5-5B-I2V	0.67	0.19	74.21	0.95	82.64	86.01	4.19	4.05	81.20	87.73	86.03	0.52	3.63	0.98	23.16	16.00	3.45	3.14	2.29
HunyuanVideo-I2V	0.58	0.20	74.86	0.95	81.30	77.77	4.13	3.72	81.34	89.14	86.08	0.52	3.36	0.99	24.90	31.44	3.58	3.11	2.43
LTXV-13B-0.9.8	0.67	0.19	74.97	0.94	81.87	76.93	4.07	3.83	81.69	88.25	86.12	0.55	3.61	0.99	24.84	35.22	3.52	2.76	2.34
S.A.+Wan2.2-I2V	0.57	0.20	86.39	0.96	71.95	65.28	3.54	2.81	80.55	88.97	81.68	0.43	0.84	0.99	29.76	56.76	3.67	2.58	1.63
S.D.+Wan2.2-I2V	0.63	0.20	75.53	0.94	61.88	55.83	3.60	2.33	79.93	88.84	82.12	0.56	0.96	1.00	30.59	73.80	3.59	2.47	1.24
Self-Forcing	0.71	0.16	83.28	0.96	85.10	69.60	3.99	3.28	81.11	88.66	83.65	0.56	2.10	0.97	32.48	52.58	3.57	3.15	1.93
StoryAdapter	0.47	0.20	93.59	0.98	67.25	54.94	3.47	1.48	80.39	88.38	79.41	0.39	0.78	1.00	33.16	0.47	3.65	0.77	0.79
StoryDiffusion	0.54	0.20	87.73	0.96	67.51	54.68	3.66	1.72	80.52	89.95	81.92	0.51	1.05	1.00	36.33	0.52	3.41	0.76	0.87
StoryGen	0.25	0.19	88.38	0.96	32.82	14.16	3.27	1.28	76.60	79.70	82.05	0.31	0.37	0.93	25.91	0.66	3.48	0.26	0.33
Wan2.2-I2V	0.65	0.20	79.36	0.92	83.22	90.90	4.07	4.05	81.48	88.85	85.10	0.56	3.62	0.97	25.16	26.64	3.53	3.44	2.13
Wan2.2-T2V	0.65	0.16	82.58	0.94	86.22	88.66	3.99	3.43	81.40	88.94	84.27	0.56	2.04	0.97	27.32	46.90	3.60	3.86	2.87
Wan2.2-TI2V	0.68	0.20	78.48	0.95	78.71	78.84	4.21	3.61	81.33	88.34	85.97	0.54	3.56	0.99	25.76	50.26	3.60	3.26	2.31
Agent-Based Frameworks																			
AniMaker	0.69	0.21	81.90	0.95	82.13	86.01	4.05	3.88	81.39	87.73	86.07	0.53	3.41	0.95	27.07	49.41	3.45	3.24	2.52
MM-StoryAgent	0.45	0.21	90.13	0.98	34.27	-	-	-	79.69	-	-	0.44	-	-	14.22	0.69	-	0.76	0.79
MovieAgent	0.61	0.18	72.50	0.96	47.84	45.71	3.54	1.29	79.14	-	-	0.45	0.77	0.97	24.90	28.43	2.89	2.54	1.68
VideoGen-of-Thought	0.56	0.21	91.75	1.00	60.11	32.07	3.56	1.68	80.51	83.58	87.70	0.42	1.22	1.00	25.11	63.36	3.55	0.71	0.59
Continuous Generation Techniques																			
LongLive	0.60	0.16	91.66	0.98	67.75	-	-	-	77.10	-	-	-	-	-	30.62	57.36	-	3.49	2.56
Commercial Solutions																			
Sora2	0.65	0.19	90.57	0.98	79.15	88.58	4.23	4.09	81.40	87.66	86.43	0.46	3.60	0.99	33.08	30.86	3.56	3.74	2.64
Veo3.1	0.64	0.21	91.04	0.98	77.91	90.67	4.18	4.10	82.24	86.64	85.77	0.51	3.66	0.99	33.27	34.80	3.68	3.67	2.78

(Phy.I) scores (< 3.0 / 5 even for Sora2 and Veo3.1) indicate an inability to model immediate causal dynamics such as collisions; globally, the degradation of Character Consistency (Cha.) and Clothing (Clo.) across multi-shot narratives reveals a failure to maintain a coherent character-level model that preserves identity and attributes. These failures suggest that current models function primarily as local visual interpolators rather than holistic world models, as they lack the capacity to maintain internal representation that governs physical laws and semantic consistency.

Trade-offs in Cross-Dimensional Model Capabilities. An inherent conflict exists between dynamic intensity and content preservation in current architectures. This tension manifests in two dimensions: First, in object dynamics, increasing Action Strength (A.S.) compromises Physical Interaction Accuracy (Phy.I), exemplified by S.D.+Wan2.2-I2V's exceptional A.S. (73.80 / 100) versus negligible Phy.I (1.24 / 5). This disparity confirms that extreme motion generation often distorts object structures, undermining the preservation of valid physical states. Second, in viewpoint dynamics, aggressive Camera Control (Cam.) disrupts Character Consistency (Cha.), as models prioritize rapid view shifts over semantic identity. These failures indicate that motion and stability mechanisms remain deeply entangled; thus, future designs that explicitly decouple motion generation from content preservation hold significant promise.

Reference Images as a Double-Edged Sword. While reference images provide dense visual guidance, enabling open-source models like Wan2.2-I2V to achieve Video Consistency comparable to commercial giants, they simultaneously act as a rigid constraint. Specifically, the static image **locks** the initial state but fails to convey depth or kinematic potential. Consequently, Wan2.2-T2V, unburdened by this 2D anchor, outperforms its I2V counterpart in Physical Plausibility (3.86 vs. 3.44). This performance gap highlights the intrinsic limitations of using 2D pixels as the sole conditioning signal; to resolve this, future generation paradigms

Table 4: Human evaluation results. We report Mean Opinion Scores across four dimensions: **V.Q.** (Visual Quality), **S.V.A.** (Story Video Alignment), **V.C.** (Video Consistency), and **M.Q.** (Motion Quality). The models are ranked by the overall average (**Avg.**). We highlight the **best**, second-best, and *third-best* results.

Model	V.Q.	S.V.A.	V.C.	M.Q.	Avg.
Veo3.1	**4.29**	**4.51**	**3.68**	3.97	**4.11**
Sora2	4.22	4.45	3.56	**4.01**	4.06
AniSora3.2	*3.98*	3.40	3.03	3.32	*3.43*
Wan2.2-I2V	3.70	3.48	2.86	*3.40*	3.36
LTXV-13B-0.9.8	3.90	3.27	3.06	3.22	3.36
AniMaker	3.04	3.42	*3.14*	3.08	3.17
Wan2.2-T2V	3.22	*3.63*	2.15	3.34	3.08
Wan2.2-TI2V	3.04	3.15	2.67	2.83	2.92
HunyuanVideo-I2V	2.71	2.92	2.82	2.58	2.76
Self-Forcing	2.96	2.77	2.36	2.65	2.69
CogVideoX1.5-5B-I2V	2.01	2.81	2.48	2.26	2.39
VideoGen-of-Thought	2.97	1.67	2.30	1.33	2.07
S.D.+Wan2.2-I2V	2.76	1.76	1.68	1.95	2.04
LongLive	2.67	1.31	2.69	1.51	2.04
S.A.+Wan2.2-I2V	2.92	1.74	1.68	1.77	2.03
StoryDiffusion	2.96	1.60	1.80	1.02	1.85
StoryAdapter	3.00	1.58	1.77	1.00	1.84
MM-StoryAgent	2.26	1.44	1.38	1.00	1.52
MovieAgent	1.89	1.12	1.39	1.03	1.36
StoryGen	1.03	1.01	1.02	1.00	1.01

must incorporate more comprehensive geometric inputs, such as 3D meshes or depth priors.

4.4 Human Preference Alignment

Human Evaluation Results We engage a group of human annotators to assess the generated videos across four key dimensions: Visual Quality (**V.Q.**), Story Video Alignment (**S.V.A.**), Video Consistency (**V.C.**), and Motion Quality (**M.Q.**). The detailed human evaluation setup, including annotator guidelines, is described in Appendix B. The results of the human evaluation are presented in Table 4.

