Rethinking Video Generation Model for the Embodied World

Yufan Deng [1 2 *] Zilin Pan [1 *] Hongyu Zhang [1 *] Xiaojie Li [2] Ruoqing Hu [2]
Yufei Ding [1] Yiming Zou [1] Yan Zeng [2] Daquan Zhou [1]

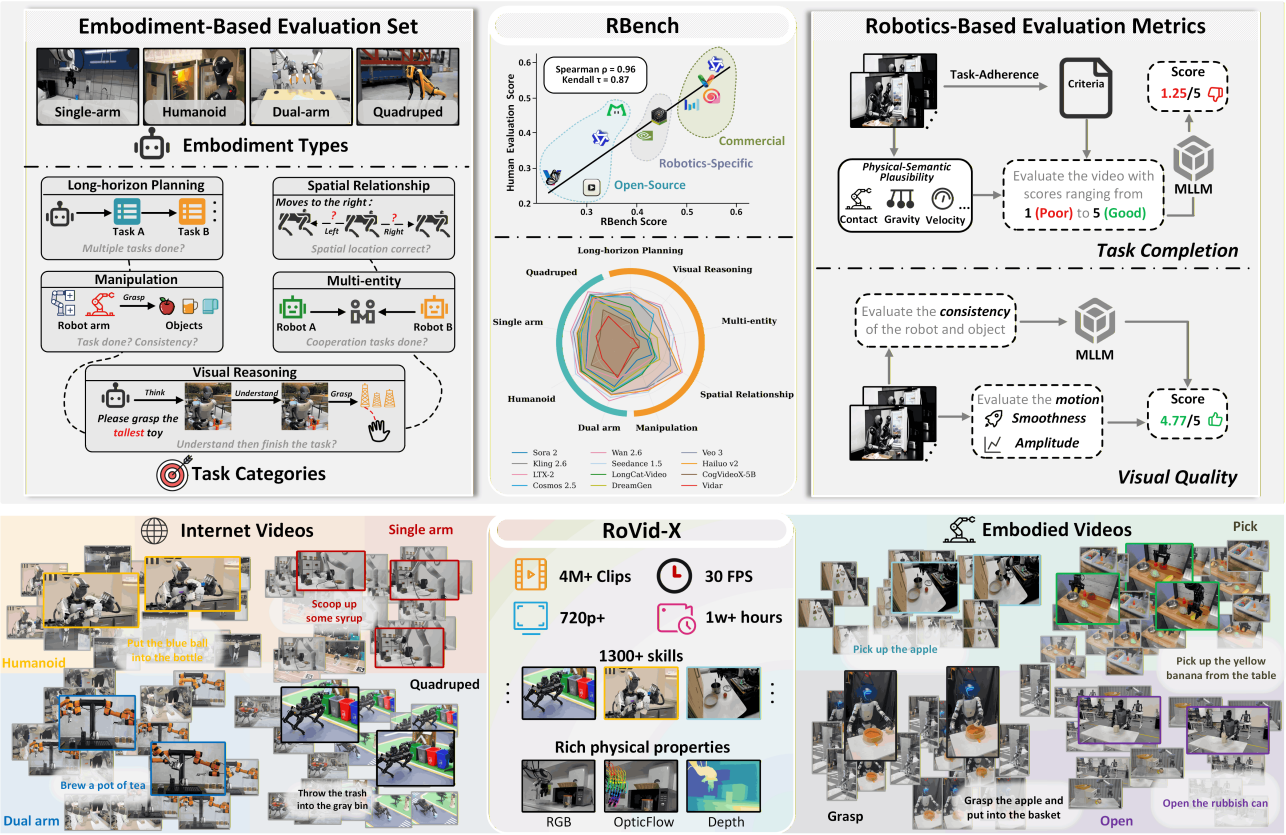

Figure 1. Overview of the comprehensive robotics benchmark and dataset for video generation. **Top:** We present **RBench** that includes the embodiment-based evaluation set and automated evaluation metrics. Our evaluation results on 25 video generation models show a high level of agreement with subjective human assessments. **Bottom:** We introduce a large-scale high-quality robotic dataset (**RoVid-X**) specifically designed for training video generation models, with data sourced from internet videos and open-source embodied videos.

Abstract

While video generation holds promise for embodied intelligence, current video models struggle with physical realism, and progress is hindered by the lack of standardized benchmarks. To address this gap, we introduce a comprehensive robotics benchmark, **RBench**, designed to evaluate robot-oriented video generation across five task domains and four distinct embodiments. By assessing task correctness and visual fidelity through reproducible metrics, our evaluation of 25 models reveals significant deficiencies in generating physically realistic robot behaviors. Furthermore, the benchmark achieves a 0.96 Spearman correlation with human judgment, validating its effectiveness. While RBench provides the necessary lens to identify these deficiencies, achieving physical realism requires moving beyond evaluation to address the critical shortage of high-quality training data. Driven by these insights, we introduce a refined four-stage data pipeline, resulting in **RoVid-X**,

[*]Equal contribution [1]Peking University Shenzhen Graduate School [2]ByteDance Seed, Beijing, China. Correspondence to: Daquan Zhou <zhoudaquan21@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

the largest open-source robotic dataset for video generation with 4 million annotated video clips, covering thousands of tasks and enriched with physical property annotations. Extensive experiments demonstrate that finetuning on RoVid-X yields consistent performance gains. Collectively, this synergistic ecosystem of evaluation and data establishes a robust foundation for rigorous assessment and scalable training of video models, accelerating the evolution of embodied AI toward physical intelligence. The code and video demos are available in the supplementary materials.

1. Introduction

Recent breakthroughs in diffusion models (Ho et al., 2020; Song et al., 2020; Peebles & Xie, 2023) and video generation (Runway, 2025; Pika, 2025; Guo et al., 2024a; Wu et al., 2025a; Wan et al., 2025) have enabled advanced applications in video editing, motion control, and multi-subject synthesis (Jiang et al., 2025c; Ju et al., 2025; Deng et al., 2025a;b; Wang et al., 2024). As video models evolve into unified foundation models for machine vision (Wiedemer et al., 2025), they demonstrate robust generalization across 3D scenes (Kim et al., 2025; Ren et al., 2025), autonomous driving (Gao et al., 2025a; Yan et al., 2025), and world modeling (Kang et al., 2024; Ball et al., 2025). Crucially, these models increasingly serve as controllable simulators for robot learning and action prediction (Guo et al., 2024b; Zhu et al., 2025; Hu et al., 2024; Zhen et al., 2025; Guo et al., 2025c; Liang et al., 2025), synthesizing robotic trajectories to mitigate the scarcity of human teleoperation data (Jang et al., 2025; Bjorck et al., 2025; Team et al., 2025a). Collectively, these advancements highlight the potential of video generation models in the perception-reasoning-action loop, paving the way for generalizable embodied intelligence.

Despite these strides, systematic evaluation for robotic video generation remains underdeveloped. Current practices primarily rely on perceptual metrics focused on visual quality (Huang et al., 2024; Liu et al., 2024; Han et al., 2025), while physics-based benchmarks often lack task-specific datasets and criteria (Meng et al., 2024a; Guo et al., 2025b; Zhang et al., 2025). Consequently, critical aspects such as task completion, action-goal alignment, and physical feasibility are frequently overlooked, leading to overly optimistic conclusions where high scores are assigned to videos with unnatural movements or incomplete tasks. To address this, evaluation protocols must go beyond perceptual metrics, incorporating physical plausibility and alignment with input instructions for rigorous, reproducible assessments.

To address these challenges, we propose **RBench**, a benchmark designed to evaluate the fidelity and utility of robotic video generation models. It is the first comprehensive benchmark with fine-grained metrics, consisting of 650 image–text pairs across five task categories and four robot types. Evaluations focus on *task completion* and *visual quality*, incorporating sub-metrics like structural consistency, physical plausibility, and execution completeness. Based on RBench, our assessment of 25 representative models reveals a persistent gap between general video models and the requirements of embodied tasks, underscoring the need for advancements in both robotic video data and training methodologies.

Advancing robotic video generation requires diverse and scalable training data (O'Neill et al., 2024; Bu et al., 2025). However, robotic interaction data remains constrained by scale and diversity compared to vision or language domains (Brohan et al., 2022; Zhao et al., 2025; Fourier ActionNet Team, 2025). Existing datasets, often with narrow distributions in environments and morphologies (Yang et al., 2025; Wang et al., 2023), **are primarily curated for training VLM or VLA models, leaving a void for datasets specifically processed for video generation.** To bridge these gaps, we integrate over 20 open-source datasets and multi-source platforms to introduce **RoVid-X**, a large-scale embodied dataset comprising 4 million annotated clips. RoVid-X is developed via a four-stage pipeline consisting of robot video collection, video quality filtering, task segmentation and captioning, and physical property annotation. As the largest dataset tailored for video generation (see Table 1), it provides the physical priors and semantic diversity essential for advancing pretrained video foundation models.

Overall, the main contributions are summarized as follows:

- A systematic benchmark tailored for robotic video generation. We propose **RBench**, which comprehensively evaluates the performance of video generation models across five robotic tasks and four robot types with 650 meticulously curated evaluation samples, while introducing reproducible automated evaluation metrics.
- Key insights into video generation for embodied research. We conduct a systematic evaluation of 25 representative video models, including open-source, commercial, and robotics-specific ones, revealing the limitations of current video foundation models and potential directions for improvement, offering new perspectives for researchers exploring the embodied domain using video world models.
- A large-scale, multi-embodiment robotic dataset for video generation. We construct **RoVid-X**, the first dataset specifically curated for training robotic video models. It comprises 4 million clips featuring diverse task descriptions and fine-grained physical property annotations, providing essential interaction priors for embodied intelligence.

Conflict of Interest Disclosure. Some authors are affiliated with ByteDance Seed, which develops Seedance, one of the models evaluated in this paper. We disclose this re-

Table 1. **Comparison of representative robotic datasets.** Unlike existing datasets designed for VLM/VLA training, RoVid-X is specifically curated for pretraining robotic video generation models, with no action labels preserved for each video, as the focus is on enhancing the physical performance of pretrained video generation models for better downstream applications in the embodied domain.

Dataset	Year	#Videos	#Skills	Resolution	Optical Flow	Diverse Robotic Forms	Diverse Captions
RoboTurk (Mandlekar et al., 2018)	2018	2.1k	2	480P	✗	✗	✗
RoboNet (Dasari et al., 2019)	2019	162k	N/A	240P	✗	✗	✗
BridgeData (Ebert et al., 2021)	2021	7.2k	4	480P	✗	✗	✗
RH20T (Fang et al., 2023)	2023	13k	33	720P	✗	✗	✗
DROID (Khazatsky et al., 2024)	2024	76k	86	720P	✗	✗	✗
Open X-Embodiment (O'Neill et al., 2024)	2024	1.4M	217	64P—720P	✗	✓	✗
RoboMIND (Wu et al., 2024a)	2024	107k	38	480P	✗	✗	✗
RoboCOIN (Wu et al., 2025c)	2025	180k	36	480P	✗	✗	✗
Galaxea (Jiang et al., 2025a)	2025	100k	58	720P	✗	✗	✗
InternData-A1 (Tian et al., 2025)	2025	630k	18	480P	✗	✗	✗
Fourier ActionNet (Fourier ActionNet Team, 2025)	2025	13k	16	800P	✗	✗	✗
Humanoid Everyday (Zhao et al., 2025)	2025	10.3k	221	320P—720P	✗	✗	✗
Agibot World (Bu et al., 2025)	2025	1M	87	480P	✗	✗	✗
RoVid-X (Ours)	2026	4M	1300+	720P	✓	✓	✓

lationship for transparency. All experiments, metrics, and analyses are designed and reported to support fair, reproducible, and model-agnostic evaluation.

2. RBench

Existing video generation benchmarks primarily focus on evaluating model performance in general scenes (Huang et al., 2024; Han et al., 2025), while other benchmarks specifically designed for physical scenarios mainly assess models' capabilities in physical reasoning (Meng et al., 2024a; Guo et al., 2025b). In this paper, we design a benchmark tailored for robotic physical scenarios, aimed at comprehensively evaluating the performance of video generation models in robotic tasks. This benchmark differs from existing general scene benchmarks by focusing on evaluating video generation models' capabilities in robotic physical environments. As shown in Figure 2, our benchmark highlights common failure modes in robotic video generation, including issues such as robot shape distortion, key action missing, non-contact attachment, and others. Section 2.1 outlines the process of benchmark construction, while Section 2.2 discusses the automatic metrics used for evaluation.

2.1. Benchmark Construction

To comprehensively evaluate video generation models for robotic scenarios, evaluation dimensions should span diverse tasks and embodiment types to reflect realistic robotic action semantics. To this end, we design a diversified benchmark with 650 evaluation cases from two aspects: task categories and embodiment types. The task-oriented categories include five representative tasks: *Common Manipulation*, *Long-horizon Planning*, *Multi-entity Collaboration*, *Spatial Relationship*, and *Visual Reasoning*, with a total of 250 image-text pairs, 50 samples for each task. The embodiment-specific categories cover four mainstream embodiment types: *Dual-arm robots*, *Humanoid robots*, *Single-arm robots*, and *Quadruped robots*, with a total of 400 image-text pairs, 100 samples for each embodiment type.

The benchmark includes a variety of text prompts and high-quality robot reference images. Each sample image is a keyframe extracted from high-quality videos sourced from public datasets or online sources, and each image is manually verified to ensure its correctness. To avoid overlap with the training data, we ensure that the selected videos in the evaluation set do not appear in the subsequent training database, and we redesign new task prompts for each reference image, effectively avoiding the risk of content overlap. All samples are verified and filtered by human annotators to ensure that the generated task prompts align with realistic logic. Appendix B Figure 5 illustrates the high aesthetic quality of the reference images (a), the broad range of testing scenarios including various objects, tasks, and action skills (b, c), and the diversity of environments in our evaluation set (d). See more details in the Appendix B.

2.2. Automatic Metrics

Existing video generation evaluation protocols, such as the representative VBench (Huang et al., 2024), primarily focus on perceptual quality, assessing aspects like frame clarity, texture fidelity, and motion smoothness. However, they lack task-specific criteria tailored to robotic scenarios. Recently, several studies (Sun et al., 2025; Gu et al., 2025; Wang et al., 2025b) have utilized multimodal large language models (MLLMs) as zero-shot evaluators for generated videos. Building upon this, we extend this approach to the domain of robotic video evaluation and propose a set of automatic evaluation metrics, incorporating manually designed indicators to assess both the visual realism and task-level validity of generated robotic videos. Following previous practices, we select the open-source Qwen3-VL (Bai et al., 2025) and the closed-source GPT-5 (OpenAI, 2025a) as our MLLM eval-

uation models. In the following sections, we introduce the evaluation methods for task completion and visual quality, respectively. Further details on metrics design and mathematical definitions are fully provided in the Appendix D.

2.2.1. Task Completion

Physical-Semantic Plausibility. This metric targets everyday physical and semantic plausibility violations that standard perception scores often miss. As shown in Figure 1, we evaluate temporal grids of uniformly sampled frames with a VQA-style protocol using MLLM. Beyond assessing physical-semantic plausibility, we place special emphasis on the following frequent failure modes: (i) *Floating/Penetration*: parts of the robot or objects are not grounded or interpenetrate with solid objects; (ii) *Spontaneous emergence*: entities appear/disappear without causal motion; (iii) *Non-contact attachment/Incorrect grasp*: objects move with the robot without visible contact or with improper gripper closure.

Task-Adherence Consistency. This metric evaluates whether the video aligns with the intent and sequence defined by the prompt. Typical deviations include missing actions (e.g., approach without grasping or placing), incorrect order (e.g., placing before grasping), semantic drift (e.g., "wiping" becomes "touching"), and non-responsiveness. We construct temporal grids and apply an MLLM-based VQA checklist, which covers: (i) *Task responsiveness*, ensuring the goal state is reached without premature interruption; (ii) *Key actions*, verifying that required actions (e.g., grasp, place, open/close) occur and align with the task prompt.

2.2.2. Visual Quality

Motion Amplitude. This metric measures the motion amplitude of the robotic subject while discounting apparent movement caused by camera motion, thereby penalizing videos that appear smooth but lack meaningful subject activity. Following VMBench (Ling et al., 2025), active subjects are localized with GroundingDINO (Liu et al., 2023a), temporally stable masks are produced by GroundedSAM (Ren et al., 2024), and salient points are tracked via CoTracker (Karaev et al., 2024). Let $\bar{D}_t$ be the mean displacement of tracked points on the subject at frame t. The Motion Amplitude Score (MAS) is

$$\mathrm{MAS} \;=\; \frac{1}{T}\sum_{t=1}^{T} \min\!\big(\bar{D}_t,\, 1\big), \tag{1}$$

where a lower MAS indicates insufficient subject motion.

Robot-Subject Stability. This metric assesses the stability of robot morphology and target object attributes over time. Typical failures include gripper/hand shape drifting into non-mechanical forms, extra/missing manipulators, link-length/topology changes, joint inversion, object attribute drift, and impossible deformation of rigid items. We adopt a contrastive VQA setup based on MLLM, which compares a reference frame and a generated frame and assigns a consistency score targeting the above failures.

Motion Smoothness. This metric quantifies temporal continuity and natural dynamics, targeting artifacts from low-level aliasing to high-level jitter/blur. We measure frame-to-frame quality stability with the Q-Align aesthetic score (Wu et al., 2023). For frames $\{f_t\}_{t=1}^{T}$ and per-frame score $Q(f_t)$, define:

$$\Delta Q_t \;=\; Q(f_{t-1}) - Q(f_t). \tag{2}$$

A temporal anomaly is flagged when ΔQ_t exceeds an adaptive threshold $\tau_s(t)$ determined by the robotic subject's motion. The Motion Smoothness Score (MSS) is

$$\mathrm{MSS} \;=\; 1 - \frac{1}{T}\sum_{t=2}^{T} \mathbb{I}\big(\Delta Q_t > \tau_s(t)\big), \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function. A higher MSS indicates smoother motion.

3. RoVid-X

In this section, we introduce the construction of a high-quality robotic video dataset, resulting in RoVid-X. The dataset is developed through a refined four-stage pipeline, as shown in Figure 3 (a). We then introduce the construction process of the dataset and provide statistical information.

3.1. Dataset Construction

Our data processing workflow consists of four distinct stages, each designed to ensure the quality, diversity, and relevance of the collected data. Detailed implementation procedures and prompt templates are deferred to Appendix C.

Robot Video Collection. We aggregate raw robotic videos from public internet platforms and over 20 open-source embodied datasets, ensuring broad diversity in robot types and scenarios, all of which are licensed for use. To ensure relevance, GPT-5 (OpenAI, 2025a) is utilized to automatically filter content based on visual and subtitle analysis, excluding clips inconsistent with robotic research objectives. This process identifies approximately 3 million high-quality raw videos covering diverse actions and morphologies.

Video Quality Filtering. We apply a rigorous filtering procedure to remove low-quality or irrelevant clips. This stage utilizes scene segmentation to exclude non-robotic content, followed by a multi-dimensional scoring system evaluating clarity, dynamic effects, aesthetics, and OCR. Each clip is assigned a quality score to ensure all retained data meets high-quality standards.

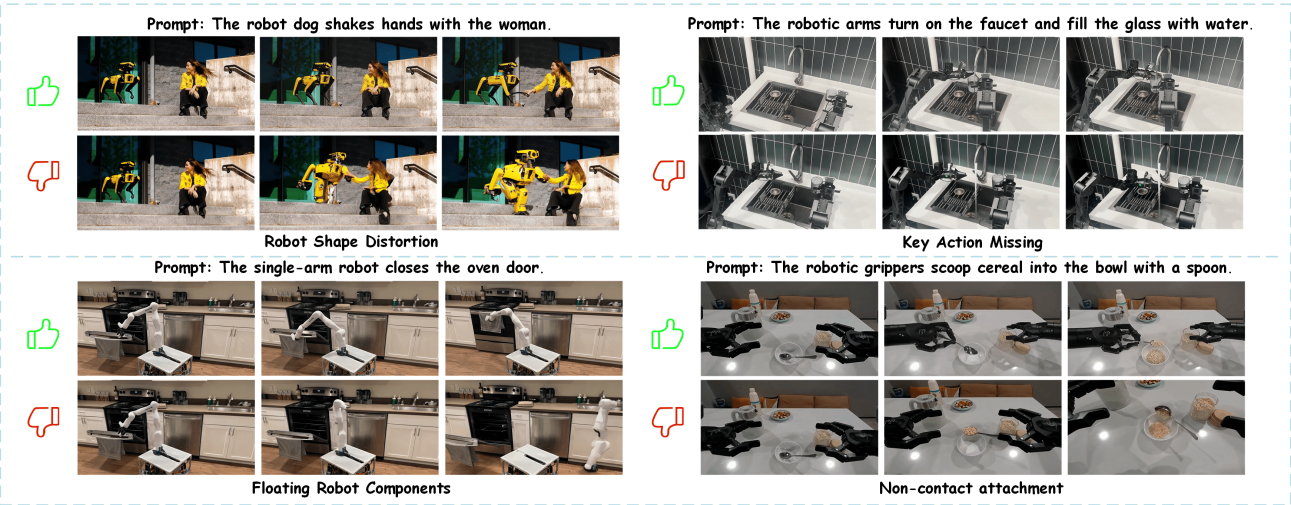

Figure 2. **Qualitative illustration of failure modes captured by RBench.** Unlike conventional metrics that focus primarily on pixel-level fidelity, RBench provides a granular evaluation across multiple dimensions, including physical plausibility and task-level consistency. These results highlight persistent challenges in robotic video generation, such as structural distortion, floating components, and key action omission, which are accurately identified by our proposed sub-metrics. More failure modes and cases are shown in the Appendix D.

Task Segmentation and Captioning. We utilize a video understanding model (Guo et al., 2025a) to analyze robot actions and segment videos based on dynamic action timestamps. This process excludes static or irrelevant scenes, ensuring precise labeling of start and end times. Subsequently, the MLLM (Guo et al., 2025a) automatically generates standardized textual descriptions for each segment, specifying the robotic subject, manipulated object, and operational details. This ensures all action descriptions are concise and strictly aligned with task requirements.

Physical Property Annotation. To ensure physical consistency and realism, we apply attribute enhancements to the videos. Specifically, we utilize FlashVSR (Zhuang et al., 2025) to improve resolution and action details, AllTracker (Harley et al., 2025) to annotate unified optical flow for subject tracking, and Video Depth Anything (Chen et al., 2025c) to generate relative depth maps for spatial modeling. These enhancements provide precise physical priors for training and evaluating robot video generation models.

3.2. Dataset Analysis

Unlike previous robotic datasets curated for VLM or VLA models, RoVid-X is the first open-source, large-scale dataset specifically designed for training and evaluating video generation models. Comprising 4 million clips, it bridges the gap between traditional video generation tasks and the unique demands of embodied robot learning, where physical interaction and real-world dynamics are pivotal.

RoVid-X spans diverse robotic actions and morphologies, providing comprehensive coverage of the physical properties required for robot training. As illustrated in Figure 3 (b), the dataset features a wide distribution across action skills, task types, and interaction objects. This variety is critical for developing robust video world models capable of simulating realistic robot behaviors in dynamic environments.