Alignment with Human Perception We employ Spearman's (ρ) and Kendall's (τ) rank correlations to evaluate the consistency between benchmarks and human ratings. The detailed methodology is provided in Appendix E. Table 5 shows that MSVBench achieves an overall ρ of **0.944** and τ of **0.836**, significantly outperforming state-of-the-art baselines like VBench ($\rho = 0.585$, $\tau = 0.504$) and ViStoryBench ($\rho = 0.836$, $\tau = 0.660$). Notably, these overall correlations surpass the scores of individual dimensions across both metrics. This validates that the proposed metrics synergistically complement each other to accurately capture the holistic nature of human judgments.

Table 5: Correlation Analysis. Comparison of Spearman's (ρ) and Kendall's (τ) correlation coefficients between objective metrics and human judgments across different dimensions. We compare MSVBench with current evaluation metrics including VBench (Huang et al., 2024), EvalCrafter (Liu et al., 2024b) and ViStoryBench (Zhuang et al., 2025). We highlight the **best** and second-best results.

Aspects	Methods	Spearman's ρ	Kendall's τ
Visual Quality	Aesthetic-Quality (VBench)	0.470	0.395
	Imaging-Quality (VBench)	0.351	0.278
	ViStoryBench	0.272	0.154
	MSVBench (Ours)	**0.604**	**0.479**
Video Story Alignment	Clip-Score (EvalCrafter)	0.425	0.352
	BLIP-BLEU (EvalCrafter)	0.175	0.118
	ViStoryBench	0.805	0.621
	MSVBench (Ours)	**0.945**	**0.839**
Video Consistency	Subject Consistency (VBench)	0.543	0.432
	Background Consistency (VBench)	0.443	0.355
	CLIP-Temp (EvalCrafter)	0.017	0.014
	ViStoryBench	**0.665**	**0.487**
	MSVBench (Ours)	0.600	0.441
Motion Quality	Motion-Smoothness (VBench)	0.511	0.430
	Dynamic-Degree (VBench)	0.529	0.366
	MSVBench (Ours)	**0.649**	**0.471**
Overall	VBench	0.585	0.504
	EvalCrafter	0.263	0.144
	ViStoryBench	0.836	0.660
	MSVBench (Ours)	**0.944**	**0.836**

Table 6: Comparison of overall human alignment (Spearman's ρ and Kendall's τ) between the lightweight Qwen3-VL models and proprietary Gemini models. We highlight the **best** and second-best results.

Model	Overall Correlation	
	Spearman's ρ	Kendall's τ
Commercial Models		
Gemini-2.5-Flash	0.792	0.628
Gemini-2.5-Pro	**0.857**	**0.702**
Ours (Lightweight)		
Qwen3-VL-4B (Base)	0.799	0.596
Qwen3-VL-4B (RL)	0.836	0.663

4.5 From Benchmark to Supervisor

To verify the effectiveness of MSVBench as a source of high-quality supervision signals, we implement a data construction pipeline that transforms raw evaluation records into high-quality instruction tuning data. Using a subset of 15 stories, we synthesize over 1,000 samples and employ a Qwen3-VL-4B backbone trained with Group Relative Policy Optimization (GRPO) for a single epoch. As shown in Table 6, the model trained on our dataset achieves Spearman's ρ of **0.836** and Kendall's τ of **0.663**, effectively outperforming Gemini-2.5-Flash. Beyond these aggregate metrics, the data successfully transfers the fine-

grained discriminative criteria of human annotators to the model: the scoring distribution shifts from a "conservative" clustering around the median (score 3) to a confident coverage of the full 1–5 range. This dual improvement—high alignment accuracy combined with human-like variance—confirms MSVBench's capability as a robust supervisor, effectively transferring comprehensive preference patterns to downstream models.

5 Conclusion

We introduce MSVBench, the first unified framework for evaluating multi-shot video generation. By synergizing specialized expert models for perceptual precision with LMMs for semantic reasoning, our hybrid protocol effectively bridges the gap between low-level visual fidelity and high-level narrative consistency. Extensive validation confirms that MSVBench achieves state-of-the-art human alignment, significantly outperforming all existing video generation benchmarks. Beyond rigorously benchmarking current systems and identifying their limitations as world models, MSVBench serves as an automated pipeline for generating high-quality supervision data, enabling a lightweight Qwen3-VL-4B model trained on its reasoning traces to achieve more human-aligned video evaluation.

Limitations

Despite the comprehensive nature of MSVBench, several limitations remain to be addressed in future work:

- **Lack of Audio-Visual Metrics:** The current framework is predominantly centered on visual evaluation. It does not yet incorporate metrics for assessing audio-visual synchronization or the quality of generated audio—both of which are pivotal components for achieving immersive video generation.

- **Limited Scale of Story Data:** The benchmark's current limited size restricts the depth of our analysis and the method's overall robustness. Expanding the dataset with diverse story types would yield a more rigorous evaluation and significantly enhance generalizability. Furthermore, this data constraint limits the evaluation trajectories available for our data reconstruction pipeline. Scaling up the dataset would provide richer trajectories, substantially improving the training and effectiveness of our lightweight automated evaluator.

- **Challenges with Continuous Generation Models:** For continuous generation approaches that synthesize long videos without explicit internal shot segmentation, our evaluation pipeline faces alignment difficulties. The absence of discrete shot boundaries hinders the accurate mapping of corresponding text and reference image prompts, thereby rendering a specific subset of shot-level metrics inapplicable or difficult to compute.

Acknowledgments

We thank the editor and reviewers for their efforts in improving our paper. This work was supported by grants: National Natural Science Foundation of China (Grant No. 62422603), National Natural Science Foundation of China (Grant No. 625B2061), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024B0101050003) and Shenzhen Science and Technology Program (Grant No. ZDSYS20230626091203008).

References

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, and 1 others. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495.

Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. 2023. Segment and track anything. *arXiv preprint arXiv:2305.06558*.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. 2025. One-minute video generation with test-time training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17702–17711.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*.

Google. 2025. Veo 3 launch.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, and 1 others. 2024. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, and 1 others. 2025. Videobench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18858–18868.

Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. 2025a. Filmaster: Bridging cinematic principles and generative ai for automated film generation. *arXiv preprint arXiv:2506.18899*.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. 2025b. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and 1 others. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818.

Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Yin, Jing Liu, Chao Xu, Siqi Wang, Yidi Wu, Bingwen Zhu, Xinwen Zhang, and 1 others. 2024. Anisora: Exploring the frontiers of animation video generation in the sora era. *arXiv preprint arXiv:2412.10255*.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, and 1 others. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Yunxin Li, Xinyu Chen, Shenyuan Jiang, Haoyuan Shi, Zhenyu Liu, Xuanyu Zhang, Nanhao Deng, Zhenran Xu, Yicheng Ma, Meishan Zhang, and 1 others. 2025a. Uni-moe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data. *arXiv preprint arXiv:2511.12609*.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, and 1 others. 2025b. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*.

Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Jiashun Zhu, Jinyi Xu, Zhen Zhao, and Min Zhang. 2024. Anim-director: A large multimodal model powered agent for controllable animation video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

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

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. 2025a. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, and 1 others. 2025b. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*.

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

Jiawei Mao, Xiaoke Huang, Yunfei Xie, Yuanqi Chang, Mude Hui, Bingjie Xu, and Yuyin Zhou. 2024. Story-adapter: A training-free iterative framework for long story visualization. *arXiv preprint arXiv:2410.06244*.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, and 1 others. 2025. Holocine: Holistic generation of cinematic multi-shot long video narratives. *arXiv preprint arXiv:2510.20822*.

Martinus Grady Naftali, Jason Sebastian Sulistyawan, and Kelvin Julian. 2022. Aniwho: A quick and accurate way to classify anime character faces in images. *arXiv preprint arXiv:2208.11012*.

OpenAI. 2024. Sora: Creating video from text.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. 2025. Animaker: Automated multi-agent animated storytelling with mcts-driven clip generation. *arXiv preprint arXiv:2506.10540*.

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

splcher. 2019. Anime Face Dataset.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708.

Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, and 1 others. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. 2023. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *International Conference on Computer Vision (ICCV)*.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. 2025. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*.

Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, and 1 others. 2024. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19801–19811.

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

Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, Jun Yu, and Min Zhang. 2025b. Filmagent: A multi-agent framework for end-to-end film automation in virtual 3d spaces. *arXiv preprint arXiv:2501.12909*.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, and 1 others. 2025. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*.

Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, and 1 others. 2024a. Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, and 1 others. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Weihao Yu, Pan Zhou, Shuicheng Yan, and Xinchao Wang. 2024. Inceptionnext: When inception meets convnext. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 5672–5683.

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

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun,

and Ming-Hsuan Yang. 2024. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, and 1 others. 2025. Kalm-embedding-v2: Superior training techniques and data inspire a versatile embedding model. *arXiv preprint arXiv:2506.20923*.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and 1 others. 2024. Videogen-of-thought: A collaborative framework for multi-shot video generation. *arXiv e-prints*, pages arXiv–2412.

Yi Zheng, Yifan Zhao, Mengyuan Ren, He Yan, Xiangju Lu, Junhui Liu, and Jia Li. 2020. Cartoon face recognition: A benchmark dataset. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2264–2272.

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

Cailin Zhuang, Ailin Huang, Wei Cheng, Jingwei Wu, Yaoqi Hu, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, and 1 others. 2025. Vistorybench: Comprehensive benchmark suite for story visualization. *arXiv preprint arXiv:2505.24862*.

A MSVBench Prompt

This section details the comprehensive set of prompts tailored for the LMM evaluator, systematically structured to evaluate performance across key dimensions: Story Video Alignment, Video Consistency, and Motion Quality.

To assess Story Video Alignment, Figures 3 and 4 illustrate the prompts for verifying whether the generated visual elements and cinematic perspectives align with the textual narrative. The evaluation of Video Consistency is detailed in Figures 5 through 8, which focus on character identity, appearance stability, and relative spatial proportions across multiple scenes. Finally, Figures 9 through 12 provide the prompts tailored for Motion Quality, encompassing action fidelity, camera control, and the plausibility of physical interactions.

```
STORY_ENTITY_EXTRACTION_PROMPT = """
    Based on the following story description, please analyze the key objects and characters that can been seen in the video, and categorize them into two types:

    Story description: {description}
    Important instructions:
    1. When referring to people/characters, DO NOT use specific names like "Lily" or general terms like "mom", "mommy", "dad", etc.
    2. Instead, describe people by their clothes color based on the following character information, such as "a girl in black", "a woman in blue", "a man in black", etc. If there are no corresponding character information, use general terms like "a girl", "a woman", etc.

    Character information: {character}
    Please output in JSON format, including the following two categories:
    1. "full_presence": Objects/characters that need to appear throughout the entire video (such as main characters, background items)
    2. "any_presence": Objects/characters that only need to appear at any point in the video (such as action-related objects)

    Output format example:
    {{
        "full_presence": ["a girl in black", "a golden dog", "green book"],
        "any_presence": ["table", "red ball"]
    }}

    Please ensure:
    - Object/character names are concise and clear
    - Use visual appearance descriptions for people instead of names or family relationships
    - Categorize by importance
    - Only output JSON, no other explanations
    """

VIDEO_PRESENCE_SCORING_PROMPT = """
    Please analyze this video and identify which of the following objects/characters are present:

    Required objects/characters to detect:
    {objects_list}

    For each object/character in the list, please determine:
    1. Whether it appears in the video (yes/no)
    2. If it appears, rate its presence on a scale of 0-5:
       - 0: Not present at all throughout the video
       - 1: Appears very briefly (less than 20% of video duration)
       - 2: Appears occasionally (20-40% of video duration)
       - 3: Appears moderately (40-60% of video duration)
       - 4: Appears frequently (60-80% of video duration)
       - 5: Appears throughout most/all of the video (80-100% of video duration)

    Please provide your analysis in the following JSON format: .....

    Be thorough in your analysis and only mark objects as present if you can clearly see them in the video.
    Only output JSON, no other explanations.
    """
```

Figure 3: Prompt for Detect & Count Score

```
SHOT_DISTANCE_DEFINITIONS = """
    Shot Distance Descriptions
    • Long Shot: Shows the relationship between characters and their environment, typically used to display the scene or environment.
    • Full Shot: Shows the full body of a character, commonly used to display movement or the full scene.
    • Medium Long Shot: Starts from above the character's knees, capturing part of the environment.
    • Medium Shot: Captures the character from the waist up.
    • Close-Up: Captures the character from the chest up.
    • Extreme Close-Up: Focuses on the character's head or face, with the background and environment typically blurred or not visible.
    """

CAMERA_ANGLE_DEFINITIONS = """
    Angle Descriptions
    • Eye Level Shot: The camera is positioned at the subject's eye level.
    • Low Angle Shot: The camera is positioned below eye level, shooting upward, emphasizing the character's power or size.
    • High Angle Shot: The camera is positioned above eye level, shooting downward, often minimizing the subject's significance.
    • Bird's Eye View: Camera shot taken from directly above, providing an overview of the scene.
    • Tilted Shot: The camera is intentionally tilted to create a sense of imbalance or tension.
    • Perspective Compression: A technique that emphasizes depth and the relationship between foreground and background through perspective.
    """

SCRIPT_TO_SHOT_GENERATION_PROMPT = """
    Based on the following script and character/environment descriptions, generate appropriate shot distance and angle information for each clip.
    {SHOT_DISTANCE_DEFINITIONS} {CAMERA_ANGLE_DEFINITIONS}
    Script Data: {script_data}
    For each clip, analyze the content and suggest the most appropriate:
    1. Shot Distance (Long Shot, Full Shot, Medium Long Shot, Medium Shot, Close-Up, Extreme Close-Up)
    2. Camera Angle (Eye Level Shot, Low Angle Shot, High Angle Shot, Bird's Eye View, Tilted Shot, Perspective Compression)

    Return the result in JSON format with shot_distance and camera_angle fields added to each clip. Keep all original fields and just add the new ones.
    Example format: .....
    """

VIDEO_SHOT_ANALYSIS_PROMPT = """
    Analyze this single video shot and identify its shot distance and camera angle.
    {SHOT_DISTANCE_DEFINITIONS} {CAMERA_ANGLE_DEFINITIONS}
    For this shot, determine:
    1. Shot Distance: Long Shot, Full Shot, Medium Long Shot, Medium Shot, Close-Up, or Extreme Close-Up
    2. Camera Angle: Eye Level Shot, Low Angle Shot, High Angle Shot, Bird's Eye View, Tilted Shot, or Perspective Compression
    Return your analysis in JSON format as a single object: .....
    """
```

Figure 4: Prompt for Shot Perspective Alignment

```
SCRIPT_CHANGE_EXTRACTION_PROMPT = """
    Analyze the following script to identify change points where character states, environment states, or other important elements undergo internal transitions.
    IMPORTANT: Focus on INTERNAL changes within existing characters/environments, NOT switching between different characters or scenes. Examples:
    - Character changes: clothing getting dirty/torn, changing clothes, getting injured, hair getting wet, etc.
    - Environment changes: sunny to rainy weather, day to night, windows breaking, objects being consumed/destroyed, furniture moving, lighting changes, etc.
    - DO NOT include: characters motion changes, new characters entering, scene transitions, camera angle changes, or switching between different locations.
    Script content: {script_json}
    Please output change points in the following format:
    1. Change type (character change/environment change/other)
    2. Specific description of the internal state change
    3. State before the change
    4. State after the change
    Please output in JSON format, for example: .....
    """

VIDEO_CHANGE_LOCALIZATION_PROMPT = """
    Based on the following script analysis results, locate these changes in the video and provide initial scoring.
    Script change analysis: {script_changes_str}
    For each change point, please:
    1. Determine if the change can be observed in the video
    2. If observable, provide the approximate timestamp (seconds) when the change occurs
    3. Describe the specific manifestation of the change in the video
    4. Provide initial scoring based on change detection:
    - Score 1: Change is completely not reflected in the video
    - Score 2: Change moment is vague and cannot be accurately located
    - Score 3-5: Change is clearly detectable (will be further evaluated in next step)
    Please output in JSON format, for example: .....
    """

VIDEO_CHANGE_PERSISTENCE_EVALUATION_PROMPT = """
    Based on the following change location results, evaluate the persistence and consistency of changes that scored 3-5 in the video, and preserve the scores of changes that scored 1-2.
    Change location results: {located_changes_str}
    For each change that received a score of 3-5 (clearly detectable), please analyze:
    1. After the change occurs, is the new state maintained consistently in subsequent segments?
    2. Are there any shots that violate the post-change state?
    3. What is the degree of change persistence?
    For changes that received scores of 1-2, simply preserve their original scores and reasons.
    Scoring criteria for detectable changes:
    - Score 5: Subsequent segments consistently maintain the change until video end or new change occurs
    - Score 4: Subsequent segments maintain part of the change, or some shots preserve the change
    - Score 3: Subsequent segments completely do not maintain the change
    - Score 1-2: Keep original scores from location step
    Please output in JSON format, for example: .....
    """
```