4. Experiment

4.1. Evaluation Setups

Evaluation Models. We evaluate 25 mainstream video generation models, grouped into three types. Specifically, the closed-source models include Veo 3 (Google DeepMind, 2025), Sora (OpenAI, 2024), Kling (Kling, 2025), Seedance (Gao et al., 2025b), and others, while the open-source models include several representative models such as HunyuanVideo (Wu et al., 2025a), LTX (HaCohen et al., 2026; 2024) and CogVideoX (Yang et al., 2024). Additionally, we assess models specifically designed for robotic tasks, such as DreamGen (Jang et al., 2025), and Cosmos 2.5 (Ali et al., 2025). The evaluations of these models cover various types of embodiments and multiple tasks, providing a comprehensive perspective on model performance.

Implementation Details. To ensure a fair comparison, all open-source models generate videos using their official default configurations to ensure consistency with the model's preset settings. For closed-source video models, we use their official APIs, strictly following the methods recommended by the developers for invoking and using the models. In the benchmark testing, we generate the videos for each image-text pair. To minimize errors, we generate three videos for each model sample and take the average as the final score for that sample. These generated videos are evaluated using the automated evaluation metrics that we propose. Further details on the model setup are provided in the Appendix E.

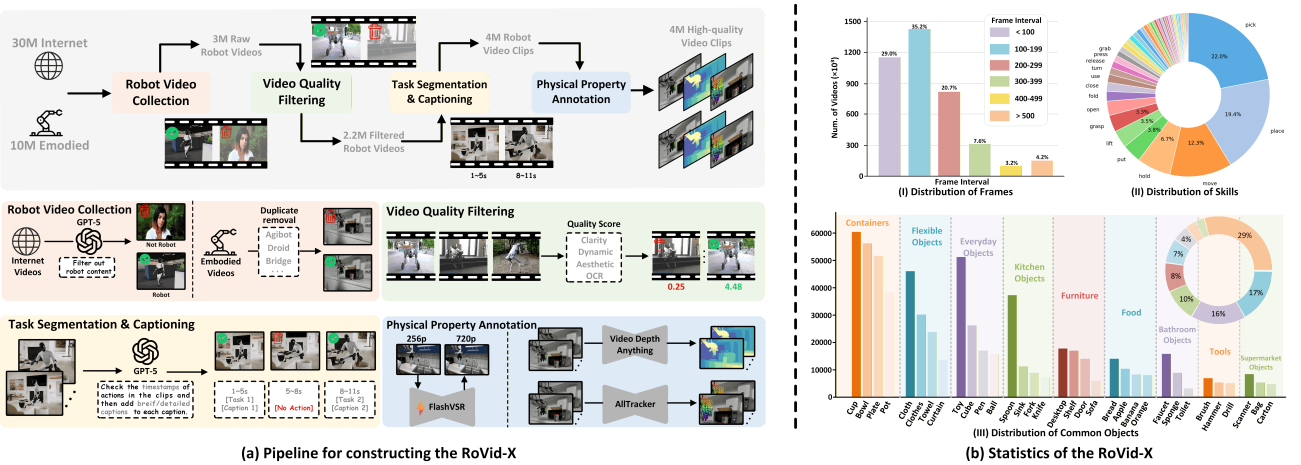

Figure 3. **Overview of RoVid-X Construction and Descriptive Statistics.** (a) shows the four-stage pipeline for constructing the RoVid-X. Additional comprehensive details are provided in Appendix C. (b) presents descriptive statistics, covering frame intervals, skill distribution, and common objects, highlighting the dataset's diversity and suitability for robotic task training and video generation.

4.2. Main Analysis

4.2.1. Quantitative Results

Table 2 presents a comprehensive quantitative evaluation across varying model architectures, tasks, and embodiments. Beyond standard performance metrics, the results reveal a pivotal paradigm shift in the video generation landscape, sparking a **rethink** of the field's focus and future direction:

From Visual Fidelity to Physical Intelligence. The most significant trend observed is the transition of video generation models from pursuing high-fidelity visualization to addressing the complex dynamics of the physical world. While traditional metrics prioritize pixel-level quality, our benchmark highlights that top-tier commercial models (e.g., Wan 2.6, Seedance 1.5 Pro) are beginning to emerge as effective ***World Simulators***. This signals a shift towards ***Physical AI***, where video generation models must simulate interaction-rich, physically challenging real-world scenarios, not just generate aesthetically pleasing videos.

Iterative Scaling Unlocks Physical Capabilities. Analyzing model evolution reveals a strong correlation between model iteration and physical capabilities. For instance, the *Wan* series exhibits a dramatic performance leap: from Wan 2.1 (Rank 14, 0.399) to Wan 2.6 (Rank 1, 0.607). Similarly, Seedance evolves from 1.0 to 1.5 Pro, climbing from Rank 6 to Rank 2. These substantial gains suggest that scaling laws and iterative optimization are not just improving visual quality but are actively refining the model's understanding of physics, distinct motion patterns, and control logic.

The "Media-Simulation" Gap in Consumer Models. Surprisingly, widely recognized consumer-oriented models like the Sora series perform suboptimally on this benchmark (Sora v2 Pro at Rank 17, Avg 0.362). This counter-intuitive result highlights a critical "domain gap": models optimized for media consumption prioritize visual smoothness and cinematic transitions, often at the expense of physical fidelity and precise motion control. This discrepancy suggests that proficiency in creative video generation does not naturally transfer to Embodied AI tasks, underlining the necessity for physically-grounded training data.

Closed-source Models Lead in Performance. Commercial closed-source models dominate the top 7 positions in our benchmark, showing a clear advantage over open-source models. The performance gap between the state-of-the-art commercial model (Wan 2.6) and the leading open-source model (Wan 2.2) highlights the need for the open-source community to focus on scaling physical training data and optimizing architectures for embodied video tasks to bridge this capability gap.

Cognitive and Fine-grained Control Bottlenecks. A consistent trend across all models is that high-level logic and precise interaction tasks pose the greatest performance bottlenecks. Top models like Wan 2.6 excel in execution-oriented tasks but show a sharp decline in *Visual Reasoning* (0.531). Additionally, a "Manipulation Gap" is observed, where models perform better on coarse-grained locomotion (Quadruped, Humanoid) than on fine-grained manipulation, indicating that mastering fine-grained contact dynamics for object interaction is more challenging than generating legged locomotion patterns.

4.2.2. Qualitative Results

We conduct a qualitative analysis of representative tasks, and the partial results are shown in Figure 4. For the vi-

Table 2. **RBench quantitative results.** Evaluations across task-oriented and embodiment-specific dimensions for 25 mainstream video generation models. The "Avg." column shows the mean score across nine indicators, with task performance in the left block and embodiment performance in the right block. The "#" next to the Sora v2 Pro model in the table indicates review limitations from the official Sora API, where 50 of 650 videos couldn't be generated. The scores derived from sub-metrics are reported in the Appendix J.

Models	Rank	Avg.	Tasks					Embodiments			
			Manipulation	Spatial	Multi-entity	Long-horizon	Reasoning	Single arm	Dual arm	Quadruped	Humanoid
Open-source											
Wan2.2_A14B (Wan et al., 2025)	8	0.507	0.381	0.454	0.373	0.501	0.330	0.608	0.582	0.690	0.648
HunyuanVideo 1.5 (Wu et al., 2025a)	10	0.460	0.442	0.316	0.312	0.438	0.364	0.513	0.526	0.634	0.595
LongCat-Video (Team et al., 2025b)	11	0.437	0.372	0.310	0.220	0.384	0.186	0.586	0.576	0.681	0.621
Wan2.1_14B (Wan et al., 2025)	14	0.399	0.344	0.268	0.282	0.335	0.205	0.464	0.497	0.595	0.599
LTX-2 (HaCohen et al., 2026)	15	0.381	0.284	0.304	0.233	0.386	0.164	0.453	0.424	0.622	0.555
Wan2.2_5B (Wan et al., 2025)	16	0.380	0.331	0.313	0.142	0.318	0.234	0.436	0.448	0.590	0.607
SkyReels (Chen et al., 2025a)	18	0.361	0.203	0.276	0.203	0.254	0.234	0.507	0.477	0.586	0.509
LTX-Video (HaCohen et al., 2024)	19	0.344	0.302	0.176	0.210	0.280	0.241	0.440	0.456	0.526	0.464
FramePack (Zhang & Agrawala, 2025)	20	0.339	0.206	0.258	0.173	0.169	0.170	0.440	0.464	0.626	0.548
HunyuanVideo (Kong et al., 2024)	21	0.303	0.177	0.180	0.108	0.147	0.035	0.454	0.480	0.625	0.524
CogVideoX_5B (Yang et al., 2024)	23	0.256	0.116	0.112	0.098	0.212	0.079	0.338	0.385	0.465	0.496
Commercial											
Wan 2.6 (Wan et al., 2025)	1	0.607	0.546	0.656	0.479	0.514	0.531	0.666	0.681	0.723	0.667
Seedance 1.5 pro (Chen et al., 2025b)	2	0.584	0.577	0.495	0.484	0.570	0.470	0.648	0.641	0.680	0.692
Wan 2.5 (Wan et al., 2025)	3	0.570	0.527	0.576	0.402	0.496	0.437	0.680	0.634	0.726	0.654
Hailuo v2 (Hailuo, 2025)	4	0.565	0.560	0.637	0.386	0.545	0.474	0.594	0.611	0.640	0.635
Veo 3 (Google DeepMind, 2025)	5	0.563	0.521	0.508	0.430	0.530	0.504	0.634	0.610	0.689	0.637
Seedance 1.0 (Gao et al., 2025b)	6	0.551	0.542	0.425	0.448	0.454	0.442	0.622	0.641	0.698	0.686
Kling 2.6 pro (Kling, 2025)	7	0.534	0.529	0.598	0.364	0.530	0.358	0.570	0.605	0.637	0.613
Sora v2 Pro# (OpenAI, 2025b)	17	0.362	0.208	0.268	0.186	0.255	0.115	0.476	0.513	0.664	0.561
Sora v1 (OpenAI, 2024)	22	0.266	0.151	0.223	0.111	0.166	0.139	0.314	0.324	0.544	0.419
Robotics-specific											
Cosmos 2.5 (Ali et al., 2025)	9	0.464	0.358	0.338	0.201	0.496	0.399	0.544	0.560	0.658	0.626
DreamGen(gr1) (Jang et al., 2025)	12	0.420	0.312	0.372	0.297	0.334	0.215	0.564	0.532	0.579	0.575
DreamGen(droid) (Jang et al., 2025)	13	0.405	0.358	0.348	0.214	0.316	0.339	0.499	0.476	0.542	0.556
Vidar (Feng et al., 2025)	24	0.206	0.073	0.106	0.050	0.054	0.050	0.382	0.410	0.374	0.357
UnifoLM-WMA-0 (Unitree, 2025)	25	0.123	0.036	0.040	0.018	0.062	0.000	0.268	0.194	0.293	0.200

sual reasoning task, Seedance 1.0 (Gao et al., 2025b) and Hailuo (Hailuo, 2025) correctly identify the blue clothing and the hollow basket, while Wan 2.5 (Wan et al., 2025) mistakenly identifies the woven basket as the hollow basket. In the long-horizon planning task, Wan 2.5 successfully completes all actions in the correct sequence, while Hailuo lacks the "turn-on" action, leading to a violation of physical logic. In the spatial relationship task, Hailuo correctly places the bok choy to the left of the pan, whereas other models mistakenly place it inside the pan. Notably, LongCat-Video introduces an unrealistic human arm intervention, disrupting physical plausibility. More detailed analysis and qualitative results can be found in the Appendix I.

These models each have their strengths, but there is still significant room for improvement in their overall performance. This further highlights the necessity of designing such a benchmark to advance video models in robotic tasks.

4.3. Human Preference Study

We conduct a human preference study with 30 participants to assess metric–human alignment. For each prompt and video, annotators compare two model outputs and choose A, B, or Tie. Per-model scores are computed by assigning 5, 3, and 1 points to wins, ties, and losses, respectively. We then compare these model-level human scores with the corresponding RBench benchmark scores. On the ten-model subset used in the study, the Spearman rank correlation score is $\rho = 0.96$ (two-sided $p < 10^{-3}$), as shown in Table 3. Overall, the top-ranked models under our benchmark largely align with human judgments. This strong consistency validates the effectiveness of our automated metrics as a reliable evaluation standard for robotic video generation. See Appendix F for more details about human evaluation.

Table 3. **Comparison between human preference scores and RBench scores.** This table demonstrates a high correlation between the two sets of scores, as reflected in the similar ranking orders. The Δr column denotes the rank difference $(r_b - r_h)$.

Model	Human	RBench	r_h	r_b	Δr
Wan 2.5	0.573	0.570	1	1	0
Veo 3	0.540	0.563	2	3	1
Hailuo v2	0.513	0.565	3	2	-1
Seedance 1.0	0.505	0.551	4	4	0
Cosmos 2.5	0.500	0.464	5	5	0
DreamGen	0.482	0.420	6	7	1
LongCat-Video	0.480	0.437	7	6	-1
Wan2.1-14B	0.378	0.399	8	8	0
CogVideoX-5B	0.333	0.256	9	10	1
LTX-Video	0.246	0.344	10	9	-1

4.4. Validation of RoVid-X

To assess the effectiveness and robustness of RoVid-X, we finetune models initialized with Wan2.1 14B and Wan2.2

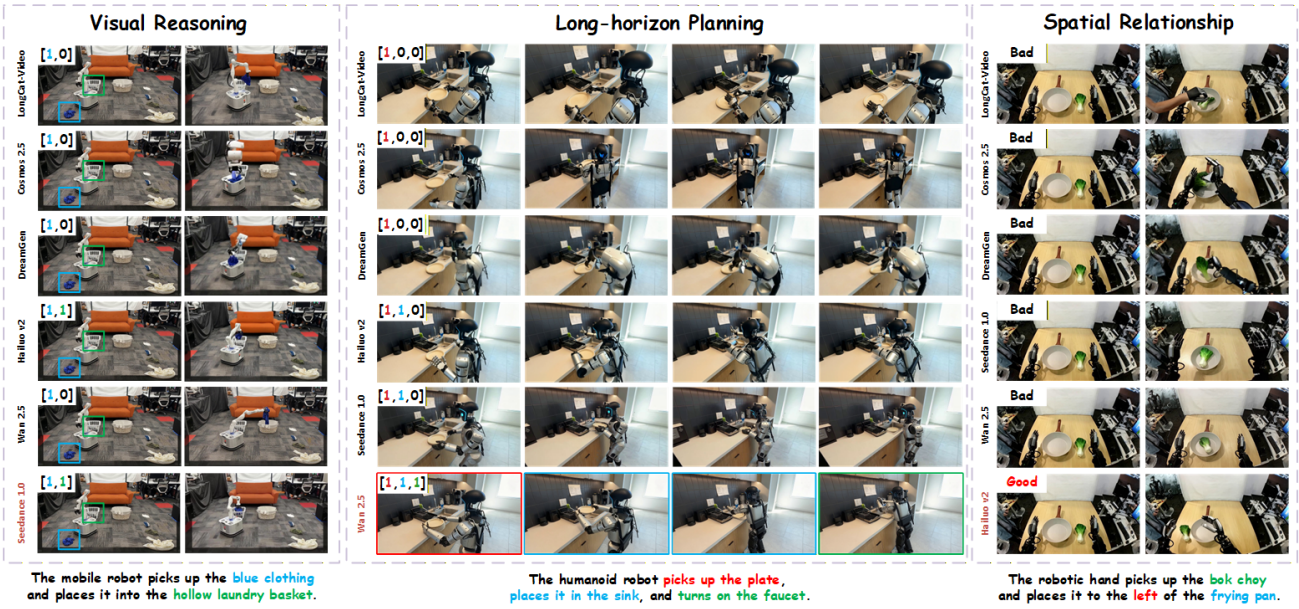

Figure 4. **Qualitative comparison across representative tasks.** We visualize the generated results for three representative tasks: **Visual Reasoning**, **Long-horizon Planning**, and **Spatial Relationship**, across six models. Each row displays temporally sampled frames from the same generated video, with captions below indicating the corresponding task instruction. More cases are shown in the Appendix I.

5B weights, using MSE loss exclusively. Due to computational constraints, we randomly sample 200k instances from the original RoVid-X dataset. The results, shown in Table 4, highlight that our dataset significantly enhances performance across five task domains and four distinct embodiments. These improvements validate both the proposed dataset and the effectiveness of our data collection pipeline.

Table 4. **RoVid-X effectiveness validation experiment.** Consistent improvements are observed across tasks and embodiments after finetuning, validating the effectiveness of the collected dataset.

Model	Manip.	Long.	Multi.	Spatial.	Reason.
Wan2.1_14B	0.344	0.335	0.282	0.268	0.205
Wan2.1_14B+Ours	**0.376**	**0.389**	**0.295**	**0.314**	**0.298**
Wan2.2_5B	0.331	0.318	0.142	0.313	0.234
Wan2.2_5B+Ours	**0.373**	**0.387**	**0.221**	**0.403**	**0.284**

Model	Single	Dual	Quad.	Humanoid	Total
Wan2.1_14B	0.464	0.497	0.595	0.599	0.399
Wan2.1_14B + Ours	**0.526**	**0.546**	**0.639**	**0.628**	**0.446**
Wan2.2_5B	0.436	0.448	0.590	0.607	0.380
Wan2.2_5B + Ours	**0.514**	**0.503**	**0.628**	**0.641**	**0.439**

5. Related Work

5.1. Video World Modeling for Robotics

Recent breakthroughs in video generation have yielded models capable of synthesizing high-quality content from text or image prompts (Runway, 2025; OpenAI, 2024; Kong et al., 2024; Wan et al., 2025). These advancements are increasingly applied to embodied intelligence (Feng et al., 2025; Unitree, 2025; Ali et al., 2025; Bruce et al., 2024) as video provides rich information for robot training (Cheang et al., 2024). Specifically, video models can synthesize robot trajectories (Jang et al., 2025; Bjorck et al., 2025; Bharadhwaj et al., 2024) to mitigate the costs of human teleoperation, with executable actions extracted via inverse dynamics models (Tian et al., 2024; Baker et al., 2022; Du et al., 2023; Zhou et al., 2024) or latent action models (Tharwat et al., 2025; Ye et al., 2025). Furthermore, video generation assists in policy learning by simulating task dynamics and predicting future states. This includes initializing robot policies (Liao et al., 2025; Jiang et al., 2025b; Wu et al., 2025b) or co-training policies with inverse dynamics models (Guo et al., 2024b; Zhu et al., 2025; Li et al., 2025). These efforts highlight the potential of video models in embodied AI. More related works are presented in the Appendix A.

6. Conclusion

We rethink video generation model for the embodied world by introducing **RBench**, a benchmark filling the gap in evaluating robot-oriented video models. Beyond perceptual metrics, RBench integrates task-level accuracy and visual fidelity through fine-grained sub-metrics. Our evaluation of 25 models reveals significant deficiencies in physical realism, while the high correlation with human judgment validates the benchmark's effectiveness. Furthermore, we present **RoVid-X**, a large-scale, diverse robotics dataset overcoming the limitations of existing resources. Together, RBench and RoVid-X establish a robust foundation for ad-

vancing video foundation models in the embodied domain.

Future Work. We aim to bridge the gap between video generation model and actionable robot policy by using Inverse Dynamics Models (IDM) to extract executable actions from generated videos, enabling closed-loop control experiments in both simulations and real-world hardware. Additionally, we plan to develop more automated, physically grounded metrics to assess generated behaviors. Our focus will also be on training video models with enhanced physical capabilities to generate high-fidelity robot actions, accelerating the development of video-driven embodied intelligence.

Acknowledgements

We are grateful to TeleAI for the resources and support that enabled this work. We also thank our coauthors for their insightful discussions, technical assistance, and constructive feedback throughout the development of this project.

Impact Statement

This paper presents work aimed at advancing the field of video generation with a focus on physical AI. There are many potential societal consequences of our work, none of which we feel need to be specifically highlighted here.

References

Aldaco, J., Armstrong, T., Baruch, R., Bingham, J., Chan, S., Draper, K., Dwibedi, D., Finn, C., Florence, P., Goodrich, S., et al. Aloha 2: An enhanced low-cost hardware for bimanual teleoperation. *arXiv preprint arXiv:2405.02292*, 2024.

Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.-W., et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., and Tang. Qwen3-vl technical report. *arXiv preprint arXiv*, 2025.

Baker, B., Akkaya, I., Zhokov, P., Huizinga, J., Tang, J., Ecoffet, A., Houghton, B., Sampedro, R., and Clune, J. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Ball, P. J., Bauer, J., Belletti, F., Brownfield, B., Ephrat, A., Fruchter, S., Gupta, A., Holsheimer, K., Holynski, A., Hron, J., Kaplanis, C., Limont, M., McGill, M., Oliveira, Y., Parker-Holder, J., Perbet, F., et al. Genie 3: A new frontier for world models. 2025.

Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.-W., and Grover, A. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Bharadhwaj, H., Dwibedi, D., Gupta, A., Tulsiani, S., Doersch, C., Xiao, T., Shah, D., Xia, F., Sadigh, D., and Kirmani, S. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al. π: A vision-language-action flow model for general robot control. *CoRR*, 2024.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Bruce, J., Dennis, M. D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Hu, X., Huang, X., et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Bugliarello, E., Moraldo, H. H., Villegas, R., Babaeizadeh, M., Saffar, M. T., Zhang, H., Erhan, D., Ferrari, V., Kindermans, P.-J., and Voigtlaender, P. Storybench: A multifaceted benchmark for continuous story visualization. *Advances in Neural Information Processing Systems*, 36: 78095–78125, 2023.

Cheang, C.-L., Chen, G., Jing, Y., Kong, T., Li, H., Li, Y., Liu, Y., Wu, H., Xu, J., Yang, Y., et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., Xiong, W., Wang, W., Pang, N., Kang, K., Xu, Z., Jin, Y., Liang, Y., Song, Y., Zhao, P., Xu, B., Qiu, D., Li, D., Fei, Z., Li, Y., and Zhou, Y. Skyreels-v2: Infinite-length film generative model, 2025a. URL https://arxiv.org/abs/2504.13074.

Chen, S., Chen, Y., Chen, Y., Chen, Z., Cheng, F., Chi, X., Cong, J., Cui, Q., Dong, Q., Fan, J., et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025b.

Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., and Kang, B. Video depth anything: Consistent depth estimation for super-long videos. *arXiv:2501.12375*, 2025c.

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 720–736, 2018.

Dasari, S., Ebert, F., Tian, S., Nair, S., Bucher, B., Schmeckpeper, K., Singh, S., Levine, S., and Finn, C. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019.

Deng, Y., Guo, X., Wang, Y., Fang, J. Z., Wang, A., Yuan, S., Yang, Y., Liu, B., Huang, H., and Ma, C. Cinema: Coherent multi-subject video generation via mllm-based guidance. *arXiv preprint arXiv:2503.10391*, 2025a.

Deng, Y., Guo, X., Yin, Y., Fang, J. Z., Yang, Y., Wang, Y., Yuan, S., Wang, A., Liu, B., Huang, H., et al. Magref: Masked guidance for any-reference video generation. *arXiv preprint arXiv:2505.23742*, 2025b.

Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J., Schuurmans, D., and Abbeel, P. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

Ebert, F., Yang, Y., Schmeckpeper, K., Bucher, B., Georgakis, G., Daniilidis, K., Finn, C., and Levine, S. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

Fang, H.-S., Fang, H., Tang, Z., Liu, J., Wang, C., Wang, J., Zhu, H., and Lu, C. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023.

Feng, Y., Tan, H., Mao, X., Xiang, C., Liu, G., Huang, S., Su, H., and Zhu, J. Vidar: Embodied video diffusion model for generalist manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

Fourier ActionNet Team, Y. M. Actionnet: A dataset for dexterous bimanual manipulation. *arXiv preprint arXiv*, 2025.