Figure 5: Prompt for State Shift & Persistence

```
CHARACTER_VISUAL_STYLE_CLASSIFICATION_PROMPT = """
    Analyze this image/video and determine if it shows:
    1. Cartoon/animated characters (animated, drawn, CGI, stylized art style)
    2. Real people (photographic, realistic human characters)

    Please respond with only one word: 'cartoon' or 'real'
    """

FACE_VISUAL_STYLE_CLASSIFICATION_PROMPT = """
    Analyze this image/video and determine if it shows:
    1. Cartoon/animated faces (animated, drawn, CGI, stylized art style)
    2. Real people (photographic, realistic human faces)

    Please respond with only one word: 'cartoon' or 'real'
    """
```

Figure 6: Prompt for Character & Face Consistency

```
CHARACTER_CLOTHING_CONSISTENCY_PROMPT = """
    Please analyze these two images based on the character description: "{character_prompt}"
    This is a two-step analysis:

    **STEP 1: Character Existence Check**
    First, determine if the described character exists in BOTH images.
    - If the character exists in both images: Continue to Step 2
    - If the character is missing from either image: Give 0 points for final score and stop
analysis

    **STEP 2: Clothing and Shoes Consistency Analysis**
    If the character exists in both images, analyze the consistency of their clothing and shoes.

    **IMPORTANT: Consider perspective and viewing angle changes**
    - The same clothing item may appear different from front vs back view
    - Different camera angles can change how colors, patterns, and details appear
    - Lighting conditions may affect color perception
    - Focus on identifying whether it's the SAME clothing item despite these variations

    Scoring criteria for Step 2 (1-5 points):
    - 5 points: Completely identical - Same clothing item, colors and styles match perfectly
(accounting for perspective)
    - 4 points: Very similar - Same clothing item, colors and styles are basically the same, minor
differences due to angle/lighting/perspective
    - 3 points: Quite similar - Likely the same clothing item, but some uncertainty due to
perspective or visible differences
    - 2 points: Partially similar - Could be the same item but significant differences make it
unclear
    - 1 point: Basically different - Different clothing items, though similar type
    - 0 points: Character missing from either image

    Please analyze in detail:
    1. Character existence: Does the described character appear in both images?
    2. Clothing analysis: Compare the character's clothing considering possible perspective
changes (front/back/side views)
    3. Shoes analysis: Compare the character's shoes considering possible viewing angle
differences
    4. Perspective considerations: How might different viewing angles affect the appearance of
the same items?
    5. Overall consistency: Final judgment accounting for perspective variations

    Please output in the following format:
    Character Existence: [Yes/No - exists in both images]
    Final Score: X points (0 if character missing from either image, 1-5 if character exists in
both)
    Clothing Analysis: [Detailed comparison of clothing features, considering perspective
changes]
    Shoes Analysis: [Detailed comparison of shoes features, considering viewing angles]
    Overall Judgment: [Final reasoning for the consistency score, accounting for perspective
variations]
    """
```

Figure 7: Prompt for Clothes & Color Consistency

```
CHARACTER_RELATIVE_SIZE_ANALYSIS_PROMPT = """
    Analyze this video and determine the relative size ratios between these character pairs:
{pairs_list}

    IMPORTANT ANALYSIS REQUIREMENTS:
    1. Examine the entire video carefully, not just single frames
    2. Consider multiple appearances of characters throughout the video
    3. Account for perspective effects (closer objects appear larger)
    4. Account for distance from camera (farther objects appear smaller)
    5. Consider natural size differences between character types
    6. Look for consistent size relationships across different scenes/shots
    7. Ignore temporary perspective distortions (extreme close-ups, etc.)
    8. Focus on the characters' actual relative sizes when they appear together

    For each character pair that appears together in the video:

    PAIR: [character1 to character2]
    RATIO: [precise numerical ratio - how many times larger character1 is than character2]
    EXPLANATION: [detailed explanation of size relationship, considering: natural size
differences, perspective effects, distance from camera, consistency across scenes, artistic
choices]
    CONFIDENCE: [0.0-1.0 based on clarity of size relationship and number of observations]

    RATIO CALCULATION GUIDELINES:
    - 1.0 means characters are the same size
    - 1.5 means character1 is 1.5 times larger than character2
    - 0.7 means character1 is 0.7 times the size of character2 (smaller)
    - Consider the characters' body sizes, not temporary positioning
    - Average the size relationship across multiple appearances
    - Account for natural proportions (e.g., adult vs child, different species)

    If characters don't appear together clearly enough to judge size:
    PAIR: [character1 to character2]
    RATIO: N/A
    EXPLANATION: Characters not visible together with sufficient clarity
    CONFIDENCE: 1.0

    Be precise with numerical ratios and provide detailed explanations for your assessments.
    """
```

Figure 8: Prompt for Relative Size Consistency

```
SCENE_CHARACTER_ACTION_ANALYSIS_PROMPT = """
    Analyze this video and the following scene description:
    Scene Description: {scene_description}

    Please determine:
    1. How many distinct MAIN CHARACTERS (protagonists) are performing clear, identifiable
actions in this video?
    2. Are these main characters performing independent actions (not interacting directly with
each other or shared objects)?
    3. If there are multiple independent main characters, what are their individual target
actions?

    Respond in the following JSON format:
    {{
      "character_count": <number>,
      "need_individual_tracking": <true/false>,
      "characters": [
        {{
          "description": "<character description for detection>",
          "target_action": "<specific action this character is performing>"
        }}
      ],
      "overall_target_action": "<overall scene action if not using individual tracking>"
    }}

    Guidelines:
    - ONLY include MAIN CHARACTERS (protagonists) - typically 1-2 characters who are the
focus of the scene
    - EXCLUDE background characters, extras, or people without clear/significant actions
    - AVOID duplicates - each character should be unique and distinct
    - Set need_individual_tracking to true only if there are more than one main characters AND
they are performing independent actions
    - Character descriptions should be simple for object detection (e.g., "person", "woman",
"child", "man")
    - Target actions should be specific action verbs (e.g., "eating", "walking", "sitting", "reading")
    - If a character only appears in background or has no clear action, DO NOT include them
    """
```

Figure 9: Prompt for Action Recognition

```
CAMERA_MOTION_EVALUATION_PROMPT = """
    You are a professional cinematographer evaluating camera motion quality. Please score how
well the actual camera motion matches the intended camera motion instruction.

    CAMERA MOTION INSTRUCTION (What was expected): {instruction}

    ACTUAL CAMERA MOTION ANALYSIS (What actually happened): {analysis}

    SCORING CRITERIA:
    - Score 0: No motion when motion was expected, or motion completely opposite to
instruction
    - Score 1: Very poor match - motion exists but wrong direction/type
    - Score 2: Poor match - some similarity but significant differences
    - Score 3: Fair match - generally correct but with notable deviations
    - Score 4: Good match - mostly correct with minor differences
    - Score 5: Excellent match - motion perfectly matches the instruction

    Please provide:
    1. A score from 0-5 (integer only)
    2. A detailed explanation of your scoring reasoning

    Format your response as:
    SCORE: [0-5]
    EXPLANATION: [Your detailed reasoning]
    """
```

Figure 10: Prompt for Camera Control

```
CHARACTER_MOTION_EXTRACTION_PROMPT = """
    Analyze the following prompt and identify ALL moving characters and their expected motions:
    Prompt: "{prompt}"
    Please identify:
    1. List ALL characters that should be moving (humans, animals, creatures, etc.)
    2. For each character, describe their expected motion type and physical requirements
    3. Key physical principles that should govern each character's movement
    Format your response as:
    CHARACTER_COUNT: [total number of moving characters]
    CHARACTER_1: [name/type of first character]
    MOTION_1: [describe expected motion and physical requirements]
    PHYSICS_1: [key physical laws that should apply]
    .....
    Focus on realistic movement patterns, balance, gravity effects, momentum, etc.
    If no moving characters are expected, state CHARACTER_COUNT: 0
    """

CHARACTER_PHYSICS_EVALUATION_PROMPT = """
    Analyze this video focusing on the physical plausibility of this specific character's movement:
    Character: {character}
    Expected Motion: {expected_motion}
    Physics Principles: {physics_principles}
    Evaluate how well this character's movement follows realistic physics:
    1. Is the character "{character}" present and moving in the video?
    2. Does their movement respect gravity, momentum, and balance?
    3. Are their body mechanics and motion dynamics realistic?
    4. Do they move in a way that follows natural physics laws?

    Give a score from 0-5 based on these detailed criteria:
    0 = Character not found in video
    - The specified character is completely absent from the video
    - Cannot locate the character in any frame
    1 = Character present but not moving
    - The specified character is visible but completely static
    - No movement to evaluate for physics
    2 = Character moving but mostly violates physics
    - Movement violates basic physical laws (gravity, momentum, etc.)
    - Character floats, moves through solid objects, or defies physics
    - Motion is mostly unrealistic and impossible
    3 = Character moving with partially correct physics
    - Basic physics followed but with noticeable inaccuracies
    - Some movements realistic, others clearly wrong
    - Mixed realistic and unrealistic motion patterns
    4 = Character moving with mostly correct physics
    - Movement is largely realistic with minor physics issues
    - Generally natural motion with small deviations
    - Overall convincing but some details feel slightly off
    5 = Character moving with completely realistic physics
    - All movement follows natural physics perfectly
    - Realistic body mechanics, gravity effects, momentum
    - Motion appears completely natural and believable

    Format response as: .....
    """
```

Figure 11: Prompt for Physical Plausibility

```
INTERACTION_EXTRACTION_PROMPT = """
    Analyze the following prompt and identify the MOST IMPORTANT physical interaction:
    Prompt: "{prompt}"
    Please identify:
    1. The main object that characters interact with (choose only ONE most important object)
    2. The expected motion or deformation of this object based on the interaction
    3. Key physical principle that should govern this interaction
    Format your response as:
    MAIN_OBJECT: [the single most important interacting object]
    EXPECTED_MOTION: [describe the expected motion/deformation of this object]
    PHYSICAL_PRINCIPLE: [key physical law or realistic behavior]
    Be specific about the type of motion (rotation, translation, deformation, etc.) and direction.
    Focus on the most critical interaction only.
    """

INTERACTION_VERIFICATION_PROMPT = """
    Based on this analysis of expected interaction:
    Main Object: {main_object} Expected Motion: {expected_motion}
    Physical Principle: {physical_principle}
    Watch this video and evaluate:
    1. Is the main object "{main_object}" present in the video?
    2. If present, how accurately does its motion/deformation match the expected behavior?
    3. Does the interaction follow the physical principle described?
    Give a score from 0-5 based on these detailed criteria:
    0 = Main object not found in video
    - The specified object is completely absent from the video
    - Example: Looking for a "ball" but no ball appears in any frame
    1 = Object present but motion completely incorrect
    - Object exists but behaves in a way that violates basic physics
    - Motion direction is opposite to expectation
    - Object ignores fundamental forces (gravity, momentum, etc.)
    2 = Object present with mostly incorrect motion
    - Object shows some realistic behavior but major aspects are wrong
    - Motion direction may be correct but speed/acceleration is highly unrealistic
    - Some physical principles followed but others completely ignored
    3 = Object present with partially correct motion
    - Object follows basic physical principles but with noticeable inaccuracies
    - Motion is generally in the right direction but with timing or magnitude issues
    - Some aspects realistic, others clearly wrong
    4 = Object present with mostly correct motion
    - Object behavior is largely realistic with minor physics violations
    - Motion follows expected patterns with small deviations
    - Overall convincing but some details feel off
    5 = Object present with completely correct motion
    - Object behavior is physically accurate and realistic
    - Motion perfectly matches expectations based on real-world physics
    - All aspects of the interaction appear natural and convincing
    Format response as:
    OBJECT_PRESENT: [Yes/No]
    MOTION_ACCURACY: [detailed analysis of the object's motion with specific observations]
    SCORE: [0-5]
    JUSTIFICATION: [explanation for the score with reference to the criteria above]
    """
```

Figure 12: Prompt for Physical Interaction Accuracy

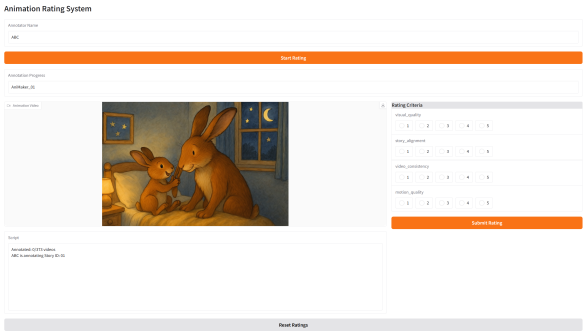

Figure 13: Human Rating System Interface.

B Human Rating Details

To evaluate the perceptual quality of the generated videos and validate the alignment between MSVBench and human judgments, we conduct a comprehensive human evaluation involving 10 participants with animation backgrounds. We assess storytelling videos synthesized by 20 models based on MSVBench scripts. As illustrated in Figure 13, we employ a specialized annotation system and establish rigorous rating guidelines, incorporating pre-evaluation training to ensure inter-rater reliability. Evaluators rate each video on a 5-point scale across four dimensions—Visual Quality, Story Video Alignment, Video Consistency, and Motion Quality—following the specific instructions detailed below.

B.1 Visual Quality

Evaluators are instructed to rate the overall aesthetic appeal and technical fidelity of the video content. The focus is on image clarity, lighting distribution, and stylistic unity. Raters must scrutinize the video for technical imperfections, identifying issues such as blurriness, noise, severe distortion, or structural collapse typical of generative artifacts.

- **1 (Unusable):** The video is filled with severe noise, distortion, or artifacts. Lighting is chaotic, and the subject or the scene is totally unrecognizable.
- **2 (Poor):** Visible artifacts (e.g., limb distortion, abnormal color blocks) are present. The image is blurry, or the style fluctuates violently, causing visual discomfort.
- **3 (Fair):** The image is generally clear and recognizable. However, there are localized issues with lighting (e.g., slight overexposure), minor discontinuities between shots, or slight stylistic dissonance between the subject and background.
- **4 (Good):** The video is clear with good aesthetic composition and natural lighting. Minor edge flickering or background blur exists but does not affect the overall viewing experience.
- **5 (Excellent):** The video exhibits cinema-grade illustrative aesthetics with rich lighting layers and a highly unified style. No obvious generation traces are visible.

B.2 Story Video Alignment

This dimension requires evaluators to determine the semantic correspondence between the generated video and the provided scripts. The assessment involves verifying whether the described characters, objects, and environments are present and possess the correct attributes (e.g., color, quantity). Additionally, evaluators check for adherence to specific camera language (e.g., close-up, overhead view) mentioned in the script.

- **1 (Irrelevant):** The content is unrelated to the prompt, exhibiting severe hallucinations (e.g., generating a city instead of a forest).
- **2 (Severe Mismatch):** The general environment is correct, but the core subject is missing, or fundamental attributes are wrong (e.g., a blue car instead of a red one).
- **3 (Partial Deviation):** The core subject and environment are correct, but there are errors in camera shot types or missing details.
- **4 (High Fidelity):** All major visual elements and attributes are accurately presented. Minor discrepancies are permitted in ambiguous details not explicitly described in the text.
- **5 (Precise Match):** The video is a precise visual translation of the script. All objects, colors, spatial relationships, and specific camera angles are perfectly executed.

B.3 Video Consistency

Evaluators assess the temporal coherence and stability of visual elements across consecutive frames and shots. The primary task is to detect any unnatural morphing or illogical discrepancies in character identities (appearance, clothing), scene layouts, and relative object scales, to determine the extent

to which visual narrative continuity is maintained throughout the sequence.

- **1 (Incoherent):** Character identity or scene layout completely changes after a shot transition, making it impossible to establish narrative continuity.
- **2 (Disjointed):** The character is recognizable between shots, but there are sudden changes in clothing color, severe facial deformation, or jumps in scene style.
- **3 (Average):** Character identity is generally preserved, but details (e.g., patterns, accessories) flicker or change during side profiles, long shots, or large movements.
- **4 (Stable):** Characters and scenes remain stable in most shots. Slight differences are acceptable only under extreme camera angles or dramatic lighting changes.
- **5 (Consistent):** Character features and spatial relationships remain stable throughout the video. There are no perceptible changes in relative scale, maintaining logical consistency regardless of rendering variations.