Fu, Z., Zhao, T. Z., and Finn, C. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., and Xu, Q. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28135–28144, 2025a.

Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025b.

Google DeepMind. Veo-3 technical report. Technical report, Google DeepMind, May 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Gu, J., Liu, X., Zeng, Y., Nagarajan, A., Zhu, F., Hong, D., Fan, Y., Yan, Q., Zhou, K., Liu, M.-Y., et al. " phyworldbench": A comprehensive evaluation of physical realism in text-to-video models. *arXiv preprint arXiv:2507.13428*, 2025.

Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025a.

Guo, X., Zheng, M., Hou, L., Gao, Y., Deng, Y., Wan, P., Zhang, D., Liu, Y., Hu, W., Zha, Z., Huang, H., and Ma, C. I2V-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024a.

Guo, X., Huo, J., Shi, Z., Song, Z., Zhang, J., and Zhao, J. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025b.

Guo, Y., Hu, Y., Zhang, J., Wang, Y.-J., Chen, X., Lu, C., and Chen, J. Prediction with action: Visual policy learning via joint denoising process. *Advances in Neural Information Processing Systems*, 37:112386–112410, 2024b.

Guo, Y., Shi, L. X., Chen, J., and Finn, C. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025c.

HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., and Bibi, O. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Hailuo. Hailuo. *Hailuo Lab*, 2025. URL https://hailuoai.video/.

Han, H., Li, S., Chen, J., Yuan, Y., Wu, Y., Deng, Y., Leong, C. T., Du, H., Fu, J., Li, Y., et al. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18858–18868, 2025.

Harley, A. W., You, Y., Sun, X., Zheng, Y., Raghuraman, N., Gu, Y., Liang, S., Chu, W.-H., Dave, A., Tokmakov, P., You, S., Ambrus, R., Fragkiadaki, K., and Guibas, L. J. AllTracker: Efficient dense point tracking at high resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hu, Y., Guo, Y., Wang, P., Chen, X., Wang, Y.-J., Zhang, J., Sreenath, K., Lu, C., and Chen, J. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Jang, J., Ye, S., Lin, Z., Xiang, J., Bjorck, J., Fang, Y., Hu, F., Huang, S., Kundalia, K., Lin, Y.-C., et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

Ji, P., Xiao, C., Tai, H., and Huo, M. T2vbench: Benchmarking temporal dynamics for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5325–5335, June 2024.

Jiang, T., Yuan, T., Liu, Y., Lu, C., Cui, J., Liu, X., Cheng, S., Gao, J., Xu, H., and Zhao, H. Galaxea open-world dataset and g0 dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025a.

Jiang, Y., Huang, S., Xue, S., Zhao, Y., Cen, J., Leng, S., Li, K., Guo, J., Wang, K., Chen, M., et al. Rynnvla-001: Using human demonstrations to improve robot manipulation. *arXiv preprint arXiv:2509.15212*, 2025b.

Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., and Liu, Y. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025c.

Jiang, Z., Xie, Y., Lin, K., Xu, Z., Wan, W., Mandlekar, A., Fan, L. J., and Zhu, Y. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16923–16930. IEEE, 2025d.

Ju, X., Wang, T., Zhou, Y., Zhang, H., Liu, Q., Zhao, N., Zhang, Z., Li, Y., Cai, Y., Liu, S., et al. Editverse: Unifying image and video editing and generation with in-context learning. *arXiv preprint arXiv:2509.20360*, 2025.

Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., and Feng, J. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Karaev, N., Makarov, I., Wang, J., Neverova, N., Vedaldi, A., and Rupprecht, C. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proc. arXiv:2410.11831*, 2024.

Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M. K., Chen, L. Y., Ellis, K., et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Kim, G., Han, J., and Cho, S. Videofrom3d: 3d scene video generation via complementary image and video diffusion models. *arXiv preprint arXiv:2509.17985*, 2025.

Kling. Image to video elements feature, 2025. URL https://klingai.com/global/.

Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Li, S., Gao, Y., Sadigh, D., and Song, S. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

Liang, J., Tokmakov, P., Liu, R., Sudhakar, S., Shah, P., Ambrus, R., and Vondrick, C. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025.

Liao, M., Ye, Q., Zuo, W., Wan, F., Wang, T., Zhao, Y., Wang, J., Zhang, X., et al. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816, 2024.

Liao, Y., Zhou, P., Huang, S., Yang, D., Chen, S., Jiang, Y., Hu, Y., Cai, J., Liu, S., Luo, J., et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Ling, X., Zhu, C., Wu, M., Li, H., Feng, X., Yang, C., Hao, A., Zhu, J., Wu, J., and Chu, X. Vmbench: A benchmark for perception-aligned video motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13087–13098, 2025.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023a.

Liu, Y., Li, L., Ren, S., Gao, R., Li, S., Chen, S., Sun, X., and Hou, L. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36: 62352–62387, 2023b.

Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., and Shan, Y. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024.

Lynch, C., Wahid, A., Tompson, J., Ding, T., Betker, J., Baruch, R., Armstrong, T., and Florence, P. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.

Mandlekar, A., Zhu, Y., Garg, A., Booher, J., Spero, M., Tung, A., Gao, J., Emmons, J., Gupta, A., Orbay, E., et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pp. 879–893. PMLR, 2018.

Mandlekar, A., Nasiriany, S., Wen, B., Akinola, I., Narang, Y., Fan, L., Zhu, Y., and Fox, D. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023.

Mao, J., Zhao, S., Song, S., Shi, T., Ye, J., Zhang, M., Geng, H., Malik, J., Guizilini, V., and Wang, Y. Learning from massive human videos for universal humanoid pose control. *arXiv preprint arXiv:2412.14172*, 2024.

Meng, F., Liao, J., Tan, X., Shao, W., Lu, Q., Zhang, K., Cheng, Y., Li, D., Qiao, Y., and Luo, P. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024a.

Meng, F., Shao, W., Luo, L., Wang, Y., Chen, Y., Lu, Q., Yang, Y., Yang, T., Zhang, K., Qiao, Y., et al. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802*, 2024b.

Nasiriany, S., Maddukuri, A., Zhang, L., Parikh, A., Lo, A., Joshi, A., Mandlekar, A., and Zhu, Y. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

OpenAI. Sora, 2024. URL https://openai.com/sora/. Accessed: 2025-02-26.

OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025a.

OpenAI. Sora2, 2025b. URL https://openai.com/zh-Hans-CN/index/sora-2/.

O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Pika. Pika art 2.0's scene ingredients: Redefining personalized video creation, 2025. URL https://pikartai.com/scene-ingredients/. Accessed: 2025-02-26.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., and Zhang, L. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., and Gao, J. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6121–6132, 2025.

Runway. Runway, 2025. URL https://runwayml.com/. Accessed: 2025-02-26.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., and Liu, X. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8406–8416, 2025.

Team, G., Ye, A., Wang, B., Ni, C., Huang, G., Zhao, G., Li, H., Li, J., Zhu, J., Feng, L., et al. Gigabrain-0: A world model-powered vision-language-action model. *arXiv preprint arXiv:2510.19430*, 2025a.

Team, M. L., Cai, X., Huang, Q., Kang, Z., Li, H., Liang, S., Ma, L., Ren, S., Wei, X., Xie, R., and Zhang, T. Longcat-video technical report, 2025b. URL https://arxiv.org/abs/2510.22200.

Tharwat, B., Nasser, Y., Abouzeid, A., and Reid, I. Latent action pretraining through world modeling. *arXiv preprint arXiv:2509.18428*, 2025.

Tian, Y., Yang, S., Zeng, J., Wang, P., Lin, D., Dong, H., and Pang, J. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2024.

Tian, Y., Yang, Y., Xie, Y., Cai, Z., Shi, X., Gao, N., Liu, H., Jiang, X., Qiu, Z., Yuan, F., et al. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. *arXiv preprint arXiv:2511.16651*, 2025.

Unitree. Unifolm-wma-0: A world-model-action (wma) framework under unifolm family, 2025.

Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wang, J., Ma, A., Cao, K., Zheng, J., Zhang, Z., Feng, J., Liu, S., Ma, Y., Cheng, B., Leng, D., et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025a.

Wang, Y., Xian, Z., Chen, F., Wang, T.-H., Wang, Y., Fragkiadaki, K., Erickson, Z., Held, D., and Gan, C. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*, 2023.

Wang, Y., He, X., Wang, K., Ma, L., Yang, J., Wang, S., Du, S. S., and Shen, Y. Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13629–13638, 2025b.

Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., and Shan, Y. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Wiedemer, T., Li, Y., Vicol, P., Gu, S. S., Matarese, N., Swersky, K., Kim, B., Jaini, P., and Geirhos, R. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025a.

Wu, H., Zhang, Z., Zhang, W., Chen, C., Li, C., Liao, L., Wang, A., Zhang, E., Sun, W., Yan, Q., Min, X., Zhai, G., and Lin, W. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi.

Wu, H., Jing, Y., Cheang, C., Chen, G., Xu, J., Li, X., Liu, M., Li, H., and Kong, T. Unleashing large-scale video generative pre-training for visual robot manipulation. In *The Twelfth International Conference on Learning Representations*, 2025b.

Wu, K., Hou, C., Liu, J., Che, Z., Ju, X., Yang, Z., Li, M., Zhao, Y., Xu, Z., Yang, G., et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024a.

Wu, P., Shentu, Y., Yi, Z., Lin, X., and Abbeel, P. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12156–12163. IEEE, 2024b.

Wu, S., Liu, X., Xie, S., Wang, P., Li, X., Yang, B., Li, Z., Zhu, K., Wu, H., Liu, Y., et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025c.

Yan, T., Han, W., Zhou, X., Zhang, X., Zhan, K., Xu, C.-z., and Shen, J. Rlgf: Reinforcement learning with geometric feedback for autonomous driving video generation. *arXiv preprint arXiv:2509.16500*, 2025.

Yang, L., Suh, H., Zhao, T., Graesdal, B. P., Kelestemur, T., Wang, J., Pang, T., and Tedrake, R. Physics-driven data generation for contact-rich manipulation via trajectory optimization. *arXiv preprint arXiv:2502.20382*, 2025.

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Ye, S., Jang, J., Jeon, B., Joo, S. J., Yang, J., Peng, B., Mandlekar, A., Tan, R., Chao, Y.-W., Lin, B. Y., et al. Latent action pretraining from videos. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuan, S., He, X., Deng, Y., Ye, Y., Huang, J., Lin, B., Luo, J., and Yuan, L. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*, 2025.

Zhang, C., Cherniavskii, D., Tragoudaras, A., Vozikis, A., Nijdam, T., Prinzhorn, D. W., Bodracska, M., Sebe, N., Zadaianchuk, A., and Gavves, E. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918*, 2025.

Zhang, L. and Agrawala, M. Packing input frame contexts in next-frame prediction models for video generation. *Arxiv*, 2025.

Zhao, Z., Jing, H., Liu, X., Mao, J., Jha, A., Yang, H., Xue, R., Zakharor, S., Guizilini, V., and Wang, Y. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. *arXiv preprint arXiv:2510.08807*, 2025.

Zhen, H., Sun, Q., Zhang, H., Li, J., Zhou, S., Du, Y., and Gan, C. Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

Zhou, S., Du, Y., Chen, J., Li, Y., Yeung, D.-Y., and Gan, C. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

Zhu, C., Yu, R., Feng, S., Burchfiel, B., Shah, P., and Gupta, A. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.

Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., and Xue, T. Flashvsr: Towards real-time diffusion-based streaming video super-resolution, 2025. URL https://arxiv.org/abs/2510.12747.

Rethinking Video Generation Model for the Embodied World

Appendix

A. More related works

A.1. Datasets for Robot Learning

A core challenge in robot learning is the lack of large-scale, diverse datasets that facilitate the training of general-purpose robots with physical interaction capabilities (O'Neill et al., 2024; Bjorck et al., 2025). Currently, datasets used in the embodied intelligence community for robot learning can be broadly classified into three categories: real-world robot data (Brohan et al., 2022; O'Neill et al., 2024; Zhao et al., 2025; Fourier ActionNet Team, 2025; Mao et al., 2024), human video data (Damen et al., 2018; Goyal et al., 2017; Grauman et al., 2022), and synthetic robot data (Yang et al., 2025; Wang et al., 2023; Nasiriany et al., 2024; Mandlekar et al., 2023; Jiang et al., 2025d; Tian et al., 2025). As a key element in training physical AI models, most existing real-world robot datasets are collected through robotic teleoperation (Wu et al., 2024b; Fu et al., 2024; Aldaco et al., 2024) or by teams of human operators (Bu et al., 2025; Lynch et al., 2023; Black et al., 2024), which leads to high collection costs and limited data scale. Furthermore, these datasets predominantly focus on similar types of robots, resulting in issues of limited diversity and restricted environments (Ebert et al., 2021; Khazatsky et al., 2024). Additionally, inconsistent data collection and storage methods across different datasets make it difficult to enable effective cross-dataset co-training. Our focus is on collecting robot data for video generation that spans various robot morphologies and entities, and providing a unified set of physical attributes for all data sources, thereby advancing cross-entity research in robot learning.

A.2. Benchmarks for Video Generation

Establishing robust evaluation frameworks is essential for measuring the progress of video generation models. Currently, evaluation methodologies can be categorized into three primary streams: visual fidelity and semantics, which assess basic clarity and text-video alignment (Liu et al., 2023b; Sun et al., 2025; Yuan et al., 2025); temporal dynamics, focusing on motion consistency and long-range narrative coherence (Ji et al., 2024; Liao et al., 2024; Bugliarello et al., 2023); and physical plausibility, which examines adherence to fundamental laws such as inertia and collision dynamics (Meng et al., 2024a; Bansal et al., 2024; Meng et al., 2024b; Wang et al., 2025a). While these benchmarks provide valuable insights into general video quality, they are largely decoupled from the specific requirements of embodied AI. Specifically, existing frameworks often rely on isolated physical constraints or local visual metrics, failing to capture the complex interplay between robotic actions and environmental responses. Furthermore, there is a distinct lack of systematic evaluation for task-level correctness and spatial constraints in multi-embodiment scenarios. To bridge this gap, we propose a comprehensive benchmark specifically tailored for robotic video generation, introducing reproducible metrics that unify physical realism with task-oriented action completeness.

B. Evaluation Set Details

B.1. Task-Oriented Evaluation Set

To systematically evaluate the multi-dimensional task execution capabilities of video generation models in robotic scenarios, RBench constructs a task-oriented evaluation set with five core task dimensions: *Common Manipulation*, *Long-Horizon Planning*, *Multi-Entity Collaboration*, *Spatial Relationship*, and *Visual Reasoning*. For each task, we collect 50 images as initial frames from open-source datasets or public web sources. Human annotators then create and verify corresponding text prompts to ensure both correctness and diversity of language descriptions. Together, these image-text pairs define a diverse evaluation corpus that covers a wide range of everyday manipulation, complex planning, multi-entity interaction, spatial reasoning, and visual-semantic reasoning scenarios.

B.1.1. Common Manipulation

This task evaluates the ability of video generation models to produce diverse manipulation behaviors in basic object interaction scenarios. The scenes cover single-arm, dual-arm, and humanoid robots performing typical manipulation actions such as grasping, placing, pushing, rotating, and pressing. The dataset focuses on whether the model can generate physically plausible, temporally coherent, and natural manipulation behaviors that achieve the specified goals for everyday object handling.

B.1.2. Long-Horizon Planning

This task evaluates the capability of video generation models to understand and generate long-horizon robotic behaviors that involve multi-stage action planning. Each instance in the evaluation set is composed of multiple sequential sub-actions, including:

- **Object Manipulation Sequences:** e.g., The robot opens

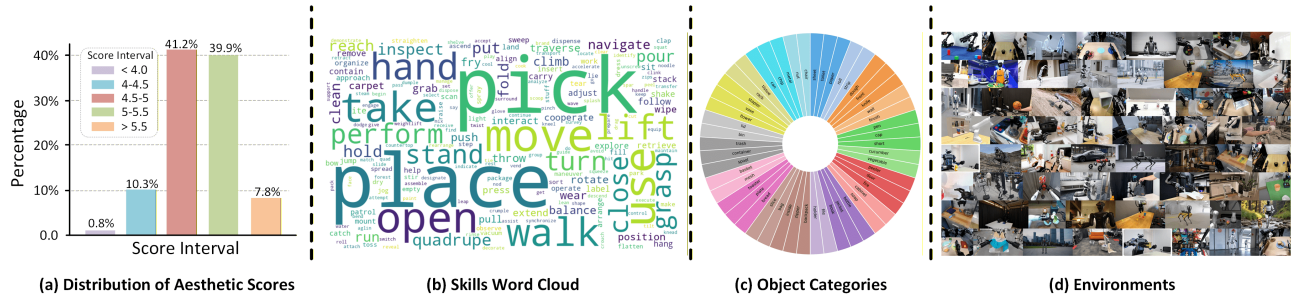

Figure 5. **Statistics in RBench.** The benchmark covers diverse tasks, object categories, and environments, demonstrating the high quality and comprehensiveness of the evaluation set, highlighting its high applicability to a wide range of robotic video generation scenarios.

the refrigerator door, takes the green box out of the refrigerator, and then closes the refrigerator door," which requires clearly delineated action stages and physically reasonable transitions.

- **Multi-Step Spatial Planning:** e.g., The humanoid robot picks up the bag, turns around, climbs up the stairs, and walks across the wooden plank," emphasizing continuous spatial transitions and modeling of multi-step action chaining.
- **Physical Motion and Body Coordination:** e.g., "The quadruped robot performs a front flip, lands steadily, then leans forward and balances upside down on its front legs," which assesses temporal coherence and physical plausibility in complex motion and body control.

Overall, the dataset spans a broad range of tasks from everyday interactions to dynamic control, focusing on a model's capability in action decomposition, stage transitions, and cross-time reasoning for comprehensive long-horizon planning.

B.1.3. Multi-Entity Collaboration

This task focuses on evaluating the capability of video generation models to depict collaborative and interactive behaviors in multi-entity robotic scenarios. Each scene contains a *Primary Entity* and a *Secondary Entity*. The Primary Entity can be a single-arm robot, a dual-arm robot, a humanoid robot, or a quadruped robot, while the Secondary Entity can be a human, an animal, or another robot. The task covers diverse interaction types such as object handover and usage, dressing assistance, collaborative task completion, following, and guidance. The dataset is designed to assess whether the model can generate natural, temporally coherent, and task-consistent multi-entity collaboration behaviors at semantic, temporal, and physical levels.

B.1.4. Spatial Relationship

This task evaluates the ability of video generation models to understand and express spatial relationships in generated videos. We construct scenes where humanoid robots, single-arm robots, and quadruped robots interact with clearly defined objects while satisfying various spatial relations, such as above/below, left/right, and front/behind. The dataset requires models to correctly present relative positions, orientations, and motion trajectories between entities, revealing their competence in spatial understanding and geometric reasoning. Consistent spatial layouts and motion patterns across time are essential to correctly reflect the described spatial relations.

B.1.5. Visual Reasoning

This task aims to evaluate the visual-semantic reasoning capabilities of video generation models in complex scenes. The evaluation set includes a wide range of visual concepts and multi-level semantic logic, such as:

- Color recognition (e.g., pick up the sky-blue book);
- Numerical and ordering reasoning (e.g., the robot places the apple, water bottle, and Rubik's cube into the bag in that order);
- Attribute and category matching (e.g., the robot gripper places the white bottle of baby powder onto the shelf, aligning it with other identical bottles in the same column);
- Geometric and object-property understanding (e.g., the robot picks up the tallest orange object and places it into the basket);
- Text and semantic understanding (e.g., the left manipulator places the cup under the white dispenser labeled 'Jasmine Tea,' and the right manipulator opens the dispenser to pour jasmine tea into the cup);
- Visual feature understanding (e.g., the robot grasps the book with a portrait of a person on its cover).

This dataset is designed to emphasize fine-grained visual grounding, logical consistency, and the ability to align robot actions with high-level visual-semantic reasoning requirements.

B.2. Embodiment-Specific Evaluation Set

Different types of robots exhibit substantial variations in morphology, degrees of freedom, control modes, and task objectives. These factors directly influence the modeling complexity and generalization challenges faced by video generation models in robotics contexts. To more systematically analyze model performance across heterogeneous robot embodiments, RBench constructs embodiment-specific evaluation subsets that encompass four representative robot categories: dual-arm robots, humanoid robots, single-arm robots, and quadruped robots. For each embodiment, 100 initial-frame images are sourced from open-access or publicly available datasets, and human annotators create and verify the corresponding prompts for accuracy and linguistic diversity.

Each subset includes a diverse range of robot models, action types, manipulated objects, scene environments, and both first-person and third-person perspectives. This design introduces embodiment-specific challenges: dual-arm robots emphasize coordinated bimanual manipulation, humanoid robots prioritize tool use and natural full-body postures, single-arm robots focus on precise object interactions, and quadruped robots predominantly test terrain adaptation and motion continuity. Along with the task-oriented splits, these embodiment-specific subsets provide a comprehensive and structured dataset for benchmarking video generation models in the embodied world.

Evaluating models across these four robot categories further uncovers the current biases and capability preferences of image-to-video generation models. For instance, due to extensive pre-training on large-scale human activity datasets, many models tend to exhibit higher task completion rates and better visual quality in humanoid-robot scenarios, while they often struggle with fine-grained single-arm manipulation. By systematically comparing performance across different embodiments, RBench makes such imbalances explicit and offers a principled framework to identify where current models excel or fail.

More broadly, embodiment-aware datasets like RBench are pivotal for advancing video foundation models in robotics. They promote the development of architectures and training strategies capable of generalizing beyond human-centric motion priors, enabling models to learn and adapt to a broader distribution of robot-specific motion patterns, rather than relying solely on human demonstrations. Furthermore, they facilitate fair and transparent comparisons between models, allowing evaluation results to clearly identify which models perform best for specific robot embodiments. Finally, such datasets help bridge the gap between generic video generation and physically grounded embodied intelligence, fostering the transition from visually appealing but brittle outputs toward robust, controllable, and deployment-ready generative models for real-world robotic systems.

C. RoVid-X Dataset Construction

RoVid-X dataset primarily comes from internet-sourced robotic videos that are public domain or non-copyrighted, as well as open-source embodied video datasets, all of which are licensed for use. Our data processing workflow consists of four distinct stages, each designed to ensure the quality, diversity, and relevance of the collected data. These stages are outlined as follows:

Robot Video Collection. In the first stage, we collect raw robotic videos from large-scale internet video platforms and over 20 open-source embodied video datasets. These datasets cover a variety of robot types and task scenarios, ensuring the breadth and diversity of the data. To improve dataset relevance and quality, we employ the GPT-5 model (OpenAI, 2025a) to automatically filter the content of each video and remove low-quality or irrelevant video clips that do not align with the research objectives. During the filtering process, GPT-5 identifies videos related to robotic tasks and actions based on visual content and subtitles, ensuring that all collected videos effectively support the training and evaluation of robotic tasks. After this filtering process, we identify approximately 3 million raw robotic video clips, covering different actions, tasks, and robot types.

Video Quality Filtering. In this stage, we perform a rigorous filtering procedure on the collected videos to remove low-quality and irrelevant video clips that do not align with the research objectives. First, we apply scene segmentation detection to remove all video data unrelated to robots. Then, we use a video quality scoring system to assess the videos from multiple dimensions, including clarity, dynamic effects, aesthetic performance, and optical character recognition (OCR), among other metrics. Each video clip is assigned a quality score based on these criteria, ensuring that the videos retained in the final dataset meet high-quality standards.