B.4 Motion Quality

This criterion evaluates the naturalness, physical plausibility, and semantic accuracy of the depicted dynamics. Evaluators examine whether the actions align with the scripts, if the movements are fluid (free from stiffness or unnatural jitter), and if physical interactions (e.g., gravity, collisions) and camera movements adhere to real-world physical laws.

- **1 (Incorrect / Broken):** Actions are incorrect (e.g., standing still instead of dancing) or violate physics (e.g., floating objects, severe clipping, non-human distortion).
- **2 (Stiff / Unnatural):** The execution deviates significantly from the semantic intent (e.g., mere mouth opening instead of laughter) or exhibits excessive rigidity, resembling a static image translating across the frame.
- **3 (Weak / Incomplete):** The action category is correct, but the magnitude is too small, lacks weight, or is incomplete. Complex interactions often result in clipping.
- **4 (Correct):** Actions largely align with the description; character movement and camera movements are smooth and adhere to ergonomics. Physics are generally correct, with minor flaws allowed only in high-speed or complex interactions.
- **5 (Vivid):** Actions are perfectly aligned with the text and visually impactful. Camera movements are cinematic and stable. Interactions possess real weight and physical feedback.

C Quantitative Results

We present the visualization results for two stories from MSVBench in Figure 14 and Figure 15. To ensure a concise yet representative comparison, we display five sampled shots for each story, showcasing the generation results from all 20 evaluated methods. These comparisons highlight performance variations regarding Visual Quality, Story Video Alignment, Video Consistency, and Motion Quality. This detailed visualization provides an intuitive understanding of each method's capability in multi-shot video generation.

Figure 14: Visualization results of Story 2 from MSVBENCH. (Continued on next page)

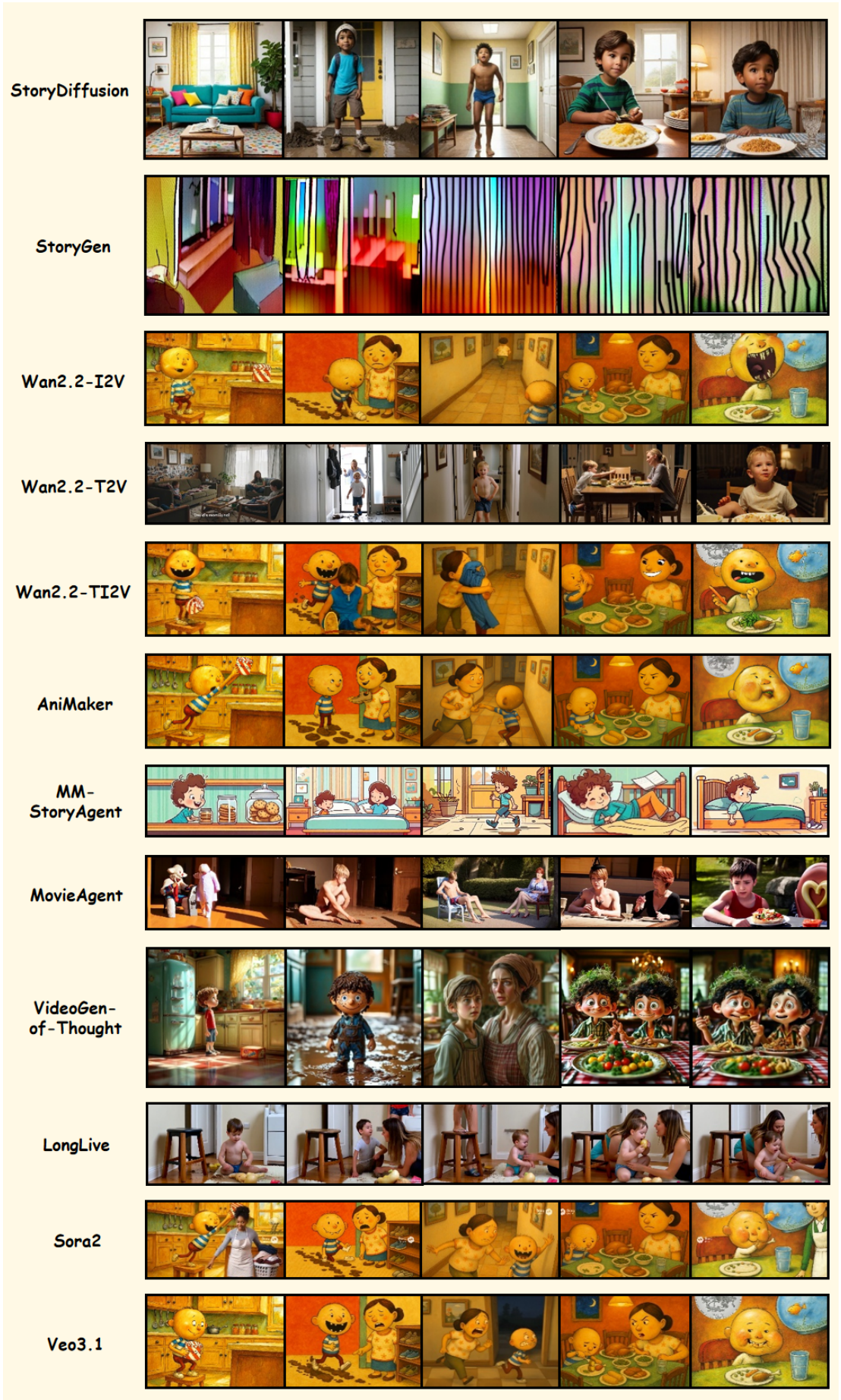

Figure 14: (Continued)

Figure 15: Visualization results of Story 14 from MSVBENCH. (Continued on next page)

Figure 15: (Continued)

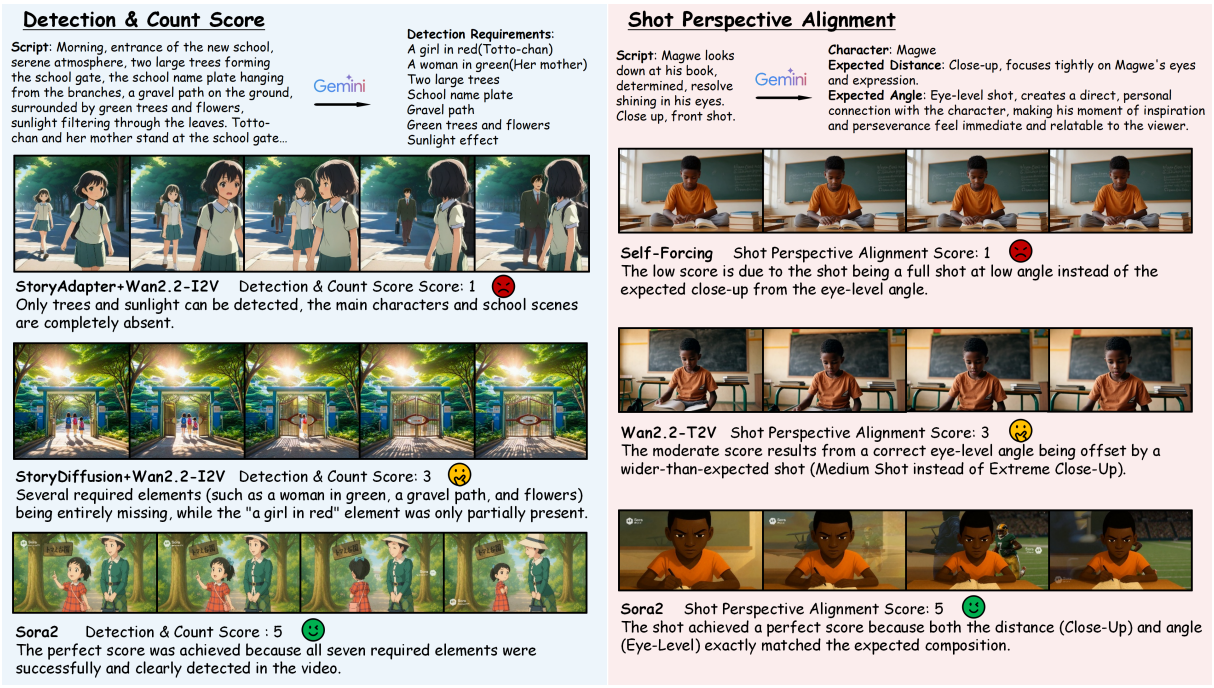

Figure 16: Case Study on **Detection & Count Score** (Left) and **Shot Perspective Alignment** (Right).