Task Segmentation and Captioning. In this stage, we use a video understanding model (Guo et al., 2025a) and a specially designed prompt template to automatically analyze the robot actions within the videos. The system segments the videos into different task segments based on timestamps, generating short subtitles for each task segment that accurately describe the robot's actions and operational details in that task.

The action recognition and description process for each task segment follows these steps: First, the system identifies all dynamic actions within the video and excludes static scenes or irrelevant actions (e.g., waiting or remaining still). The time range for each action (start and end

times) is precisely labeled to ensure accuracy. Next, using the MLLM model (Guo et al., 2025a), textual descriptions of each action are automatically generated, including the action subject (e.g., "right arm" or "left gripper"), the object being manipulated (e.g., "nameplate" or "box"), and the specific operation details (e.g., "grasp and move" or "remove from the table"). Finally, the subtitles for each task segment are output in a standardized format, ensuring that the action descriptions for each video clip are clear, concise, and aligned with the task requirements.

Physical Property Annotation. To ensure consistency and realism of robot actions within physical space, we apply physical attribute enhancement to the videos. Specifically, we use FlashVSR (Zhuang et al., 2025) to improve the video resolution, making the images clearer and enhancing the details of the actions. Then, using the AllTracker tool (Harley et al., 2025), we annotate a unified optical flow for the subjects in the videos, ensuring consistency in tracking and recording robot actions across different scenes. Additionally, using Video Depth Anything (Chen et al., 2025c), we generate relative depth maps to accurately describe the spatial relationships and depth information of objects in the scene. The goal of these physical attribute annotations is to provide researchers with more precise reference data, aiding in the training and evaluation of robot video generation models and offering richer physical data support for future research.

D. Automatic Metrics Details

To quantitatively assess the core capabilities of different models in robot video generation, we design five fine-grained metrics: *Physical-Semantic Plausibility*, *Task-Adherence Consistency*, *Motion Amplitude*, *Robot-Subject Stability*, and *Motion Smoothness*. These metrics are evaluated using an MLLM-based, VQA-style protocol applied to grid images composed of key frames sampled from each generated video. Additionally, the evaluation is supplemented by low-level, non-MLLM computational indicators that capture pixel-level motion statistics and temporal dynamics. Together, these two layers of evaluation provide a comprehensive assessment of both task completion and visual quality in robotic video generation.

D.1. Physical-Semantic Plausibility

In robotics video generation, models often produce physically implausible or commonsense-violating artifacts, such as grippers passing through objects, floating objects, or the sudden appearance of irrelevant entities. These errors are typically undetectable by standard visual perception metrics, yet they directly highlight limitations in a model's understanding of physical laws and semantic causality.

To capture these issues, we introduce the *Physical-Semantic Plausibility* metric, implemented via a VQA-style evaluation pipeline. The MLLM receives a grid image composed of key frames from the generated video and is prompted to detect the following types of violations:

- **Floating and unsupported entities.** As illustrated in Figure 6, the metallic spoon and the single-arm robot's manipulator are suspended in mid-air without any physically plausible support.
- **Non-contact attachment and incorrect grasping.** For example, in Figure 6(a), the metallic spoon moves rigidly with the gripper even though there is no clear contact or gripper closure, resulting in an unrealistic "sticking" effect.
- **Sudden appearance, disappearance, or duplication.** As demonstrated in Figure 7, (a) the robotic arm suddenly disappears in later frames, (b) human hands and a new notebook suddenly appear, and (c) the number of instant noodle packs is spuriously duplicated.
- **Interpenetration.** As shown in Figure 8, the humanoid robot hand unrealistically penetrates the box, indicating a severe violation of rigid-body constraints.

These anomalies are treated as severe physical violations that significantly reduce the credibility of the generated video. Beyond local error detection, the evaluator is also required to assess whether the overall action sequence and causal progression are reasonable, thereby characterizing the extent to which the model produces videos that are consistent with basic physical laws and human common sense.

D.2. Task-Adherence Consistency

Robotics video generation models often exhibit task-level deviations, such as ignoring the specified objective or omitting critical action stages. To measure this behavior, we design the *Task-Adherence Consistency* metric, using a VQA-style evaluation protocol. The MLLM inspects the grid of key frames and assesses the following:

- **Task responsiveness.** As shown in Figure 9(a), the failure case illustrates a robot gripper that does not respond to the instruction to grasp the mouse; the gripper remains static, and the intended task is never initiated or completed.
- **Key action completeness.** As illustrated in Figure 9(b), the failure case omits crucial actions such as *turn on* and *fill*: the faucet is never visibly operated, yet water still flows from the tap, disrupting the causal chain between actions and outcomes.

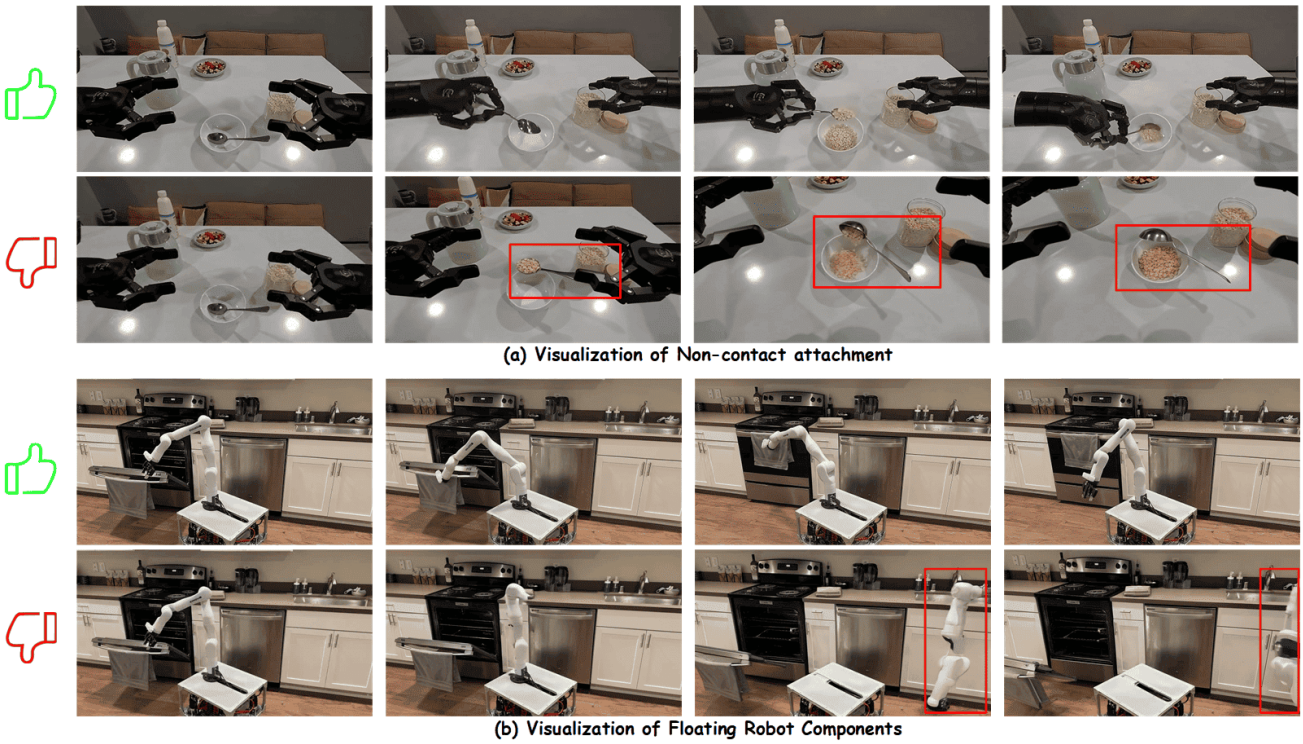

Figure 6. **Visualization of robot and subject floating.**

These phenomena reflect shortcomings in semantic understanding, action planning, and execution consistency with respect to the prompt. Importantly, they are also difficult to capture with conventional low-level perception metrics, highlighting the necessity of explicit task-adherence evaluation in robotics contexts.

Concretely, Task-Adherence Consistency is instantiated with task-specific criteria for the five task families introduced in Section B.1, with a focus on the following:

- **Common Manipulation.** Task adherence is primarily assessed through: (i) *Task Completion*, which checks whether the robot successfully accomplishes the manipulation objective described in the prompt while exhibiting reasonable intermediate phases (e.g., approach $\rightarrow$ grasp $\rightarrow$ move $\rightarrow$ place); and (ii) *Action Effectiveness*, which evaluates the physical plausibility and dynamic coherence of the manipulation, including natural gripper closure, appropriate contact locations, and smooth trajectories. Attempts with obviously discontinuous, incomplete, or physically implausible actions are regarded as failures.
- **Multi-Entity Collaboration.** For collaborative scenes involving a Primary and a Secondary Entity, task adherence is assessed through two aspects: (i) *Task Completion*, requiring that both entities execute their respective roles and complete all required interaction steps in a temporally coherent and logically consistent manner; and (ii) *Action Effectiveness*, which measures the completeness and coordination of interaction behaviors. For contact-based interactions, a full sequence of "approach $\rightarrow$ contact $\rightarrow$ release/transfer" is expected; for non-contact interactions (e.g., following, joint motion), a coherent process of "initiation $\rightarrow$ alignment $\rightarrow$ sustained coordination" is required. Missing stages, asynchronous responses, or logically inconsistent behaviors are treated as unsuccessful.
- **Spatial Relationship.** In spatial reasoning scenarios, task adherence is assessed through: (i) *Spatial Relation Accuracy*, which checks whether the spatial relations between entities (e.g., above/below, left/right, front/behind) match the textual description with consistent orientation, scale, and viewpoint; and (ii) *Manipulation Feasibility*, which examines whether the direction, trajectory, and intent of the robot's motion are compatible with the described spatial relations (e.g., moving leftward when instructed to move "to the left of"). Trajectories that contradict the described direction or result in physically unreasonable motions are considered incorrect.
- **Visual Reasoning.** In visually and semantically complex scenes, task adherence is assessed through two aspects: (i) *Visual Reasoning Accuracy*, evaluated

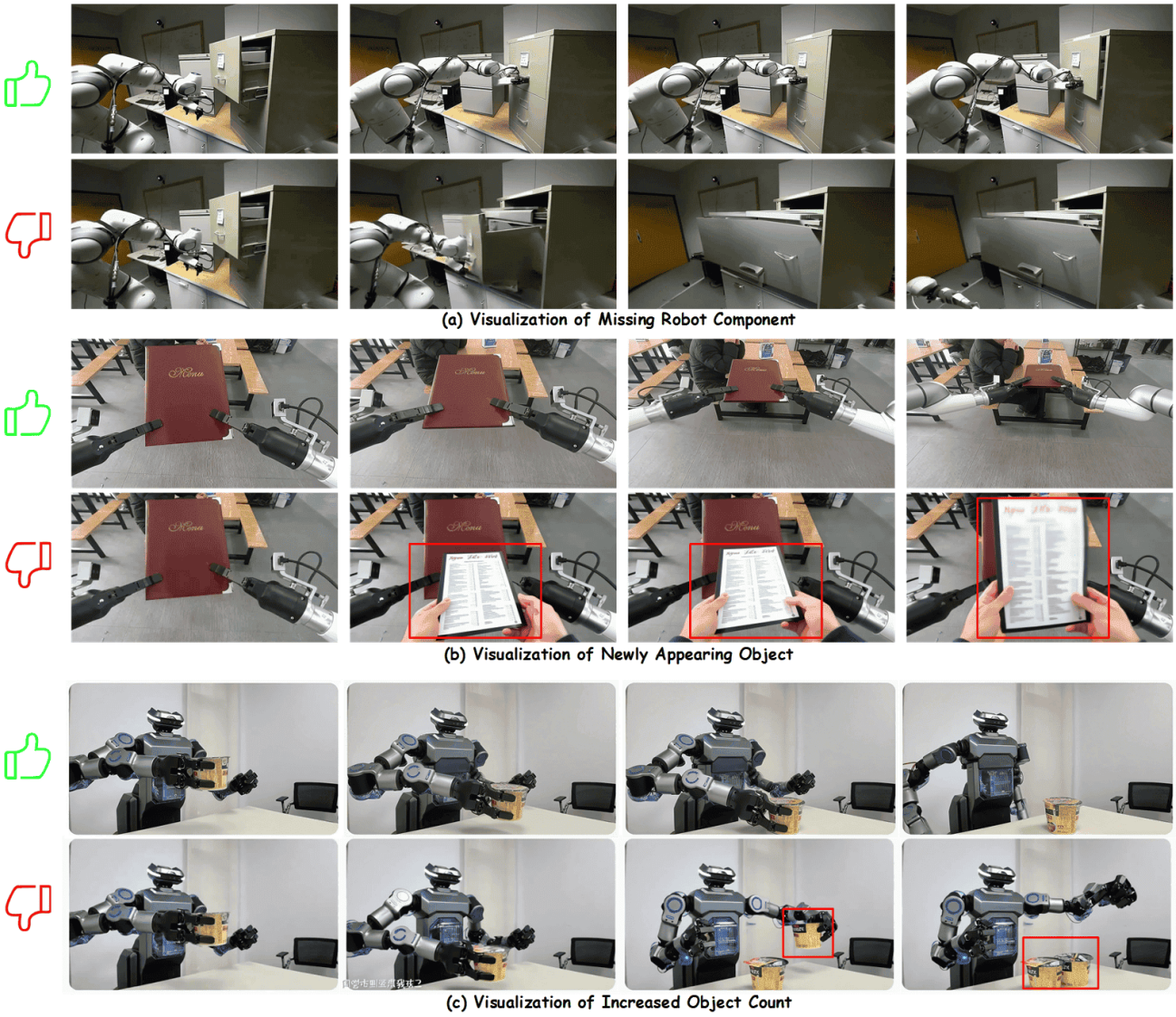

Figure 7. **Visualization of robot/subject sudden appearance, disappearance, or duplication.**

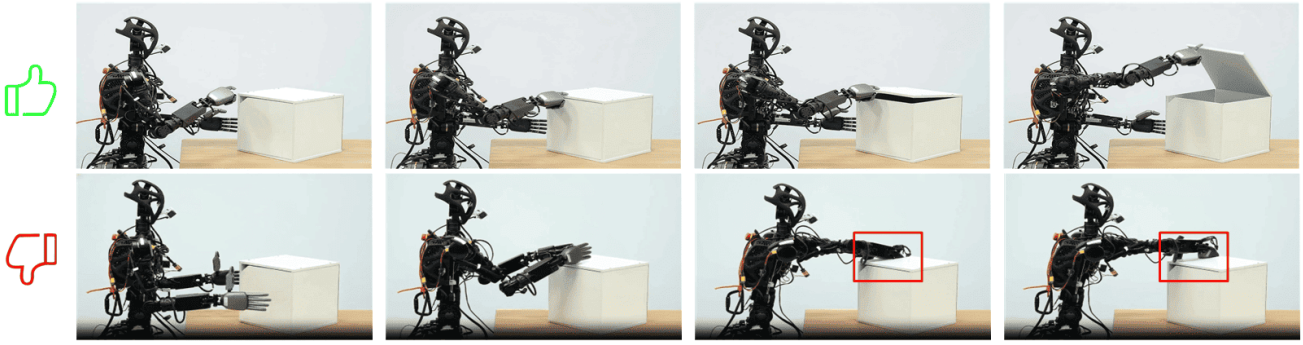

Figure 8. **Visualization of robot interpenetration.**

via an automatic Question Chain mechanism: given the prompt, MLLM first generates a set of stepwise verification questions covering the trigger-feedback-outcome logic. The same MLLM then answers these questions based on the generated video, and a score is computed as follows:

$$\text{Score} = 5 \times \frac{\text{completed questions}}{\text{total questions}}, \tag{4}$$

where missing or incorrect events are treated as unfulfilled steps. This encourages the video to satisfy both the visual and logical requirements of the task. Additionally, (ii) *Action Effectiveness* measures the physical plausibility and dynamic coherence of the robot's motions, penalizing clearly discontinuous, incomplete, or physically implausible actions even if some high-level reasoning appears correct.

- **Long-Horizon Planning.** For long-horizon tasks composed of multiple ordered sub-events, task adherence is assessed through: (i) *Event Completion Rate*. For each sample, an event list is defined as an ordered set of events. This list is reformulated into a numbered sequence, e.g., *1. open the refrigerator door; 2. take out the green box; 3. close the door*, and the final score is computed as follows:

$$\text{Score} = 5 \times \frac{\text{completed events}}{\text{total events}}, \tag{5}$$

this measures how completely the required event sequence is executed. Complementarily, (ii) *Action Effectiveness* again assesses the physical plausibility and temporal coherence of the underlying motions (e.g., natural body coordination, stable landing, and smooth transitions between stages), ensuring that partially correct high-level event ordering without valid execution is not over-rewarded.

D.3. Robot-Subject Stability

In robotics video generation, maintaining stable structure and appearance for both the robot and the manipulated objects is essential for assessing generation quality. In practice, models often exhibit abnormal changes in robot morphology or severe distortions of subject attributes. To systematically evaluate these issues, we propose the *Robot-Subject Stability* metric, which separately measures the visual and semantic consistency of the robot and the target subject throughout the generation process.

We adopt a comparative VQA mechanism: the system simultaneously observes two frames, with the left frame as a reference image and the right frame as a generated frame, and focuses on a specified entity (e.g., the *robotic gripper* or the *target subject*). The MLLM is prompted to judge how well the entity's appearance, structure, and semantics are preserved between the two frames. Specifically, the evaluator identifies:

- **Robot structural stability.** As shown in Figure 10, (a) a humanoid robot degenerates into a single-arm robot, (b) a quadruped robot morphs into a small humanoid robot, (c) a single-arm robot transforms into a humanoid robot, and (d) a parallel gripper deforms into a dexterous robotic hand. These cases reveal structural drift and inconsistency in robot embodiment over time.

- **Subject appearance stability.** As illustrated in Figure 11, (a) a rectangular knitted sleeve transforms into a long-sleeve sweater, and (b) a green plastic cup on the table becomes a round yellow mug, indicating a loss of identity-preserving appearance.

Beyond these explicit examples, we also observe a range of additional anomalies: changes in the number of robot links or arms during task execution, the spontaneous generation of extra manipulators, and unnatural variations in arm length, connectivity, or joint bending direction over time. Target objects may also undergo unrealistic material changes, such as a rigid object bending like a deformable one. The Robot-Subject Stability metric is designed to capture such inconsistencies, providing a focused measure of whether the model can preserve both robot morphology and object identity across the video sequence.

D.4. Motion Amplitude

Motivation. A common failure mode in robotic video generation is that the robot remains nearly static while the generated frames appear visually smooth, as illustrated in Figure 12(a)(b). This makes pure smoothness-based metrics insufficient. Following the perceptual motion estimation idea introduced in VMBench (Ling et al., 2025), a *Motion Amplitude Score (MAS)* is used to measure the perceptible dynamic behavior of the robot while explicitly compensating for camera motion.

Robot Localization and Tracking. The robot is first localized using GroundingDINO, and temporally stable segmentation masks are obtained via SAM2. CoTracker is then used to track a dense grid of keypoints inside the robot mask, ensuring that the estimated motion truly reflects robot articulation rather than background drift or mask leakage.

Frame-Level Motion. Let $\mathbf{p}_{t,k}$ denote the 2D location of the k-th tracked point at frame t. The raw frame-to-frame

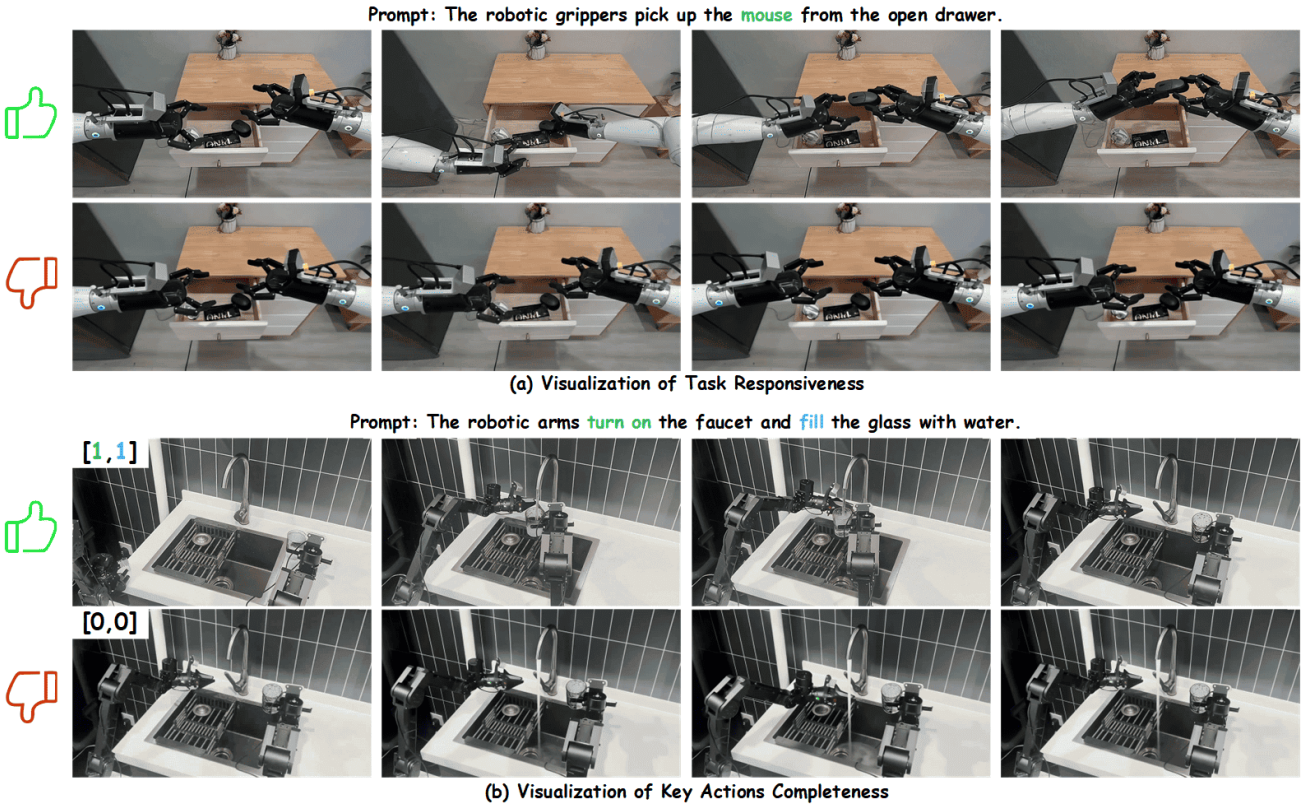

Figure 9. **Visualization of task responsiveness and key actions completeness.**

displacement is computed as:

$$\bar{D}_t = \frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{p}_{t,k} - \mathbf{p}_{t-1,k}\right\|_2. \tag{6}$$

To ensure consistency across resolutions, the motion is normalized by the video diagonal:

$$\tilde{D}_t = \frac{\bar{D}_t}{\sqrt{W^2 + H^2}}. \tag{7}$$

Camera-Motion Compensation. To estimate camera-induced movement, the robot mask is inverted and the same tracking procedure is applied to the background region. Let $\tilde{D}_t^{\text{bg}}$ denote the normalized background motion.

A *soft-zero* strategy is adopted: if the robot motion does not exceed the background motion, the small residual value is retained:

$$\hat{D}_t = \begin{cases} \tilde{D}_t - \tilde{D}_t^{\text{bg}}, & \tilde{D}_t > \tilde{D}_t^{\text{bg}}, \\ \tilde{D}_t, & \tilde{D}_t \le \tilde{D}_t^{\text{bg}}. \end{cases} \tag{8}$$

This behavior matches our implementation and improves robustness against tracking noise or partial occlusion, while effectively treating the robot as "static".

Final Score. Finally, following VMBench, the compensated displacement is clipped to stabilize extreme values:

$$\text{MAS} = \frac{1}{T}\sum_{t=1}^{T}\min\left(\hat{D}_t,\, 1\right). \tag{9}$$

Discussion. MAS captures whether the robot exhibits meaningful articulation rather than merely inheriting background or camera movement. By incorporating localization, mask-based tracking, background compensation, and a soft-zero strategy, MAS remains stable across scenes, tracking configurations, and robotic embodiments.

D.5. Motion Smoothness

This metric evaluates the temporal continuity and naturalness of motion, aiming to detect frame-level discontinuities such as low-level temporal artifacts and high-level motion blur. As illustrated in Figure 13, various robot embodiments, including quadruped robots, humanoids, and single-arm manipulators, exhibit different degrees of motion-induced distortion that substantially degrade perceived video quality.

The assessment is based on the motion-smoothness principle introduced in VMBench (Ling et al., 2025), with temporal consistency estimated using Q-Align aesthetic quality

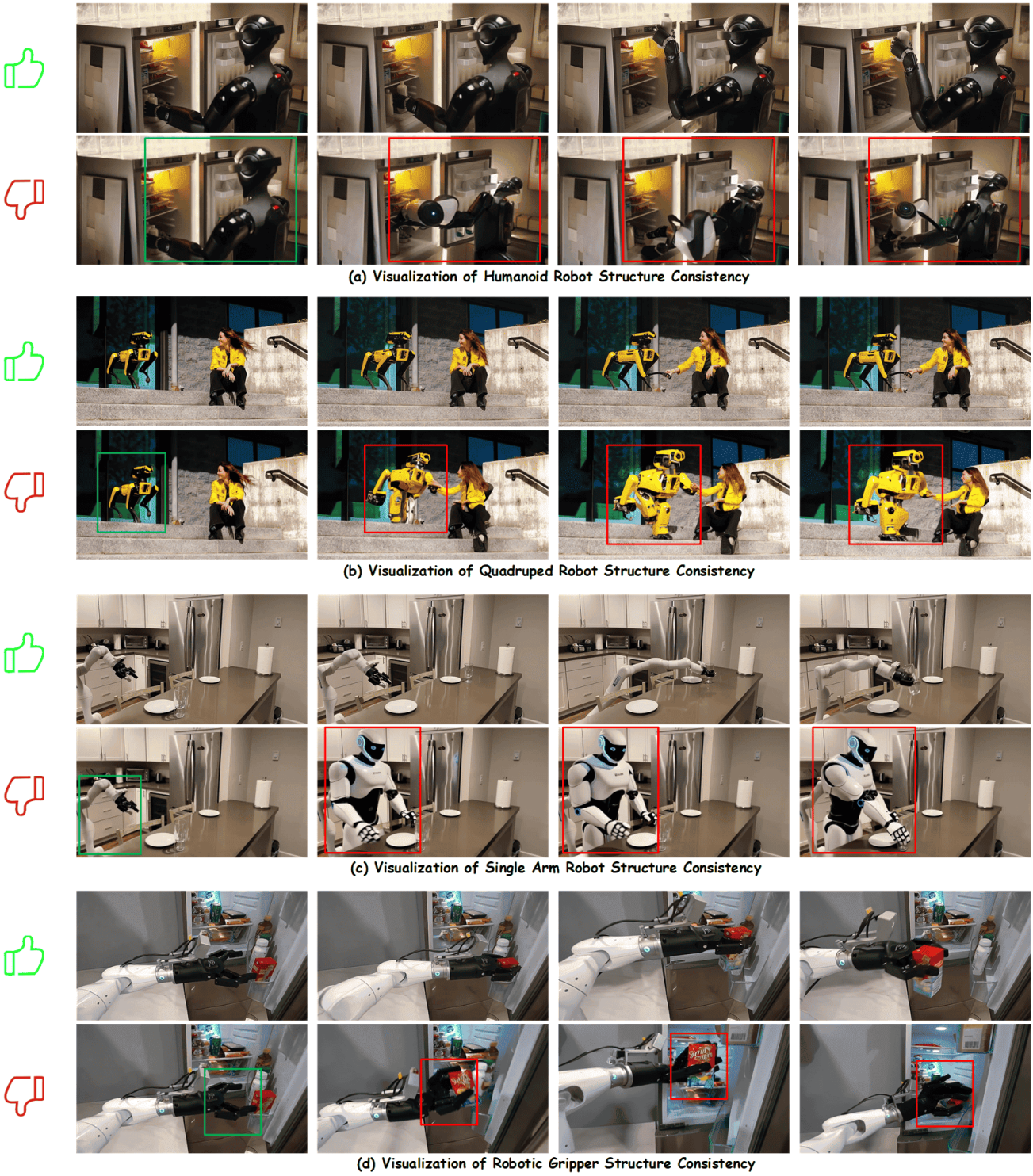

Figure 10. **Visualization of robot structural stability.**

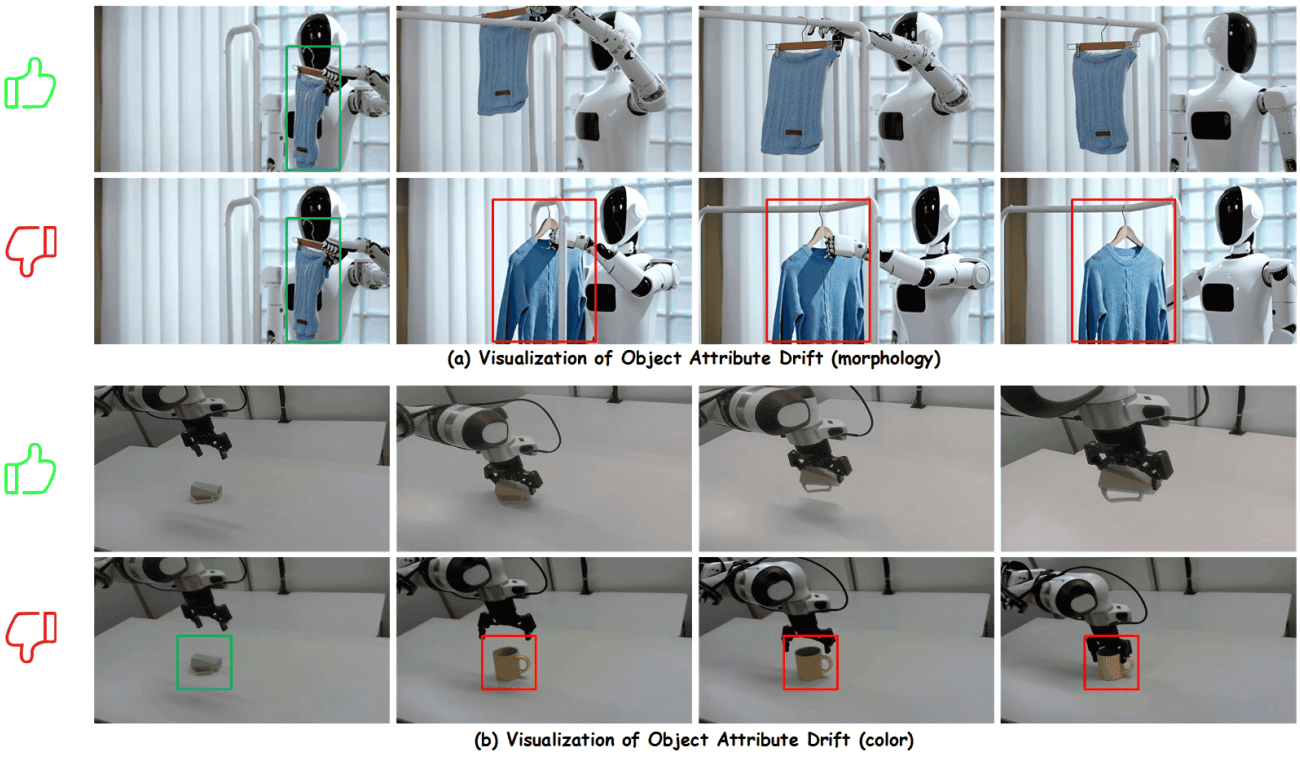

Figure 11. **Visualization of subject appearance stability.**

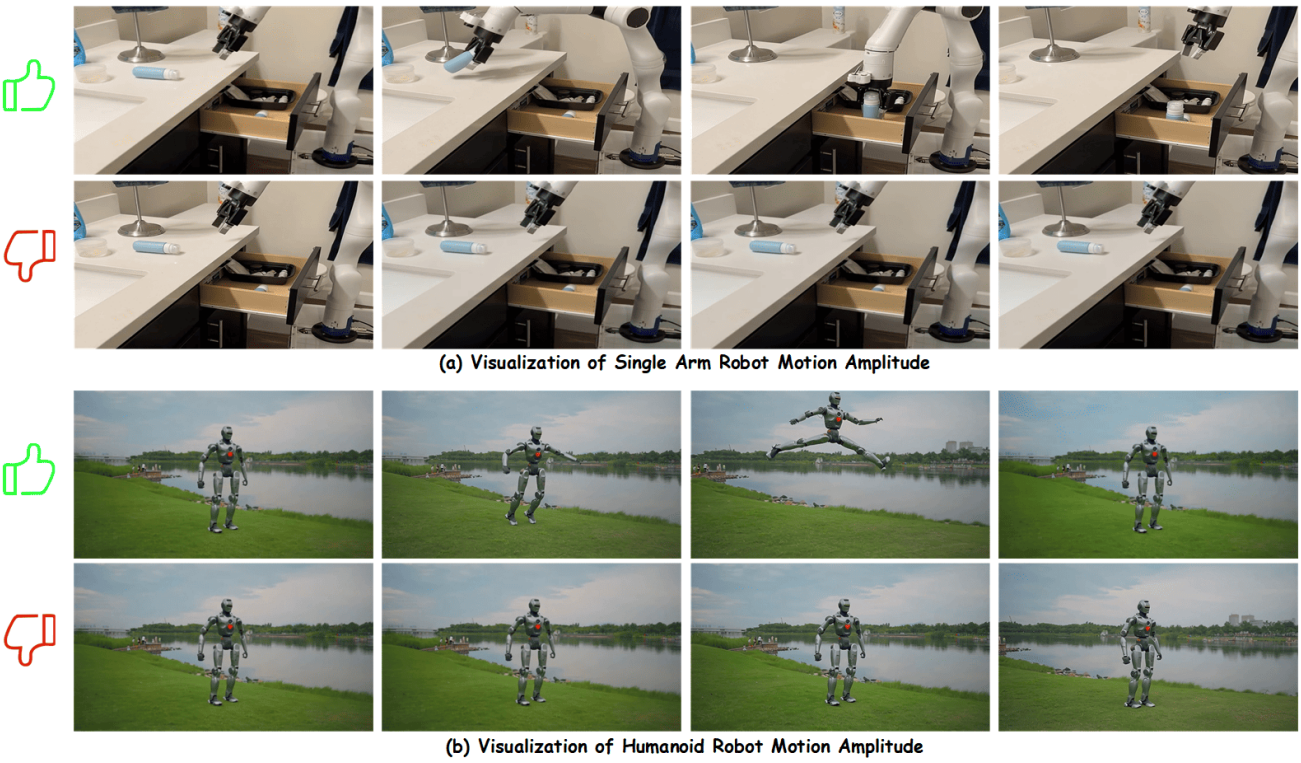

Figure 12. **Visualization of robot motion amplitude.**

scores. For each video, frames are processed with a sliding window of size w (default $w = 3$). Each window is fed into Q-Align to obtain a per-frame quality score sequence $\{Q_t\}_{t=1}^{T}$. Temporal quality fluctuation is then measured by the magnitude of adjacent-frame differences:

$$\Delta Q_t \;=\; Q(f_{t-1}) - Q(f_t). \tag{10}$$

To ensure comparability across videos with different motion intensities, the threshold for detecting abnormal temporal variations is determined by the *Motion Amplitude* value m defined in Section D.4. A piecewise adaptive threshold function is used:

$$\tau_s(m) = \begin{cases} 0.01, & m < 0.1, \\ 0.015, & 0.1 \le m < 0.3, \\ 0.025, & 0.3 \le m < 0.5, \\ 0.03, & m \ge 0.5. \end{cases} \tag{11}$$

Lower-motion videos therefore adopt a stricter threshold for detecting subtle temporal inconsistencies, while higher-motion videos receive a relaxed threshold to avoid penalizing naturally rapid movements. The function is determined through grid search on a validation split to ensure reproducibility.

A frame t is marked as temporally abnormal if its score fluctuation exceeds the adaptive threshold:

$$I_t = \mathbb{I}[\Delta Q_t > \tau_s(m)], \tag{12}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. To robustly capture abrupt artifacts such as frame drops or transient distortions, adjacent frames of each abnormal index are also flagged.

The final Motion Smoothness Score (MSS) is computed as the proportion of "normal" frames in the entire sequence:

$$\mathrm{MSS} = 1 - \frac{1}{T}\sum_{t=2}^{T} I_t. \tag{13}$$

A higher MSS indicates smoother and more temporally coherent motion, whereas videos with frequent jitter, abrupt discontinuities, or artifact-heavy transitions yield lower MSS values.

D.6. Score Aggregation

We consolidate five fine-grained evaluation signals into two final indicators, *Task Completion* and *Visual Quality*.

Notation. We denote the normalized values of the five fine-grained metrics as follows: (1) **PSS**: Physical-Semantic Plausibility, (2) **TAC**: Task-Adherence Consistency, (3) **RSS**: Robot-Subject Stability, (4) **MS**: Motion Smoothness, (5) **MA**: Motion Amplitude.

Normalization. Given a raw metric value s defined over range $[s_{\min}, s_{\max}]$, its normalized value is

$$s \leftarrow \mathrm{clip}_{[0,1]}\left(\frac{s - s_{\min}}{s_{\max} - s_{\min}}\right). \tag{14}$$

Penalty terms. Two penalty terms are used to down-weight videos with insufficient subject motion or unstable visual composition.

Motion-amplitude penalty. Let MA denote the normalized motion amplitude. A soft penalty is applied when MA falls below the threshold t:

$$P_{\mathrm{MA}}(\mathrm{MA}) = \begin{cases} (t - \mathrm{MA}) + \delta, & \mathrm{MA} < t_{\mathrm{low}}, \\ t - \mathrm{MA}, & t_{\mathrm{low}} \le \mathrm{MA} < t, \\ 0, & \mathrm{MA} \ge t, \end{cases} \tag{15}$$

with $t = 0.1$, $t_{\mathrm{low}} = 0.05$, and $\delta = 0.1$.

Stability-consistency penalty. Robot and object level stability grades are mapped to penalty magnitudes:

$$p(g) \in \{0.2, 0.4, 0.6, 0.8\} \quad \text{for grades } g \in \{B, C, D, E\}, \tag{16}$$

while grade A incurs zero penalty. Let g_r and g_o denote robot- and object-related stability grades:

$$P_{\mathrm{RSS}} = \begin{cases} \dfrac{p(g_r) + p(g_o)}{2}, & \text{if both exist}, \\ p(g_r), & \text{if only } g_r \text{ exists}, \\ 0, & \text{otherwise}. \end{cases} \tag{17}$$

Final indicators.

Task Completion (TC). Task correctness is computed from Physical-Semantic Plausibility (PSS) and Task-Adherence Consistency (TAC):

$$\mathrm{TC} = \frac{\mathrm{PSS} + \mathrm{TAC}}{2}. \tag{18}$$

Visual Quality (VQ). Visual realism and temporal coherence are expressed as a weighted combination of RSS and MS, penalized by low motion amplitude and visual instability:

$$\mathrm{VQ} = \max\Big(0,\; 0.8\cdot\mathrm{RSS} + 0.2\cdot\mathrm{MS} - P_{\mathrm{MA}}(\mathrm{MA}) - P_{\mathrm{RSS}}\Big). \tag{19}$$

Model-level aggregation. For each model, TC and VQ are computed for all evaluation samples. The final model score corresponds to the mean values of TC and VQ across samples, which are used for quantitative comparison and ranking in our benchmark.

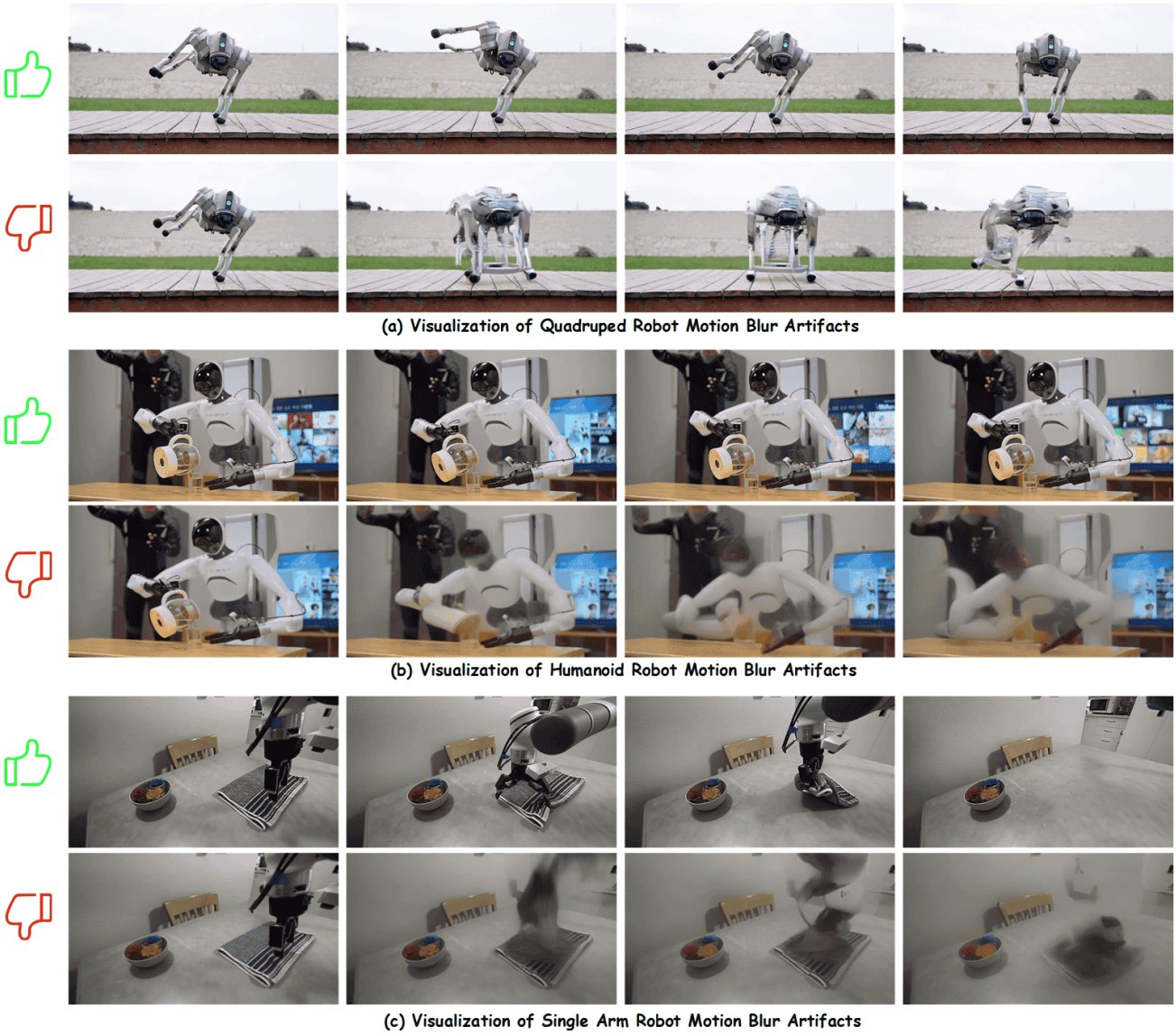

Figure 13. **Visualization of robot motion smoothness.**

E. Model Descriptions and Implementation Setups

E.1. Commercial Models

Wan 2.6. Wan is a comprehensive family of open-source foundational video generative models built on the Diffusion Transformer architecture. It supports multiple downstream tasks including T2V, I2V, editing, inpainting, and video-to-audio. We use the official API to generate 5-second 720P videos at 30 fps.

Wan 2.5. Wan is a comprehensive family of open-source foundational video generative models built on the Diffusion Transformer architecture. It supports multiple downstream tasks including T2V, I2V, editing, inpainting, and video-to-audio. We use the official API to generate 5-second 720P videos at 24 fps.

Hailuo. Hailuo provides multimodal models for T2V, I2V, and T2A tasks, supporting resolutions up to 1080p and long-duration outputs with high temporal coherence. We generate 6 second videos at 1364 $\times$ 768 and 24 fps using the official Hailuo API.

Veo3. Veo 3 is Google's latest foundational video generation model supporting high-resolution (1080p), long-duration (up to 60 seconds), and audio-integrated video synthesis using a large-scale Diffusion Transformer. We use the official Veo 3 API to generate videos up to 720p, 8 seconds, and 24 fps.

Kling 2.6 pro. We use the official Kling 2.6 pro model with default parameters to generate a 5-second video at 1920 $\times$ 1080 resolution and 24 fps.

Seedance 1.0. Seedance 1.0 is a large-scale video generation model from ByteDance, supporting text-to-video and image-to-video generation with high aesthetic quality and temporal stability. It integrates a 3D causal VAE with a 4$\times$16$\times$16 compression ratio. We use the Seedance 1.0 model to generate 5-second videos at 1280 $\times$ 720 resolution and 24 fps.

Seedance 1.5. Seedance 1.5 is a large-scale video generation model from ByteDance, supporting text-to-video and image-to-video generation with high aesthetic quality and temporal stability. We use the Seedance 1.5 model to generate 5-second videos at 1280 $\times$ 720 resolution and 24 fps.

Sora v1. We use the official Sora v1 model with default parameters to generate a 5-second video at 1280 $\times$ 720 resolution and 30 fps.

Sora v2 Pro. We use the official Sora v2 Pro model with default parameters to generate a 4-second video at 1280 $\times$ 720 resolution and 30 fps.

E.2. Open-source Models

Wan2.2_A14B. Wan2.2_A14B is an open-source large-scale video generation model that incorporates a Mixture-of-Experts (MoE) architecture. This architecture dynamically allocates specialized expert networks to enhance the model's capacity and temporal understanding. It supports multimodal inputs (text and images) for generating open-domain videos at 1280 $\times$ 720 resolution (720P), with a typical duration of 5 seconds (120 frames) at 24 fps. Compared to its predecessors, Wan2.2_A14B demonstrates superior performance in generalizing across diverse scenes, modeling complex motions, and achieving fine-grained aesthetic control. The technical report and model weights are publicly available. We utilize the official Wan2.2_A14B model with its default parameters to generate a 5-second video (81 frames) at a spatial resolution of 1280 $\times$ 720 and a frame rate of 16 fps. The model supports both text and image inputs.

LongCat-Video. LongCat-Video is an open-source foundational video generation model with 13.6B parameters, developed by the Meituan LongCat Team. It unifies text-to-video (T2V), image-to-video (I2V), and video continuation (VC) tasks within a single Diffusion Transformer (DiT) architecture, supporting efficient minute-long video generation without quality degradation. The model employs multi-reward RLHF optimization (Group Relative Policy Optimization) to enhance visual quality, motion coherence, and text alignment. We employ the official LongCat-Video model with default configurations, generating 1280 $\times$ 704 resolution videos at 15 fps using a coarse-to-fine generation strategy.