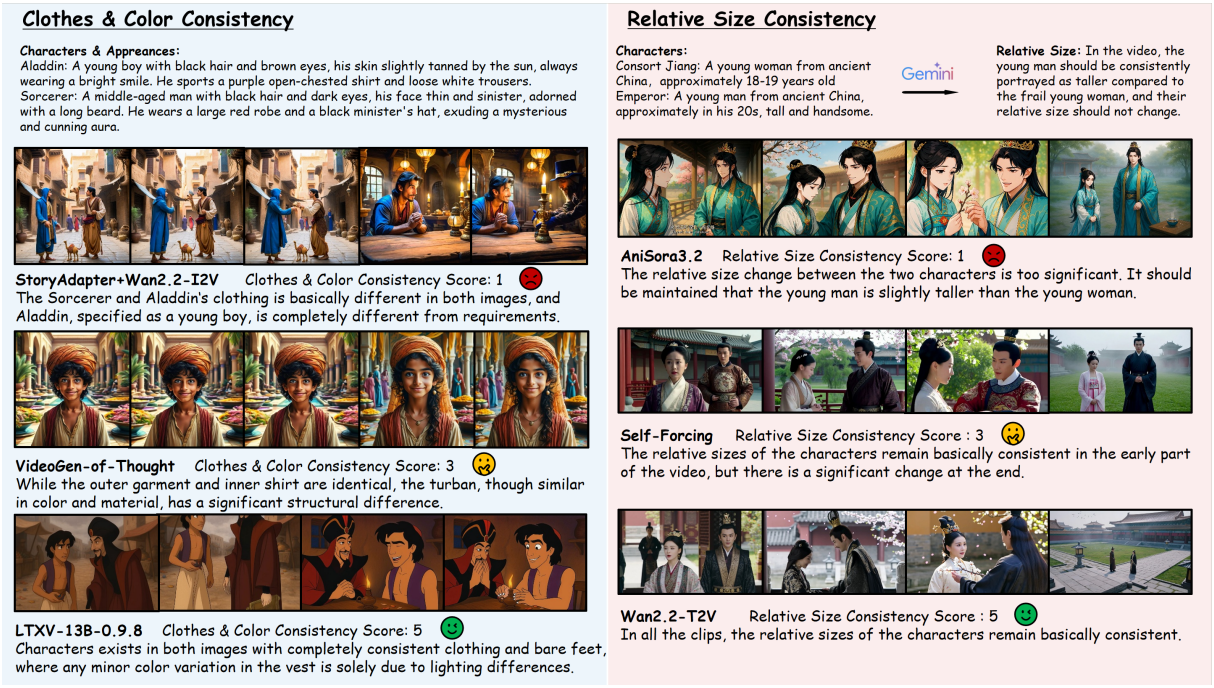

Figure 17: Case Study on **Clothes & Color Consistency** (Left) and **Relative Size Consistency** (Right).

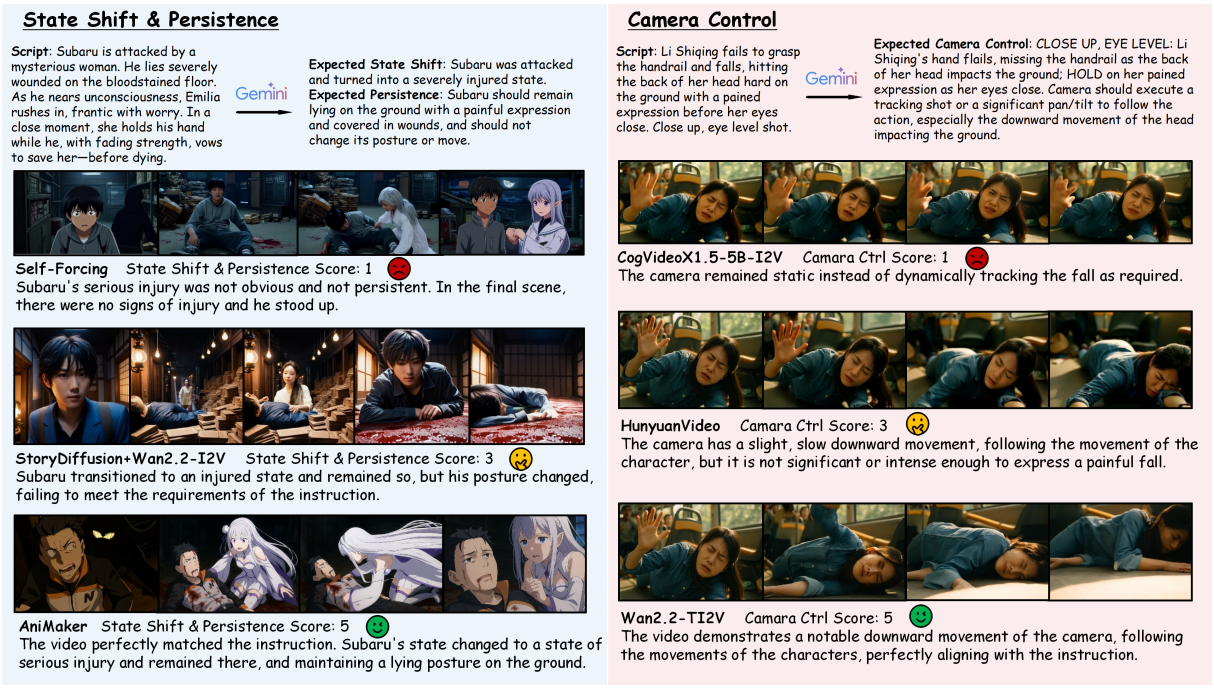

Figure 18: Case Study on **State Shift & Persistence** (Left) and **Camera Control** (Right).

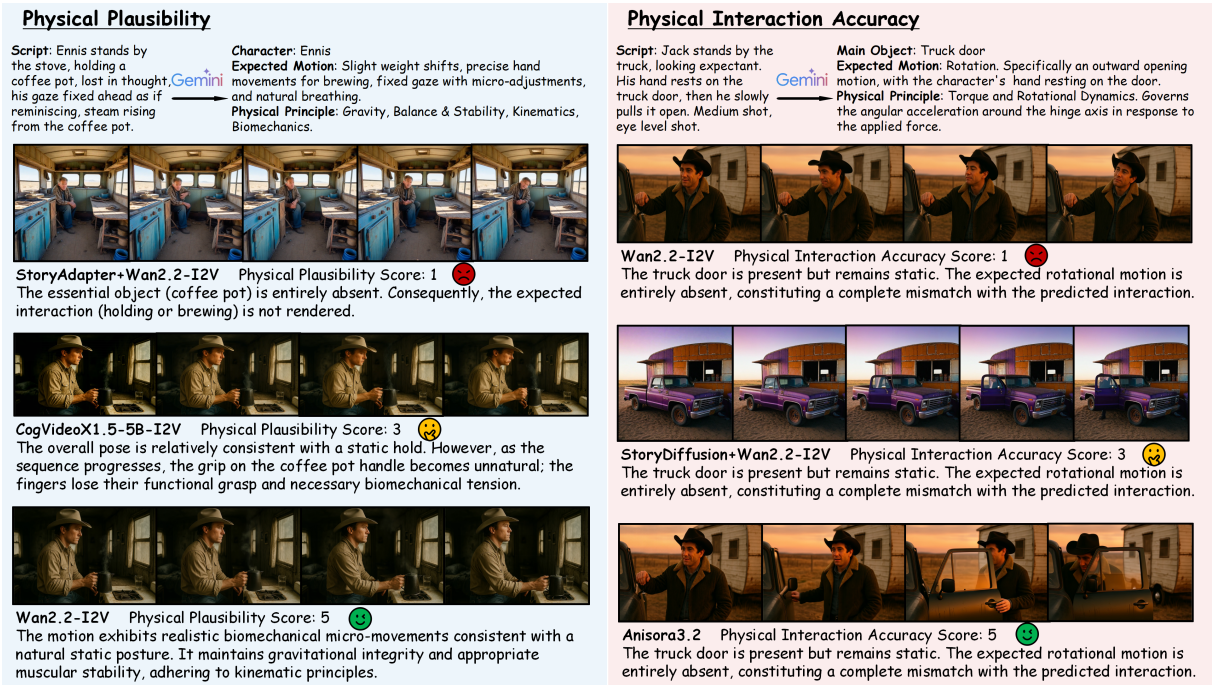

Figure 19: Case Study on **Physical Plausibility** (Left) and **Physical Interaction Accuracy** (Right).