Wan2.2_5B. Wan2.2_5B is a medium-scale model in the Wan2.2 series with 5 billion parameters, utilizing a Transformer backbone architecture. It supports multi-modal video generation from text and images, generating 5-second (121 frames) open-domain videos at 1280 $\times$ 720 and 24 fps. Model weights and documentation are fully open-sourced. We use the official Wan2.2_5B model with default settings to generate a 5-second video (120 frames) at a resolution of 1248 $\times$ 704 and 24 fps.

Wan2.1_14B. Wan2.1_14B is an early large-parameter video generation model in the Wan series with 14 billion parameters, based on a multi-modal diffusion architecture. It generates 5-second (120-frame) open-domain videos at 1280 $\times$ 720 and 24 fps, emphasizing complex scene modeling and object motion understanding. We use the official Wan2.1_14B model with default parameters to generate a 5-second video (81 frames) at 832 $\times$ 480 resolution and 16 fps.

HunyuanVideo. We use the official HunyuanVideo model with default parameters to generate a 5-second video

at 1248 × 704 resolution and 24 fps.

HunyuanVideo 1.5. We use the official HunyuanVideo 1.5 model with default parameters to generate a 5-second video at 848 × 480 resolution and 24 fps.

SkyReels-V2. SkyReels-V2 is an open-source infinite-length film generative model developed by Skywork AI. It supports text-to-video, image-to-video, and video continuation tasks, with modules including SkyCaptioner-V1, multi-stage pretraining, RL for motion quality, and a diffusion forcing framework for long video synthesis. We use the official SkyReels-V2 model with default settings to generate 960 × 544 resolution videos at 24 fps for approximately 4 seconds.

LTX-Video. LTX-Video is an open-source transformer-based latent diffusion model developed by Lightricks. It integrates a Video-VAE and denoising transformer with a 1:192 compression ratio using spatiotemporal downscaling, enabling efficient latent-space processing. It supports both text-to-video and image-to-video generation. We use the official LTX-Video model with default parameters to generate a 5-second video at 832 × 480 resolution and 16 fps.

LTX-2. LTX-2 is an open-source transformer-based latent diffusion model. We use the official LTX-2 model with default parameters to generate a 5-second video at 1536 × 1024 resolution and 24 fps.

FramePack. FramePack is a neural structure for next-frame prediction designed to avoid forgetting and drifting in video generation. It compresses input frames by importance to maintain a fixed transformer context, enabling long video synthesis with low computational cost. It supports T2V and I2V and can be integrated with models such as HunyuanVideo or Wan. We employ FramePack with a base model to generate 5-second videos at 832 × 480 resolution.

CogVideoX. CogVideoX is an open-source text-to-video generation model employing a diffusion transformer and a 3D VAE. It introduces expert-adaptive LayerNorm and multi-resolution frame packing to enhance motion consistency. It produces up to 10-second videos at 720 × 1280 and 16 fps. We use the official CogVideoX weights and inference scripts with default settings to generate 6-second videos (160 frames) at 720 × 480 and 8 fps.

E.3. Robotics-specific Models

Cosmos 2.5. Cosmos-Predict2.5 and Cosmos-Transfer2.5 are NVIDIA's world simulation foundation models for physical AI, supporting Text2World, Image2World, and Video2World simulations with flow-based architectures. We generate videos at 1280 × 720 resolution and 16 fps using official NVIDIA codebase.

DreamGen (GR). GRL is DreamGen's generalizable robot learner model integrating synthetic neural trajectories and real teleoperation data for policy learning. We generate all trajectory videos at 768 × 432 resolution and 16 fps.

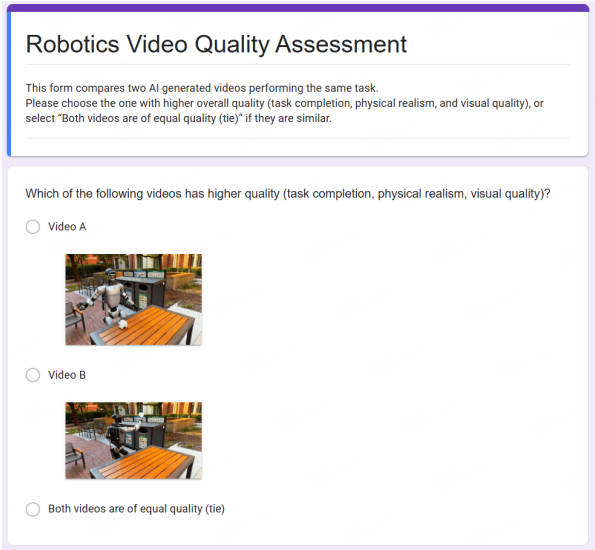

Figure 14. **Visualization of the Questionnaire for User Study**

DreamGen (DROID). DROID is a dynamics-aware imitation learning model using video diffusion with inverse dynamics modeling for realistic physical trajectories. We generate all videos at up to 768 × 432 resolution and 16 fps.

UnifoLM-WMA-0. UnifoLM-WMA-0 integrates video world modeling and action policy learning for generalizable robotic reasoning and control. We generate videos up to 320 × 512 resolution, 30 fps.

Vidar. Vidar is an embodied video diffusion model designed for robotic manipulation, integrating a strong video diffusion prior and a Masked Inverse Dynamics Model (MIDM). Vidar generates 4-second multi-view interactive videos at 704 × 480 resolution and 15 fps with automatically annotated action trajectories.

F. Human Preference Study Details

To complement the main paper's human preference analysis, this section provides additional statistical examinations that further characterize the agreement between human judgments and our automatic benchmark. The questionnaire used in our human preference study is shown in Figure 14. As described in the main paper, thirty participants compared pairs of generated videos for the same prompt and selected the better one (or "Tie"). These votes were converted into per-model human scores using the $5/3/1$ win-tie-loss scheme, and the resulting model-level ranking exhibited a strong correlation with RBench ($\rho = 0.96$).

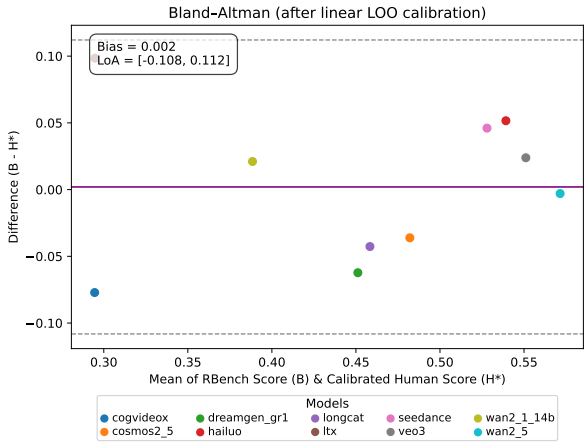

Figure 15. Bland–Altman plot after linear leave-one-out calibration (H^*). Points are models; x-axis $m_i = \frac{B_i+H_i^*}{2}$, y-axis $d_i = B_i - H_i^*$. The solid line indicates the bias ($\bar{d}$); dashed lines show the 95% limits of agreement (LoA). In our study the legend reports Bias $= 0.002$ and LoA $= [-0.108,\ 0.112]$.

While rank correlation quantifies the consistency of *ordering*, it does not measure whether the two scoring methods agree in an *absolute* sense (i.e., whether they assign comparable magnitudes). To assess this complementary notion of agreement, we conduct a Bland-Altman analysis between human scores and benchmark scores. Because the two score scales may differ by a systematic offset or slope, we additionally apply a leave-one-out (LOO) linear calibration to correct for potential scale mismatch without overfitting. The following subsections detail the calibration procedure, Bland-Altman computation, accompanying figures, and per-model statistics.

Bland–Altman basics. Bland–Altman analysis assesses *agreement* between two measurement methods by plotting, for each item i, the difference $d_i = B_i - H_i$ against the mean $m_i = \frac{B_i+H_i}{2}$, where H and B denote human and automatic scores, respectively. As shown in Fig. 15, the horizontal solid line represents the *bias* $\bar{d} = \frac{1}{n}\sum_i d_i$ (average difference). The dashed lines denote the *95% limits of agreement (LoA)*: $\bar{d} \pm 1.96\, s_d$, where s_d is the standard deviation of the differences. Narrower LoA and small bias indicate stronger agreement beyond mere correlation.

Linear LOO calibration. Human and automatic scores can exhibit *scale* mismatch. To mitigate such systematic differences, we apply a *leave-one-out (LOO) linear calibration* to the benchmark.

Mathematical formulation. Let $\{(B_i, H_i)\}_{i=1}^n$ be the automatic benchmark and human scores for n models. For each $i \in \{1, \ldots, n\}$, define the training index set $S_{-i} = \{j : j \neq i\}$ and estimate the OLS (Ordinary Least Squares) calibration parameters by

$$(\hat{\alpha}_{-i}, \hat{\beta}_{-i}) \;=\; \arg\min_{\alpha,\beta} \sum_{j \in S_{-i}} \big(H_j - \alpha - \beta B_j\big)^2. \quad (20)$$

The calibrated score for model i is

$$\hat{H}_i \;=\; \hat{\alpha}_{-i} + \hat{\beta}_{-i}\, B_i. \quad (21)$$

Bland–Altman with calibrated scores. We then perform Bland–Altman analysis on (B, H^*) by computing, for each i,

$$d_i \;=\; B_i - H_i^*, \qquad m_i \;=\; \frac{B_i+H_i^*}{2}. \quad (22)$$

The bias $\bar{d}$ and the 95% LoA $\bar{d} \pm 1.96\, s_d$ are computed from $\{d_i\}_{i=1}^n$ as in the basics above.

Figure 15 shows the Bland-Altman plot after linear LOO calibration, with the bias and LoA reported in the legend. Table 5 lists, for each model, the mean m_i, the difference d_i, and the corresponding LOO calibration parameters $\hat{\alpha}_{-i}$ and $\hat{\beta}_{-i}$ used to obtain H_i^*. As a complement to the correlation analysis (Spearman $\rho = 0.96$ between human scores and the automatic benchmark), the Bland-Altman view quantifies absolute agreement: the small bias and tight LoA indicate that the calibrated benchmark scores are in close agreement with human judgments.

Model	$\mathrm{mean}(B, H^*)$	$\mathrm{diff}(B - H^*)$	α_{LOO}	β_{LOO}
LTX-Video	0.295	0.098	0.141	0.613
Hailuo	0.539	0.052	0.205	0.496
Seedance 1.0	0.528	0.046	0.203	0.500
Veo 3	0.551	0.024	0.204	0.503
Wan2.1 14b	0.388	0.021	0.193	0.524
Wan2.5	0.571	-0.003	0.199	0.517
Cosmos2.5	0.482	-0.036	0.200	0.523
LongCat-Video	0.458	-0.043	0.202	0.520
Dreamgen	0.451	-0.062	0.203	0.522
Cogvideox-5B	0.295	-0.077	0.242	0.446

Table 5. Bland–Altman statistics after linear LOO calibration. $m_i = \frac{B_i+H_i^*}{2}$, $d_i = B_i - H_i^*$. α_{LOO} and β_{LOO} are the intercept and slope learned from all *other* models when calibrating H_i^*. All values are rounded to three decimals.

G. Additional Analysis of VLM-based Evaluators

This section provides additional analysis of the VLM-based evaluators used in RBench, including evaluator selection, cross-evaluator consistency, ranking stability, robustness to prompt variations, and potential evaluator bias.

G.1. Evaluator Selection

RBench adopts two VLM-based evaluators, GPT-5 and Qwen3-VL, to assess video generation quality under structured physical and semantic criteria. We select these two

models because they represent strong closed-source and open-source VLMs, respectively, and are developed by different organizations with different model architectures, training data, and deployment settings. This design reduces the risk that the benchmark outcome depends on a single model family or a specific training ecosystem.

We also conducted preliminary trials with other mainstream VLMs, including InternVL and LLaVA-series models. In our robotic video evaluation setting, these models showed weaker fine-grained physical reasoning ability, especially when detecting subtle violations such as implausible contacts, incorrect object motion, unstable robot-object interactions, or temporally inconsistent manipulation outcomes. These limitations introduced larger score noise in complex robotic scenarios. Therefore, we use GPT-5 and Qwen3-VL as the primary VLM evaluators to improve evaluation reliability.

G.2. Cross-evaluator Consistency

To verify whether the benchmark results are sensitive to the choice of VLM evaluator, we compare the rankings obtained from Qwen3-VL and GPT-5. As shown in Table 6, the two evaluators produce highly consistent model scores and rankings, with a Pearson correlation of $r = 0.9911$ and a Spearman correlation of $\rho = 0.9860$. These results indicate that the evaluation outcome is largely invariant to the evaluator choice.

This consistency also suggests that RBench captures stable physical and semantic quality differences among video generation models, rather than evaluator-specific preferences. In particular, if the closed-source evaluator systematically favored models from a specific ecosystem, we would expect those models to receive noticeably higher rankings under GPT-5 than under Qwen3-VL. However, models such as Sora v1 and Sora v2 Pro remain lower ranked under both evaluators, which does not support the assumption of systematic evaluator bias.

G.3. Prompt Sensitivity Analysis

We further study the robustness of VLM-based evaluation to prompt variations. Specifically, we construct three semantically equivalent variants of the evaluation prompt and evaluate five representative models on a subset of 50 samples. The resulting model rankings remain unchanged across the prompt variants, with Kendall's $\tau = 1.0$. The absolute score fluctuations are within ± 0.01, which is smaller than the score gaps between adjacent models.

These results show that the structured VQA protocol effectively mitigates prompt sensitivity. Since the evaluator is guided by predefined questions and scoring criteria, minor changes in prompt wording do not substantially affect the final ranking.

Table 6. Comparison of model scores and rankings under Qwen3-VL and `gpt-5-2025-08-07`. The two evaluators show strong agreement, with Pearson correlation $r = 0.9911$ and Spearman correlation $\rho = 0.9860$.

Model	Qwen Score	GPT-5 Score	Rank Qwen	Rank GPT-5	Diff. Abs.
Wan 2.5	0.7806	0.5702	1	1	0
Hailuo v2	0.7616	0.5647	2	2	0
Veo 3	0.7604	0.5626	3	3	0
Seedance 1.0	0.7552	0.5509	4	4	0
Wan2.2_A14B	0.6983	0.5074	5	5	0
Cosmos 2.5	0.6324	0.4644	6	6	0
LongCat-Video	0.6086	0.4373	7	7	0
DreamGen(gr1)	0.5747	0.4200	8	8	0
Wan2.2_5B	0.5508	0.3799	9	11	2
Wan2.1_14B	0.5416	0.3988	10	10	0
SkyReels	0.5314	0.3610	11	12	1
DreamGen(droid)	0.5139	0.4053	12	9	3
LTX-Video	0.4861	0.3439	13	14	1
Sora v2 Pro	0.4813	0.3607	14	13	1
FramePack	0.4528	0.3393	15	15	0
Sora v1	0.3723	0.2657	16	16	0
CogVideoX_5B	0.3218	0.2557	17	17	0
Vidar	0.2074	0.2062	18	18	0
UnifoLM-WMA-0	0.1040	0.1234	19	19	0

G.4. Bootstrap Sensitivity Analysis

We further evaluate the stability of RBench rankings under prompt resampling. Specifically, we perform 1,000 bootstrap iterations, where each iteration randomly samples 80% of the prompts and recomputes the model rankings. As shown in Table 7, the top-ranked models exhibit high Top-5 retention, and the lowest-ranked models also show high Bottom-5 retention. This indicates that the rankings at both ends of the leaderboard are stable under prompt perturbations. In contrast, middle-ranked models show slightly larger rank fluctuations, suggesting that small score differences in this region should be interpreted with appropriate caution.

G.5. Reproducibility and API-version Control

We acknowledge that closed-source VLM APIs may change over time, which can affect evaluation reproducibility. To mitigate this issue, we fix the closed-source evaluator to a specific API version, `gpt-5-2025-08-07`, and keep the evaluation parameters fixed throughout all experiments. For the open-source evaluator, we use local Qwen3-VL inference, which avoids dependency on external API changes and improves reproducibility.

We also observe that evaluation failures are more likely to occur in complex scenarios, such as multi-object interaction,

Table 7. Bootstrap sensitivity analysis of RBench ranking stability under prompt resampling. We perform 1,000 bootstrap iterations by randomly sampling 80% of prompts in each iteration and recomputing the rankings.

Model	Score	Rank	95% CI	Med. Rank	Top-5 Ret.	Bot.-5 Ret.
Wan 2.6	0.607	1	[0.588, 0.626]	1	100%	0.0%
Seedance 1.5 pro	0.584	2	[0.565, 0.603]	2	98.9%	0.0%
Wan 2.5	0.570	3	[0.551, 0.589]	3	92.6%	0.0%
Hailuo	0.565	4	[0.546, 0.584]	4	90.8%	0.0%
Veo 3	0.563	5	[0.542, 0.580]	5	82.4%	0.0%
Seedance 1.0	0.551	6	[0.532, 0.570]	6	15.7%	0.0%
Kling 2.6 pro	0.534	7	[0.515, 0.553]	7	14.8%	0.0%
Wan2.2_A14B	0.507	8	[0.488, 0.527]	8	3.6%	0.0%
Cosmos 2.5	0.464	9	[0.445, 0.484]	9	1.2%	0.0%
HunyuanVideo 1.5	0.460	10	[0.441, 0.480]	10	0.0%	0.0%
LongCat-Video	0.437	11	[0.418, 0.457]	11	0.0%	0.0%
DreamGen(gr1)	0.420	12	[0.401, 0.439]	12	0.0%	0.0%
DreamGen(droid)	0.405	13	[0.386, 0.424]	13	0.0%	0.0%
Wan2.1_14B	0.399	14	[0.380, 0.418]	14	0.0%	0.0%
LTX-2	0.381	15	[0.362, 0.401]	15	0.0%	0.0%
Wan2.2_5B	0.380	16	[0.361, 0.399]	16	0.0%	0.0%
Sora v2 Pro	0.362	17	[0.342, 0.379]	18	0.0%	3.4%
SkyReels	0.361	18	[0.342, 0.380]	17	0.0%	3.0%
LTX-Video	0.344	19	[0.325, 0.362]	19	0.0%	17.3%
FramePack	0.339	20	[0.321, 0.358]	20	0.0%	19.8%
HunyuanVideo	0.303	21	[0.285, 0.321]	21	0.0%	70.5%
Sora v1	0.266	22	[0.249, 0.283]	22	0.0%	88.7%
CogVideoX_5B	0.256	23	[0.239, 0.272]	23	0.0%	97.3%
Vidar	0.206	24	[0.190, 0.222]	24	0.0%	100%
UnifoLM-WMA-0	0.123	25	[0.111, 0.136]	25	0.0%	100%

long-horizon planning, and tasks involving subtle physical state changes. These cases remain challenging for current VLM evaluators. Nevertheless, the high cross-evaluator consistency, low prompt sensitivity, and strong human alignment indicate that the current evaluation protocol provides a reliable basis for comparing robotic video generation models.

H. Prompt Template

This section introduces the prompt design used for the Visual Reasoning task in our MLLM-based evaluation pipeline. Among all tasks, Visual Reasoning involves the most structured form of reasoning and thus provides a representative example for illustrating our prompt design. We provide the MLLM with essential contextual information recorded during dataset construction, including the video viewpoint, a high-level content description, and the identities of the robotic manipulator and manipulated object. This background knowledge serves as the foundation for the model's subsequent reasoning and scoring process.

Unlike other tasks, Visual Reasoning requires explicit verification of logical dependencies between robot actions. Therefore, it adopts a two-part structure consisting of a question-chain constructor and a video assessment prompt. The full templates are shown below.

Question-chain construction. The first component converts the original text instruction into a sequence of binary verification questions (Figure 16). The model analyzes the semantics of the instruction and generates a short chain of stepwise questions that reflect the intended causal and temporal structure of the robot's behavior. This question chain acts as an explicit reasoning scaffold and is subsequently fed into the main evaluation prompt.

Video assessment prompt. The second component integrates the generated question chain with a 3×2 grid of chronologically ordered frames extracted from the video (Figures 17 and 18). It also incorporates the contextual background information provided at the beginning—namely, the view perspective, video content summary, robotic manipulator type, and manipulated object identity. With these inputs, the prompt instructs the model to determine whether each reasoning step has been successfully completed and to evaluate the stability, consistency, and physical plausibility of the robot and objects throughout the sequence. A strict scoring protocol and a structured JSON output format ensure reproducible and interpretable evaluation across models.

I. Additional Qualitative Comparisons

This section provides supplementary qualitative results that extend the "Qualitative Comparison Across Representative Tasks" analysis presented in the main paper. For each of the five task categories in RBench, namely Common Manipulation, Long-Horizon Planning, Multi-Entity Collaboration, Spatial Relationship, and Visual Reasoning, we select two representative cases and visualize the generated videos from ten state-of-the-art image-to-video (I2V) models. As shown in Figures 19, 20, 21, 22, 23, these examples offer a more detailed inspection of model behaviors under diverse embodied scenarios, complementing the quantitative results discussed in the main text.

J. Comprehensive Quantitative Results

This section presents the complete quantitative evaluation results obtained from both GPT-based and Qwen-based evaluators. We report detailed per-model scores across all five tasks (Common Manipulation, Long-Horizon Planning, Multi-Entity Collaboration, Spatial Relationship, and Visual Reasoning) and four robot embodiments (Dual Arm, Humanoid, Single Arm, and Quadruped). To ensure clarity and readability, we summarize below the abbreviations used in all tables. For task-level metrics, each abbreviation corresponds to a specific VQA-style prompt used in our MLLM-based evaluation; the construction of these prompts and the computation of the total score (**TS**) are illustrated with representative example in Section H. The robot-level metrics are defined in Section D.

```
"""
   ### Task Description:
   You are an expert in video generation evaluation.
   The text prompt for the video generation model is {prompt}, and the domain is "visual reasoning in manipulation
video generation".
   Please strictly follow the requirements and examples below to generate a "step-by-step verification" question chain.
Evaluate whether the video perfectly implements the text prompt.

   ### Question Generation Rules:
   1. Each question must be based on clearly observable phenomena and allow for binary judgment (yes/no).
   2. Questions should cover: trigger → feedback → result, and be output in sequence.

   ### Example:
   [Input Prompt]:
   "The robot arranges the blocks on the table in the order of green, red, blue, and yellow from left to right from its own
perspective."

   [Output]:
   1. Does the robot pick up the green block first?
   2. Is the green block placed on the table on the far left?
   3. Does the robot pick up the red block next and place it to the right of the green block?
   4. Does the blue block follow the red block?
   5. Is the yellow block placed to the rightmost of the sequence?

   ### Generate Output:
   Now, for the given target text, generate a set of evaluation questions (3-5 questions).
"""
```

Figure 16. **Question-chain construction for Visual Reasoning.** This component analyzes the original instruction and transforms it into a sequence of binary verification questions that define causal and temporal dependencies in the robot's intended actions.

Task-level Metrics (per-task evaluation):

- **AES**: Action Execution Score
- **TCS**: Task Completion Score
- **OCS**: Object Consistency Score
- **RCS**: Robot Consistency Score
- **PSS**: Physical Semantic Score
- **ECR**: Event Completion Ratio
- **ECS**: Entity Consistency Score
- **ACS**: Action Coordination Score
- **SRS**: Spatial Relation Score
- **MFS**: Manipulation Feasibility Score
- **VRS**: Visual Reasoning Score
- **TS**: Total Score

Robot-level Metrics (per-robot embodiment evaluation):

- **PSS**: Physical Semantic Score
- **TAC**: Task-Adherence Consistency
- **RSS**: Robot–Subject Stability
- **MSS**: Motion Smoothness Score
- **MAS**: Motion Amplitude Score
- **TC**: Task Completion
- **VQ**: Visual Quality
- **TS**: Total Score

The following tables report the complete results for all models evaluated using both GPT and Qwen3-VL. As shown in Table 8, Table 9, Table 10, Table 11, and Table 12, the GPT-based evaluator provides assessments across five tasks.

Similarly, the Qwen-based results are presented in Table 17, Table 18, Table 19, Table 20, and Table 21.

For completeness, we also provide per-embodiment results. The detailed evaluation tables for dual-arm, humanoid, single-arm, and quadruped robots under GPT are shown in Table 13, Table 14, Table 15, and Table 16. The corresponding Qwen-based tables are provided in Table 22, Table 23, Table 24, and Table 25.

```
'''
You are shown a single image that is a 3 × 2 grid of chronologically ordered frames (read row-wise).
These frames are extracted from an AI-generated video recorded from a {view} perspective.
Video content: {description}
Robotic manipulator: {robotic_manipulator}
Manipulated object: {manipulated_object}

Your evaluation focus is:  Visual Reasoning.

Please evaluate the video from the following five aspects.
Each aspect receives a score from 1 to 5:
- If the aspect is judged as "No", assign 1 point.
- If "Yes", assign 2–5 points depending on quality (5 = perfect).
- If any aspect in Category A (Action Execution and Visual Reasoning) equals 1, the total score = 1.
- Otherwise, compute the mean of all five scores as the final score.
- BE STRICT WHEN SCORING — if any issue or imperfection is detected, assign 1 or 2 points decisively.

---

### Category A — Action Execution and Visual Reasoning
These aspects focus on how effectively and coherently the robot executes behaviors and whether the visual reasoning process aligns with the described goals.
1) Action Effectiveness
 - Check whether the robot's motion is physically reasonable (e.g., proper gripper closure, contact location, trajectory smoothness).
 - Reference scoring:
   1 = Motion discontinuous, incomplete, or physically implausible.
   2 = Basically correct and understandable motion.
   3 = Generally reasonable with slight inaccuracy.
   4 = Smooth motion and natural contact.
   5 = Fully consistent with physical and logical principles.

2) Visual Reasoning Accuracy
  Given the following visual reasoning questions, please determine whether each question has been accurately completed,
  calculate a completion rate, and ensure the evaluation of each question's completion is extremely strict.
 - Visual Reasoning Questions: {Questions-chain}
 - Collect ALL questions that are NOT completed into an array called missing_events. Excluding missed questions and obtain the completed_questions.
 - BE STRICT WHEN JUDGING WHETHER THE QUESTION IS COMPLETED
 - Reference scoring:
   score = 5 * (completed_questions ÷ total_questions), rounded to 1 decimals.

---
```

Figure 17. **Visual Reasoning evaluation prompt (Part I).** The first part of the evaluation prompt integrates the structured reasoning chain with contextual information extracted during dataset construction.

Category B — Visual and Physical Consistency
These aspects evaluate whether the visual and physical properties of the robot and objects remain stable and realistic throughout the video.
3) Manipulated Object Consistency
- manipulated object: **{manipulated_object}**
- Check whether the manipulated object maintains a consistent shape, structure, and outline over time.
- Note: Evaluate this aspect by comparing all frames to the first frame.
- Reference scoring:
  1 = Noticeable changes in appearance such as color, shape, or material; consistency not maintained.
  2 = Moderate differences but still consistent.
  3 = Minor deformation; overall consistency maintained.
  4 = Very small jitter or local artifacts only.
  5 = Completely stable and perfectly consistent appearance.

4) Robotic Manipulator Consistency
- Robotic Manipulator: **{robotic_manipulator}**
- Check whether the robotic entity, arm or gripper maintains stable geometry and articulation without disconnection or self-intersection.
- Note: Evaluate this aspect by comparing all frames to the first frame.
- Reference scoring:
  1 = Changes in the form, structure, or appearance of the manipulator or its subcomponents; consistency not maintained.
  2 = Moderate differences but still consistent.
  3 = Minor deformation; overall consistency maintained.
  4 = Very small jitter or local artifacts only.
  5 = Completely stable and perfectly consistent appearance.

5) Anomaly Check
- Examine whether any of the following issues are avoided:
  a) Floating or penetration (violation of physical grounding).
  b) Objects or robot parts suddenly appear or disappear between frames.
  c) Non-contact attachment or false grasp (object sticks to gripper without visible closure).
- Reference scoring:
  1 = Occurrence of any of the above anomalies.
  2 = No obvious anomalies but with noticeable artifacts or noise.
  3 = Slight noise affecting visual quality.
  4 = Only tiny visual imperfections.
  5 = No above anomalies.

[Final Scoring Rule]

If any aspect in Category A (Action Execution and Visual Reasoning) equals 1, the total score = 1.
Otherwise, compute the mean of all five scores as the final score.

[Output Format (Strict JSON, No Other Text)]
Each aspect and the total must include the following fields:
- "reason": a brief justification for the score
- "score": score

Example JSON structure:
```
{{
  "action_execution": {{"reason": "...", "score": 5}},
  "visual_reasoning_accuracy": {{"reason": "...", "score": 5}},
  "object_consistency": {{"reason": "...", "score": 4}},
  "manipulator_consistency": {{"reason": "...", "score": 3}},
  "anomaly_check": {{"reason": "...", "score": 3}},
  "total": {{"reason": "...", "score": 4}}
}}
  '''
```

Figure 18. **Visual Reasoning evaluation prompt (Part II).** The second part specifies the structured scoring protocol and the JSON output format used to ensure consistent and interpretable evaluation results.

Figure 19. **Visualization of Common Manipulation.** The first row contains the reference image, and the accompanying text shows the input prompt.

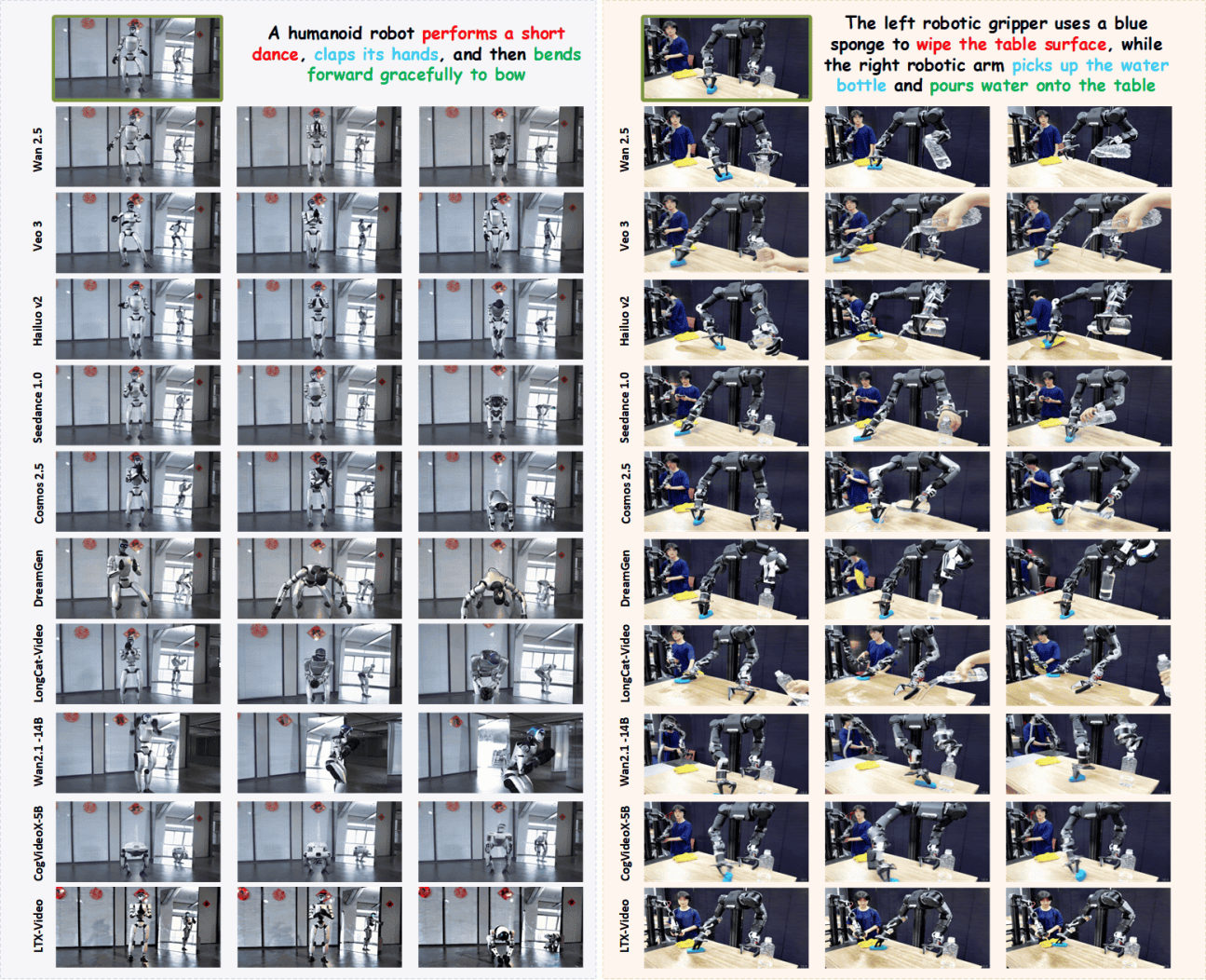

Figure 20. **Visualization of Long-Horizon Planning.** The first row contains the reference image, and the accompanying text shows the input prompt.

Figure 21. **Visualization of Multi-Entity Collaboration.** The first row contains the reference image, and the accompanying text shows the input prompt.

Figure 22. **Visualization of Spatial Relationship.** The first row contains the reference image, and the accompanying text shows the input prompt.

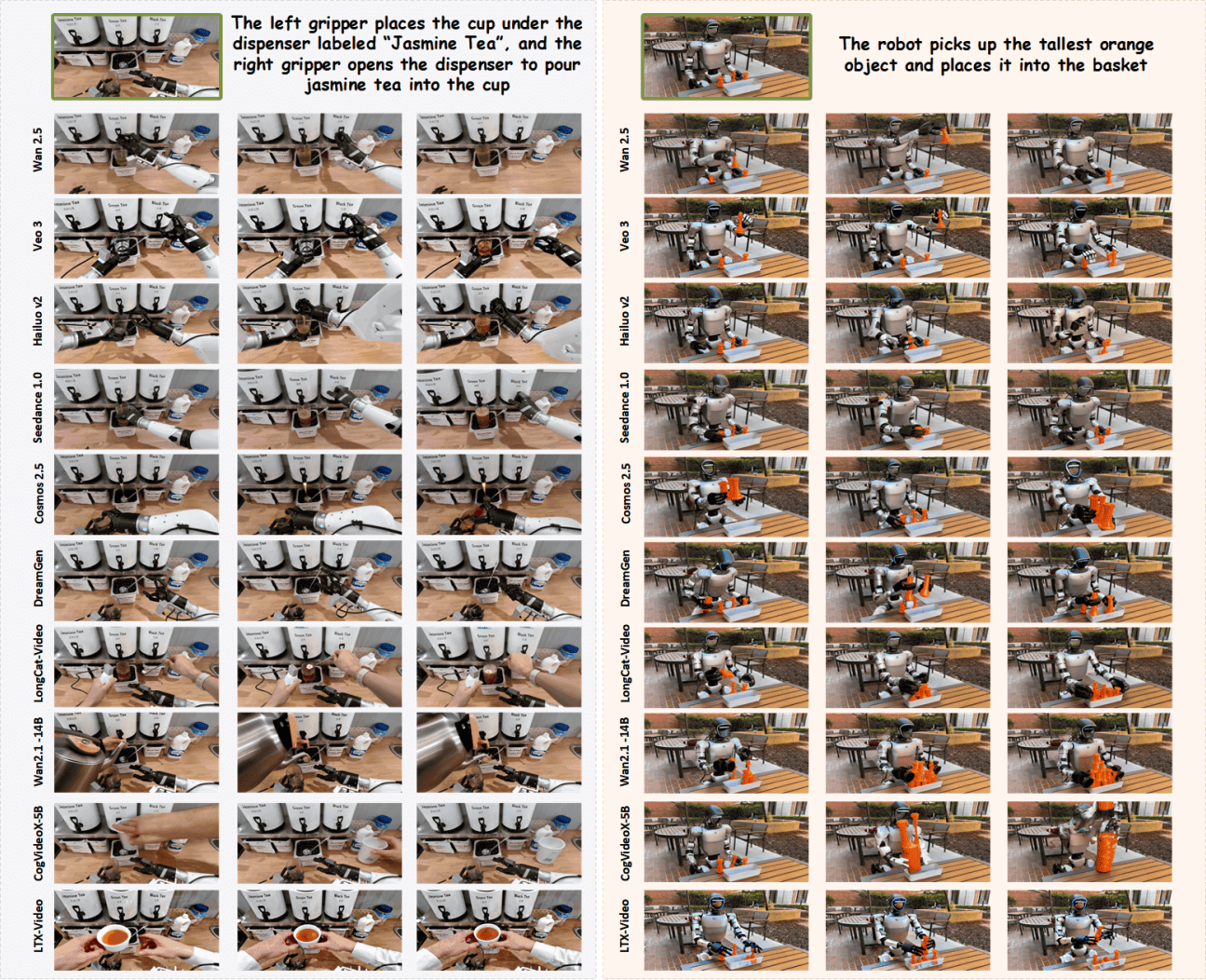

Figure 23. **Visualization of Visual Reasoning.** The first row contains the reference image, and the accompanying text shows the input prompt.

Table 8. **Results on Common Manipulation with GPT**

Model	AES	TCS	OCS	RCS	PSS	TS
Open-source						
Wan2.2_A14B	0.479	0.410	0.635	0.765	0.520	0.381
HunyuanVideo 1.5	0.505	0.480	0.575	0.695	0.490	0.442
LongCat-Video	0.469	0.408	0.591	0.739	0.510	0.371
Wan2.1_14B	0.446	0.375	0.552	0.692	0.510	0.344
LTX-2	0.340	0.330	0.515	0.620	0.380	0.284
Wan2.2_5B	0.416	0.395	0.540	0.670	0.495	0.331
Skyreels	0.348	0.230	0.545	0.740	0.465	0.202
LTX-Video	0.414	0.307	0.552	0.718	0.484	0.302
FramePack	0.346	0.188	0.637	0.739	0.556	0.205
HunyuanVideo	0.307	0.190	0.660	0.755	0.545	0.177
CogVideoX-5B	0.245	0.140	0.490	0.635	0.335	0.115
Closed-source						
Wan 2.6	0.581	0.596	0.637	0.750	0.668	0.545
Seedance 1.5 pro	0.654	0.642	0.591	0.750	0.556	0.576
Wan 2.5	0.565	0.600	0.635	0.770	0.605	0.527
Hailuo v2	0.576	0.625	0.625	0.745	0.595	0.559
Veo 3	0.572	0.602	0.607	0.729	0.540	0.520
Seedance 1.0	0.591	0.590	0.640	0.730	0.620	0.542
Kling 2.6 pro	0.561	0.565	0.610	0.760	0.590	0.528
Sora v2 Pro	0.354	0.229	0.637	0.719	0.561	0.207
Sora v1	0.280	0.170	0.445	0.605	0.360	0.151
Robotics-specific						
Cosmos 2.5	0.495	0.370	0.585	0.705	0.505	0.358
DreamGen(gr1)	0.420	0.341	0.566	0.683	0.469	0.311
DreamGen(droid)	0.420	0.369	0.604	0.640	0.421	0.358
Vidar	0.180	0.090	0.430	0.470	0.310	0.073
UnifoLM-WMA-0	0.105	0.045	0.400	0.300	0.220	0.036
Mean	**0.427**	**0.371**	**0.572**	**0.685**	**0.492**	**0.338**

Table 9. **Results on Long-Horizon Planning with GPT**

Model	AES	ECR	OCS	RCS	PSS	TS
Open-source						
Wan2.2_A14B	0.569	0.626	0.715	0.722	0.590	0.500
HunyuanVideo 1.5	0.537	0.527	0.606	0.725	0.537	0.437
LongCat-Video	0.507	0.485	0.598	0.742	0.598	0.384
Wan2.1_14B	0.482	0.404	0.589	0.687	0.553	0.335
LTX-2	0.485	0.446	0.566	0.676	0.529	0.386
Wan2.2_5B	0.444	0.517	0.601	0.666	0.425	0.317
Skyreels	0.394	0.311	0.673	0.740	0.500	0.253
LTX-Video	0.433	0.347	0.566	0.633	0.458	0.279
FramePack	0.301	0.145	0.655	0.732	0.560	0.168
HunyuanVideo	0.241	0.101	0.651	0.714	0.553	0.147
CogVideoX-5B	0.301	0.223	0.543	0.629	0.448	0.212
Closed-source						
Wan 2.6	0.640	0.545	0.701	0.743	0.634	0.514
Seedance 1.5 pro	0.638	0.710	0.677	0.763	0.625	0.569
Wan 2.5	0.603	0.519	0.743	0.737	0.615	0.495
Hailuo v2	0.600	0.677	0.725	0.706	0.637	0.544
Veo 3	0.608	0.681	0.709	0.729	0.641	0.530
Seedance 1.0	0.606	0.603	0.712	0.727	0.628	0.454
Kling 2.6 pro	0.618	0.685	0.710	0.750	0.697	0.530
Sora v2 Pro	0.422	0.296	0.646	0.715	0.603	0.255
Sora v1	0.250	0.133	0.629	0.689	0.543	0.166
Robotics-specific						
Cosmos 2.5	0.568	0.556	0.706	0.656	0.593	0.495
DreamGen(gr1)	0.444	0.284	0.750	0.731	0.611	0.333
DreamGen(droid)	0.491	0.505	0.655	0.637	0.517	0.316
Vidar	0.108	-0.023	0.550	0.491	0.375	0.054
UnifoLM-WMA-0	0.037	0.026	0.398	0.175	0.120	0.061
Mean	**0.453**	**0.413**	**0.643**	**0.677**	**0.544**	**0.349**

Table 10. **Results on Multi-Entity Collaboration with GPT**

Model	ACS	TCS	ECS	OCS	PSS	TS
Open-source						
Wan2.2_A14B	0.351	0.441	0.712	0.632	0.654	0.373
HunyuanVideo 1.5	0.301	0.375	0.687	0.625	0.585	0.311
LongCat-Video	0.190	0.244	0.696	0.642	0.654	0.220
Wan2.1_14B	0.261	0.318	0.693	0.545	0.551	0.282
LTX-2	0.243	0.289	0.651	0.572	0.578	0.233
Wan2.2_5B	0.125	0.181	0.687	0.568	0.537	0.141
Skyreels	0.189	0.195	0.695	0.664	0.621	0.203
LTX-Video	0.207	0.255	0.686	0.574	0.563	0.209
FramePack	0.186	0.186	0.709	0.598	0.529	0.173
HunyuanVideo	0.100	0.100	0.672	0.683	0.650	0.107
CogVideoX-5B	0.128	0.113	0.630	0.482	0.440	0.098
Closed-source						
Wan 2.6	0.443	0.541	0.738	0.708	0.654	0.478
Seedance 1.5 pro	0.456	0.591	0.743	0.689	0.689	0.483
Wan 2.5	0.392	0.453	0.750	0.633	0.662	0.401
Hailuo v2	0.378	0.422	0.744	0.678	0.619	0.385
Veo 3	0.450	0.432	0.731	0.737	0.621	0.430
Seedance 1.0	0.422	0.488	0.750	0.702	0.696	0.447
Kling 2.6 pro	0.357	0.392	0.738	0.698	0.676	0.364
Sora v2 Pro	0.155	0.191	0.733	0.558	0.591	0.186
Sora v1	0.128	0.107	0.595	0.494	0.523	0.111
Robotics-specific						
Cosmos 2.5	0.203	0.244	0.727	0.659	0.670	0.201
DreamGen(gr1)	0.262	0.347	0.707	0.664	0.628	0.296
DreamGen(droid)	0.211	0.260	0.657	0.548	0.548	0.214
Vidar	0.056	0.062	0.454	0.431	0.329	0.049
UnifoLM-WMA-0	0.017	0.034	0.392	0.301	0.295	0.018
Mean	**0.248**	**0.290**	**0.679**	**0.603**	**0.583**	**0.256**

Table 11. **Results on Spatial Relationship with GPT**

Model	SRS	MFS	OCS	RCS	PSS	TS
Open-source						
Wan2.2_A14B	0.604	0.596	0.709	0.758	0.669	0.454
HunyuanVideo 1.5	0.370	0.344	0.629	0.750	0.534	0.315
LongCat-Video	0.390	0.375	0.679	0.734	0.648	0.310
Wan2.1_14B	0.316	0.338	0.669	0.713	0.610	0.267
LTX-2	0.401	0.383	0.616	0.687	0.562	0.304
Wan2.2_5B	0.441	0.389	0.602	0.705	0.485	0.312
Skyreels	0.388	0.416	0.601	0.685	0.574	0.276
LTX-Video	0.224	0.215	0.612	0.681	0.491	0.176
FramePack	0.364	0.321	0.628	0.742	0.557	0.257
HunyuanVideo	0.163	0.192	0.605	0.740	0.586	0.179
CogVideoX-5B	0.193	0.241	0.451	0.653	0.370	0.111
Closed-source						
Wan 2.6	0.787	0.704	0.719	0.734	0.734	0.655
Seedance 1.5 pro	0.675	0.608	0.700	0.716	0.666	0.494
Wan 2.5	0.750	0.598	0.757	0.750	0.674	0.576
Hailuo v2	0.764	0.654	0.742	0.750	0.720	0.636
Veo 3	0.601	0.553	0.726	0.767	0.684	0.508
Seedance 1.0	0.484	0.445	0.710	0.742	0.679	0.425
Kling 2.6 pro	0.757	0.681	0.727	0.757	0.696	0.598
Sora v2 Pro	0.392	0.392	0.654	0.714	0.571	0.267
Sora v1	0.351	0.305	0.546	0.583	0.481	0.223
Robotics-specific						
Cosmos 2.5	0.419	0.395	0.661	0.693	0.620	0.338
DreamGen(gr1)	0.467	0.411	0.677	0.725	0.677	0.371
DreamGen(droid)	0.400	0.433	0.625	0.691	0.591	0.348
Vidar	0.163	0.250	0.517	0.500	0.405	0.105
UnifoLM-WMA-0	0.065	0.141	0.445	0.532	0.315	0.040
Mean	**0.437**	**0.415**	**0.640**	**0.700**	**0.584**	**0.342**