D Case Study

To explicitly demonstrate the effectiveness and discriminative capability of our proposed MSVBench metrics, we present detailed case studies across all LMM based metrics. Figures 16 through 19 visualize how the LMM extracts semantic requirements from the input scripts and assesses the generated videos.

D.1 Story Video Alignment Assessment

Figure 16 illustrates the evaluation of **Detection & Count Score** and **Shot Perspective Alignment**. For detection, the LMM first parses the script to identify necessary entities (e.g., "A girl in red," "School name plate"). It then verifies both the presence and the precise quantity of these elements in each video; for instance, it assigns a low score to StoryAdapter (Score 1) for failing to generate required characters, while confirming that Sora2 (Score 5) achieves perfect recall of all mandated elements. For cinematography, the evaluator assesses the realized shot against the intended camera language without reference to other models. As shown, it accurately penalizes Self-Forcing (Score 1) for generating a full shot instead of the required close-up, while validating Sora2's (Score 5) perfect adherence to the "Close-Up" and "Eye-Level" constraints.

D.2 Video Consistency Assessment

Figure 17 demonstrates the **Clothes & Color Consistency** and **Relative Size Consistency** metrics. In multi-shot generation, maintaining visual attributes is critical. The LMM identifies specific violations in individual model outputs, such as the structural change in the turban in VideoGen-of-Thought (Score 3) or complete clothing mismatches in StoryAdapter (Score 1). Conversely, it acknowledges the consistent output generated by LTXV (Score 5). Furthermore, the Relative Size metric detects logical inconsistencies in character proportions across shots, such as penalizing AniSora3.2 (Score 1) for significantly altering the height difference between the protagonist and the consort.

D.3 Complex Logic and Camera Control

Figure 18 showcases the evaluation of temporal logic via **State Shift & Persistence** and **Camera Control**. The State Shift metric evaluates whether narrative-driven changes (e.g., an injury) persist logically within a specific video. The LMM successfully identifies that Self-Forcing (Score 1) violates the narrative when the character "Subaru" stands up uninjured in the final scene, whereas it recognizes that AniMaker (Score 5) successfully maintains the "severely injured" state throughout. Simultaneously, the Camera Control metric assesses dynamic movement against the prompt. The evaluator assigns a low score to CogVideoX (Score 1) for producing a static shot, while separately validating the complex tracking shot in Wan2.2-TI2V (Score 5), demonstrating its sensitivity to motion dynamics.

D.4 Motion Quality and Physics

Figure 19 focuses on **Physical Plausibility** and **Physical Interaction Accuracy**. These metrics assess adherence to real-world physics for each generation. For plausibility, the LMM detects subtle biomechanical issues, such as the unnatural grip on a coffee pot in CogVideoX (Score 3), and separately confirms the natural static posture in Wan2.2 (Score 5). Regarding interactions, the system verifies causal logic; for instance, it penalizes Wan2.2-I2V (Score 1) where a "truck door" remains static despite a prompt explicitly describing it opening, while recognizing the correct rotational dynamics in other successful generations like AniSora3.2 (Score 5).

Collectively, these cases above confirm that MSVBench provides fine-grained, objective scores for each model. By evaluating generated content strictly against the input script and physical principles rather than performing relative rankings, the framework effectively penalizes hallucinations, continuity errors, and physical violations.

E Correlation Coefficient Details

To evaluate the alignment between the proposed MSVBench metrics and human perception, we compute correlation coefficients between objective metric scores and human ratings. This involves three steps: (1) metric aggregation, which normalizes raw sub-metrics into ranks to compute dimension-level scores via negative averaging; (2) statistical analysis, utilizing Spearman's rank correlation and Kendall's rank correlation coefficients to quantify the alignment between MSVBench and human ratings; and (3) overall score calculation, which derives a unified model ranking by averaging the scores across the four evaluation dimensions.

E.1 Metric Aggregation

Given that MSVBench sub-metrics vary widely in scale and units, we normalize them by ranking. For N models, let $v_{i,k}$ denote the raw value of the k-th sub-metric for model i. Each value is converted into a rank $r_{i,k} \in [1, N]$, where $r = 1$ represents the best performance. If multiple models share the same value, we assign them the average rank of the tied group.

For a dimension (e.g., Visual Quality) with K sub-metrics, the aggregated dimension-level score for model i is:

$$M_i = -\frac{1}{K} \sum_{k=1}^{K} r_{i,k}, \tag{1}$$

where the negative sign ensures that higher-quality models (with smaller ranks) receive larger M_i values, aligning the rating direction with human evaluation scores.

E.2 Statistical Analysis

We employ two non-parametric rank-based correlation coefficients to quantify the alignment between MSVBench scores and human ratings. Spearman's rank correlation (ρ) is utilized to assess the global monotonic consistency between the two ranking lists, reflecting how well the overall ranking trend matches human perception. Complementarily, Kendall's rank correlation (τ) focuses on the accuracy of pairwise relative orderings, providing a more rigorous evaluation of measuring ordinal association and robustness against ranking noise.

Spearman's Rank Correlation (ρ): Let $\hat{r}_{m,i} = R(M_i) - \bar{R}_M$ and $\hat{r}_{h,i} = R(H_i) - \bar{R}_H$ denote centered ranks for MSVBench and human scores, respectively. The coefficient is:

$$\rho = \frac{\sum_i \hat{r}_{m,i}\, \hat{r}_{h,i}}{\sqrt{\sum_i \hat{r}_{m,i}^2} \sqrt{\sum_i \hat{r}_{h,i}^2}}. \tag{2}$$

A high ρ indicates consistent ordering between MSVBench and human judgments.

Kendall's Rank Correlation (τ): Using pairwise consistency,

$$\tau = \frac{N_c - N_d}{\sqrt{(N_0 - N_1)(N_0 - N_2)}}, \tag{3}$$

where N_c and N_d are concordant and discordant pairs, $N_0 = \frac{N(N-1)}{2}$ is the total number of model pairs, and N_1, N_2 correct for ties in MSVBench and human scores, respectively. This formulation ensures robust ranking comparison when ties occur.

Table 7: MSVBench Leaderboard: Overall average ranking of all evaluated methods.

Rank	Model	Overall Ranking$^{\downarrow}$
1	Veo3.1	5.60
2	Sora2	6.67
3	Wan2.2-TI2V	7.50
4	AniSora3.2	7.90
5	AniMaker	8.00
6	Wan2.2-T2V	8.12
7	Wan2.2-I2V	8.45
8	Self-Forcing	8.57
9	LTXV-13B-0.9.8	9.15
10	HunyuanVideo-I2V	9.60
11	CogVideoX1.5-5B-I2V	9.87
12	LongLive	10.21
13	S.D.+Wan2.2-I2V	10.95
14	S.A.+Wan2.2-I2V	11.05
15	VideoGen-of-Thought	11.42
16	StoryDiffusion	12.17
17	StoryAdapter	12.42
18	MovieAgent	15.19
19	MM-StoryAgent	15.30
20	StoryGen	16.70

E.3 Overall Score Calculation

Besides dimension-level evaluation, MSVBench provides a unified *overall score* for each model. For each model i, let M_i^{VQ}, M_i^{SA}, M_i^{VC}, and M_i^{MQ} denote the aggregated (negative) rank scores of the four MSVBench dimensions. The overall score, also referred to as the model's **Average Rank**, is defined as the mean of these four dimension-level rank scores:

$$R_i^{\text{overall}} = -\frac{1}{4}\left(M_i^{\text{VQ}} + M_i^{\text{SA}} + M_i^{\text{VC}} + M_i^{\text{MQ}}\right). \tag{4}$$

Table 7 presents the MSVBench Leaderboard. Notably, the derived overall ranking demonstrates strong alignment with human perception, validating the reliability of our MSVBench.