Table 12. **Results on Visual Reasoning with GPT**

Model	AES	VRS	OCS	RCS	PSS	TS
Open-source						
Wan2.2_A14B	0.424	0.401	0.651	0.709	0.552	0.330
HunyuanVideo 1.5	0.453	0.456	0.604	0.709	0.447	0.364
LongCat-Video	0.271	0.211	0.559	0.722	0.516	0.186
Wan2.1_14B	0.250	0.263	0.565	0.646	0.429	0.204
LTX-2	0.267	0.224	0.500	0.610	0.360	0.163
Wan2.2_5B	0.283	0.304	0.583	0.700	0.438	0.233
Skyreels	0.290	0.267	0.587	0.668	0.447	0.233
LTX-Video	0.283	0.287	0.644	0.688	0.516	0.241
FramePack	0.243	0.203	0.570	0.743	0.397	0.169
HunyuanVideo	0.096	0.058	0.647	0.744	0.522	0.035
CogVideoX-5B	0.136	0.120	0.428	0.577	0.261	0.079
Closed-source						
Wan 2.6	0.544	0.622	0.700	0.733	0.605	0.530
Seedance 1.5 pro	0.494	0.570	0.635	0.701	0.570	0.470
Wan 2.5	0.482	0.488	0.693	0.732	0.539	0.437
Hailuo v2	0.511	0.541	0.666	0.727	0.566	0.473
Veo 3	0.511	0.610	0.633	0.711	0.577	0.504
Seedance 1.0	0.505	0.505	0.705	0.733	0.644	0.441
Kling 2.6 pro	0.477	0.410	0.627	0.733	0.555	0.357
Sora v2 Pro	0.193	0.159	0.642	0.738	0.556	0.115
Sora v1	0.219	0.158	0.475	0.646	0.250	0.139
Robotics-specific						
Cosmos 2.5	0.482	0.493	0.664	0.744	0.590	0.399
DreamGen(gr1)	0.317	0.262	0.652	0.743	0.585	0.215
DreamGen(droid)	0.412	0.401	0.616	0.738	0.529	0.338
Vidar	0.090	0.050	0.556	0.659	0.431	0.050
UnifoLM-WMA-0	0.017	-0.056	0.062	0.306	0.176	0.000
Mean	**0.330**	**0.320**	**0.587**	**0.686**	**0.482**	**0.268**

Table 13. **Results on Dual Arm with GPT**

Model	PSS	TAC	RSS	MS	MA	TC	VQ	TS
Open-source								
Wan2.2_A14B	0.638	0.570	0.764	0.915	0.204	0.604	0.561	0.582
HunyuanVideo 1.5	0.612	0.622	0.649	0.951	0.370	0.618	0.434	0.526
LongCat-Video	0.620	0.540	0.771	0.937	0.244	0.580	0.572	0.576
Wan2.1_14B	0.600	0.540	0.650	0.850	0.261	0.570	0.424	0.497
LTX-2	0.488	0.415	0.637	0.848	0.378	0.451	0.396	0.423
Wan2.2_5B	0.575	0.498	0.606	0.940	0.269	0.536	0.360	0.448
Skyreels	0.598	0.498	0.658	0.884	0.252	0.548	0.406	0.477
LTX-Video	0.530	0.442	0.698	0.812	0.143	0.486	0.425	0.455
FramePack	0.550	0.395	0.712	0.885	0.103	0.472	0.457	0.464
HunyuanVideo	0.510	0.280	0.794	0.959	0.107	0.395	0.564	0.479
CogVideoX-5B	0.480	0.358	0.638	0.752	0.143	0.419	0.352	0.385
Closed-source								
Wan 2.6	0.655	0.708	0.819	0.984	0.333	0.681	0.680	0.680
Seedance 1.5 pro	0.668	0.800	0.721	0.960	0.399	0.734	0.547	0.640
Wan 2.5	0.670	0.740	0.758	0.970	0.347	0.705	0.563	0.633
Hailuo v2	0.658	0.720	0.751	0.983	0.312	0.689	0.534	0.611
Veo 3	0.665	0.682	0.711	0.973	0.262	0.674	0.546	0.610
Seedance 1.0	0.668	0.648	0.801	0.972	0.294	0.658	0.623	0.640
Kling 2.6 pro	0.672	0.615	0.770	0.965	0.210	0.644	0.567	0.605
Sora v2 Pro	0.565	0.364	0.776	0.950	0.272	0.465	0.560	0.512
Sora v1	0.422	0.312	0.531	0.880	0.210	0.368	0.279	0.323
Robotics-specific								
Cosmos 2.5	0.665	0.575	0.746	0.930	0.127	0.620	0.500	0.560
DreamGen(gr1)	0.630	0.492	0.779	0.939	0.123	0.561	0.503	0.532
DreamGen(droid)	0.610	0.542	0.668	0.863	0.201	0.576	0.375	0.475
Vidar	0.450	0.238	0.781	0.933	0.025	0.344	0.475	0.409
UnifoLM-WMA-0	0.348	0.252	0.348	0.497	0.120	0.300	0.089	0.194
Mean	**0.581**	**0.513**	**0.701**	**0.901**	**0.228**	**0.547**	**0.471**	**0.509**

Table 14. **Results on Humanoid with GPT**

Model	PSS	TAC	RSS	MS	MA	TC	VQ	TS
Open-source								
Wan2.2_A14B	0.678	0.748	0.787	0.966	0.105	0.712	0.584	0.647
HunyuanVideo 1.5	0.660	0.800	0.703	0.922	0.206	0.730	0.460	0.595
LongCat-Video	0.652	0.668	0.801	0.969	0.089	0.660	0.583	0.621
Wan2.1_14B	0.635	0.660	0.765	0.938	0.167	0.648	0.550	0.599
LTX-2	0.622	0.608	0.736	0.880	0.231	0.615	0.495	0.554
Wan2.2_5B	0.668	0.695	0.755	0.963	0.152	0.681	0.533	0.607
Skyreels	0.610	0.598	0.725	0.808	0.086	0.604	0.414	0.509
LTX-Video	0.602	0.630	0.671	0.628	0.050	0.616	0.312	0.463
FramePack	0.598	0.572	0.776	0.864	0.069	0.585	0.511	0.548
HunyuanVideo	0.595	0.465	0.763	0.933	0.096	0.530	0.517	0.523
CogVideoX-5B	0.570	0.565	0.698	0.647	0.127	0.568	0.424	0.496
Closed-source								
Wan 2.6	0.665	0.818	0.794	0.980	0.132	0.741	0.593	0.667
Seedance 1.5 pro	0.682	0.838	0.793	0.955	0.158	0.760	0.623	0.691
Wan 2.5	0.698	0.782	0.779	0.981	0.122	0.740	0.568	0.653
Hailuo v2	0.682	0.815	0.741	0.970	0.133	0.749	0.521	0.635
Veo 3	0.655	0.785	0.776	0.968	0.132	0.720	0.554	0.637
Seedance 1.0	0.675	0.752	0.833	0.964	0.166	0.714	0.658	0.686
Kling 2.6 pro	0.672	0.762	0.750	0.959	0.113	0.718	0.508	0.613
Sora v2 Pro	0.638	0.565	0.774	0.936	0.085	0.602	0.520	0.561
Sora v1	0.542	0.450	0.597	0.900	0.216	0.496	0.342	0.419
Robotics-specific								
Cosmos 2.5	0.650	0.720	0.797	0.925	0.071	0.685	0.566	0.625
DreamGen(gr1)	0.652	0.595	0.781	0.885	0.079	0.624	0.526	0.575
DreamGen(droid)	0.620	0.700	0.704	0.843	0.137	0.660	0.453	0.556
Vidar	0.445	0.298	0.694	0.855	0.025	0.371	0.343	0.357
UnifoLM-WMA-0	0.270	0.282	0.386	0.512	0.112	0.276	0.125	0.200
Mean	**0.617**	**0.646**	**0.735**	**0.886**	**0.122**	**0.632**	**0.491**	**0.561**

Table 15. **Results on Single Arm with GPT**

Model	PSS	TAC	RSS	MS	MA	TC	VQ	TS
Open-source								
Wan2.2_A14B	0.638	0.582	0.783	0.942	0.263	0.610	0.607	0.608
HunyuanVideo 1.5	0.510	0.622	0.651	0.949	0.355	0.566	0.460	0.513
LongCat-Video	0.562	0.530	0.807	0.945	0.298	0.546	0.625	0.585
Wan2.1_14B	0.542	0.472	0.677	0.849	0.282	0.507	0.422	0.464
LTX-2	0.490	0.410	0.681	0.919	0.464	0.450	0.456	0.453
Wan2.2_5B	0.518	0.480	0.619	0.943	0.313	0.499	0.372	0.435
Skyreels	0.525	0.495	0.712	0.911	0.286	0.510	0.504	0.507
LTX-Video	0.492	0.408	0.686	0.804	0.158	0.450	0.431	0.440
FramePack	0.445	0.318	0.760	0.888	0.104	0.381	0.498	0.439
HunyuanVideo	0.445	0.265	0.809	0.963	0.118	0.355	0.552	0.453
CogVideoX-5B	0.405	0.335	0.582	0.815	0.256	0.370	0.307	0.338
Closed-source								
Wan 2.6	0.652	0.710	0.796	0.983	0.392	0.681	0.651	0.666
Seedance 1.5 pro	0.635	0.832	0.752	0.960	0.419	0.734	0.561	0.647
Wan 2.5	0.668	0.802	0.787	0.969	0.412	0.735	0.624	0.679
Hailuo v2	0.665	0.752	0.680	0.989	0.396	0.709	0.479	0.594
Veo 3	0.642	0.755	0.750	0.977	0.362	0.699	0.568	0.633
Seedance 1.0	0.658	0.655	0.769	0.967	0.306	0.656	0.589	0.622
Kling 2.6 pro	0.622	0.640	0.714	0.958	0.305	0.631	0.508	0.569
Sora v2 Pro	0.490	0.310	0.784	0.962	0.324	0.400	0.552	0.476
Sora v1	0.350	0.320	0.532	0.884	0.247	0.335	0.293	0.314
Robotics-specific								
Cosmos 2.5	0.632	0.592	0.737	0.888	0.206	0.612	0.475	0.543
DreamGen(gr1)	0.618	0.620	0.716	0.932	0.263	0.619	0.508	0.563
DreamGen(droid)	0.568	0.568	0.691	0.930	0.214	0.568	0.430	0.499
Vidar	0.415	0.272	0.726	0.929	0.068	0.344	0.420	0.382
UnifoLM-WMA-0	0.428	0.315	0.437	0.709	0.389	0.371	0.164	0.267
Mean	**0.544**	**0.522**	**0.705**	**0.918**	**0.288**	**0.533**	**0.482**	**0.507**

Table 16. **Results on Quadruped Robot with GPT**

Model	PSS	TAC	RSS	MS	MA	TC	VQ	TS
Open-source								
Wan2.2_A14B	0.760	0.712	0.800	0.888	0.196	0.736	0.643	0.689
HunyuanVideo 1.5	0.715	0.738	0.713	0.940	0.403	0.726	0.542	0.634
LongCat-Video	0.742	0.628	0.827	0.923	0.179	0.685	0.676	0.680
Wan2.1_14B	0.722	0.670	0.706	0.850	0.303	0.696	0.495	0.595
LTX-2	0.715	0.670	0.758	0.845	0.287	0.692	0.552	0.622
Wan2.2_5B	0.712	0.678	0.688	0.921	0.273	0.695	0.486	0.590
Skyreels	0.732	0.605	0.722	0.853	0.163	0.669	0.503	0.586
LTX-Video	0.698	0.678	0.668	0.676	0.122	0.688	0.364	0.526
FramePack	0.720	0.575	0.827	0.876	0.069	0.648	0.605	0.626
HunyuanVideo	0.730	0.602	0.788	0.953	0.127	0.666	0.584	0.625
CogVideoX-5B	0.655	0.560	0.624	0.618	0.220	0.608	0.322	0.464
Closed-source								
Wan 2.6	0.755	0.792	0.813	0.970	0.316	0.774	0.672	0.723
Seedance 1.5 pro	0.748	0.820	0.746	0.884	0.407	0.784	0.577	0.680
Wan 2.5	0.785	0.792	0.809	0.948	0.322	0.789	0.664	0.726
Hailuo v2	0.748	0.738	0.711	0.961	0.354	0.742	0.538	0.640
Veo 3	0.745	0.695	0.798	0.961	0.214	0.720	0.658	0.689
Seedance 1.0	0.768	0.735	0.791	0.945	0.334	0.751	0.645	0.698
Kling 2.6 pro	0.740	0.738	0.736	0.861	0.258	0.739	0.535	0.637
Sora v2 Pro	0.731	0.670	0.789	0.922	0.239	0.701	0.626	0.663
Sora v1	0.700	0.672	0.620	0.863	0.282	0.686	0.401	0.543
Robotics-specific								
Cosmos 2.5	0.752	0.622	0.808	0.892	0.137	0.688	0.629	0.658
DreamGen(gr1)	0.712	0.655	0.706	0.788	0.164	0.684	0.474	0.579
DreamGen(droid)	0.705	0.568	0.687	0.854	0.160	0.636	0.448	0.542
Vidar	0.528	0.472	0.552	0.749	0.074	0.500	0.247	0.373
UnifoLM-WMA-0	0.475	0.390	0.410	0.497	0.132	0.432	0.154	0.293
Mean	**0.711**	**0.659**	**0.723**	**0.857**	**0.229**	**0.685**	**0.521**	**0.603**

Table 17. **Results on Common Manipulation with Qwen**

Model	AES	TCS	OCS	RCS	PSS	TS
Open-source						
Wan2.2_A14B	0.804	0.815	0.934	0.913	0.923	0.708
LongCat-Video	0.681	0.704	0.840	0.920	0.897	0.677
Wan2.2_5B	0.620	0.630	0.940	0.890	0.860	0.597
Wan2.1_14B	0.663	0.653	0.903	0.932	0.807	0.687
Skyreels	0.547	0.559	0.928	0.916	0.952	0.546
LTX-Video	0.630	0.666	0.845	0.904	0.880	0.450
FramePack	0.336	0.347	0.913	0.956	0.934	0.455
CogVideoX-5B	0.352	0.340	0.625	0.704	0.647	0.289
Closed-source						
Wan 2.5	0.928	0.946	0.946	0.964	0.973	0.887
Hailuo v2	0.916	0.952	0.940	0.940	0.976	0.843
Veo 3	0.953	0.962	0.962	0.981	0.990	0.896
Seedance 1.0	0.946	0.928	0.955	0.946	0.964	0.856
Robotics-specific						
Cosmos 2.5	0.687	0.708	0.895	0.916	0.864	0.687
DreamGen(gr1)	0.625	0.656	0.906	0.937	0.906	0.507
DreamGen(droid)	0.630	0.690	0.785	0.845	0.773	0.465
UnifoLM-WMA-0	0.043	0.043	0.532	0.478	0.369	0.028
Vidar	0.142	0.166	0.833	0.833	0.761	0.117
Mean	**0.597**	**0.611**	**0.867**	**0.858**	**0.835**	**0.559**

Table 18. **Results on Long-Horizon Planning with Qwen**

Model	AES	ECS	OCS	RCS	PSS	TS
Open-source						
Wan2.2_A14B	0.836	0.883	0.942	0.971	0.951	0.680
LongCat-Video	0.650	0.706	0.925	0.950	0.912	0.489
Wan2.2_5B	0.520	0.594	0.947	0.937	0.895	0.449
Wan2.1_14B	0.681	0.669	0.931	0.931	0.875	0.465
Skyreels	0.416	0.511	0.944	0.958	0.861	0.324
LTX-Video	0.352	0.414	0.779	0.808	0.705	0.357
FramePack	0.212	0.252	0.862	0.987	0.937	0.196
CogVideoX-5B	0.276	0.292	0.855	0.763	0.684	0.096
Closed-source						
Wan 2.5	0.714	0.836	0.964	0.964	0.955	0.719
Hailuo v2	0.808	0.903	0.950	0.991	0.908	0.705
Veo 3	0.812	0.903	0.984	0.968	0.945	0.854
Seedance 1.0	0.824	0.889	0.990	0.990	0.925	0.715
Robotics-specific						
Cosmos 2.5	0.731	0.810	0.953	0.981	0.935	0.596
DreamGen(gr1)	0.475	0.587	0.950	0.937	0.887	0.353
DreamGen(droid)	0.602	0.717	0.882	0.882	0.852	0.301
UnifoLM-WMA-0	0.000	-0.041	0.645	0.500	0.500	0.000
Vidar	0.050	0.058	0.800	0.866	0.666	0.019
Mean	**0.505**	**0.565**	**0.893**	**0.886**	**0.833**	**0.417**

Table 19. **Results on Multi-Entity Collaboration with Qwen**

Model	ACS	TCS	ECS	OCS	PSS	TS
Open-source						
Wan2.2_A14B	0.941	0.941	0.985	0.992	1.000	0.920
LongCat-Video	0.887	0.912	0.968	0.962	0.968	0.814
Wan2.2_5B	0.763	0.796	0.993	0.960	0.967	0.721
Wan2.1_14B	0.750	0.786	0.975	0.945	0.957	0.702
Skyreels	0.743	0.736	1.000	1.000	0.993	0.686
LTX-Video	0.806	0.806	0.975	0.975	0.993	0.734
FramePack	0.710	0.703	0.980	0.993	0.947	0.630
CogVideoX-5B	0.522	0.536	0.933	0.882	0.889	0.426
Closed-source						
Wan 2.5	0.896	0.908	0.987	0.975	1.000	0.915
Hailuo v2	0.986	0.993	0.952	0.972	0.986	0.892
Veo 3	0.914	0.914	0.993	0.987	1.000	0.924
Seedance 1.0	0.960	0.967	0.973	0.960	1.000	0.898
Robotics-specific						
Cosmos 2.5	0.864	0.878	0.993	1.000	0.972	0.768
DreamGen(gr1)	0.878	0.878	0.957	0.963	0.969	0.848
DreamGen(droid)	0.763	0.819	0.958	0.951	0.930	0.591
UnifoLM-WMA-0	0.044	0.044	0.507	0.477	0.301	0.025
Vidar	0.083	0.097	0.743	0.736	0.583	0.082
Mean	**0.714**	**0.724**	**0.933**	**0.928**	**0.910**	**0.657**

Table 20. **Results on Spatial Relationship with Qwen**

Model	SRS	MFS	OCS	RCS	PSS	TS
Open-source						
Wan2.2_A14B	0.833	0.833	1.000	1.000	1.000	0.660
LongCat-Video	0.636	0.636	1.000	1.000	1.000	0.465
Wan2.2_5B	0.600	0.625	1.000	0.900	0.900	0.402
Wan2.1_14B	0.636	0.636	1.000	1.000	1.000	0.400
Skyreels	0.750	0.750	1.000	0.875	0.875	0.400
LTX-Video	0.750	0.750	1.000	1.000	1.000	0.382
FramePack	0.400	0.400	1.000	1.000	1.000	0.240
CogVideoX-5B	0.500	0.535	1.000	0.857	0.714	0.240
Closed-source						
Wan 2.5	0.916	0.916	1.000	0.979	1.000	0.825
Hailuo v2	1.000	1.000	1.000	1.000	1.000	0.840
Veo 3	0.933	0.933	1.000	1.000	1.000	0.740
Seedance 1.0	0.666	0.666	1.000	1.000	1.000	0.665
Robotics-specific						
Cosmos 2.5	0.875	0.875	1.000	1.000	1.000	0.512
DreamGen(gr1)	0.642	0.642	1.000	1.000	1.000	0.500
DreamGen(droid)	0.545	0.545	1.000	1.000	1.000	0.505
UnifoLM-WMA-0	0.200	0.250	0.625	0.575	0.500	0.065
Vidar	0.367	0.382	0.691	0.647	0.573	0.140
Mean	**0.666**	**0.674**	**0.962**	**0.935**	**0.920**	**0.464**

Table 21. **Results on Visual Reasoning with Qwen**

Model	AES	VRS	OCS	RCS	PSS	TS
Open-source						
Wan2.2_A14B	0.727	0.701	1.000	1.000	1.000	0.550
LongCat-Video	0.361	0.358	0.944	0.958	0.847	0.354
Wan2.2_5B	0.291	0.218	0.906	0.906	0.906	0.420
Wan2.1_14B	0.397	0.401	0.886	0.897	0.784	0.269
Skyreels	0.420	0.401	0.960	0.930	0.830	0.357
LTX-Video	0.390	0.410	0.859	0.937	0.859	0.285
FramePack	0.350	0.343	0.980	0.990	0.950	0.345
CogVideoX-5B	0.073	0.044	0.779	0.632	0.544	0.030
Closed-source						
Wan 2.5	0.809	0.770	0.988	1.000	0.952	0.737
Hailuo v2	0.790	0.882	1.000	1.000	0.900	0.820
Veo 3	0.847	0.853	0.945	1.000	0.989	0.750
Seedance 1.0	0.927	0.945	0.979	0.989	1.000	0.789
Robotics-specific						
Cosmos 2.5	0.593	0.632	0.984	1.000	1.000	0.506
DreamGen(gr1)	0.437	0.425	0.937	0.958	0.822	0.404
DreamGen(droid)	0.600	0.575	0.800	0.825	0.787	0.386
UnifoLM-WMA-0	0.000	-0.098	0.177	0.348	0.289	0.000
Vidar	0.000	-0.062	0.947	0.937	0.906	0.029
Mean	**0.452**	**0.438**	**0.890**	**0.882**	**0.825**	**0.395**

Table 22. **Results on Dual Arm Robot with Qwen**

Model	PSS	TAC	RSS	MS	MA	TC	VQ	TS
Open-source								
Wan2.2_A14B	0.852	0.760	0.767	0.915	0.204	0.806	0.550	0.678
LongCat-Video	0.738	0.638	0.741	0.937	0.244	0.688	0.517	0.602
Wan2.2_5B	0.688	0.645	0.658	0.940	0.269	0.666	0.402	0.534
Wan2.1_14B	0.790	0.730	0.639	0.850	0.261	0.760	0.364	0.562
Skyreels	0.755	0.700	0.675	0.884	0.252	0.728	0.419	0.573
LTX-Video	0.615	0.535	0.692	0.812	0.143	0.575	0.399	0.487
FramePack	0.612	0.495	0.719	0.885	0.103	0.554	0.445	0.499
CogVideoX-5B	0.505	0.540	0.613	0.752	0.143	0.522	0.323	0.422
Closed-source								
Wan 2.5	0.920	0.880	0.761	0.970	0.347	0.900	0.588	0.744
Hailuo v2	0.908	0.848	0.744	0.983	0.312	0.878	0.534	0.706
Veo 3	0.870	0.802	0.777	0.973	0.262	0.836	0.581	0.708
Seedance 1.0	0.895	0.810	0.801	0.972	0.294	0.852	0.608	0.730
Robotics-specific								
Cosmos 2.5	0.792	0.708	0.791	0.930	0.127	0.750	0.543	0.646
DreamGen(gr1)	0.780	0.638	0.801	0.939	0.123	0.709	0.555	0.632
DreamGen(droid)	0.722	0.678	0.711	0.863	0.201	0.700	0.441	0.570
UnifoLM-WMA-0	0.110	0.222	0.266	0.497	0.120	0.166	0.046	0.106
Vidar	0.295	0.240	0.804	0.933	0.025	0.268	0.511	0.389
Mean	**0.685**	**0.622**	**0.694**	**0.880**	**0.205**	**0.651**	**0.449**	**0.550**

Table 23. **Results on Humanoid Robot with Qwen**

Model	PSS	TAC	RSS	MS	MA	TC	VQ	TS
Open-source								
Wan2.2_A14B	0.935	0.800	0.806	0.966	0.105	0.898	0.557	0.727
LongCat-Video	0.918	0.765	0.826	0.969	0.089	0.841	0.579	0.710
Wan2.2_5B	0.880	0.758	0.791	0.963	0.152	0.819	0.544	0.681
Wan2.1_14B	0.895	0.712	0.785	0.938	0.167	0.804	0.524	0.664
Skyreels	0.842	0.670	0.803	0.808	0.086	0.756	0.500	0.628
LTX-Video	0.852	0.670	0.808	0.628	0.050	0.761	0.445	0.603
FramePack	0.815	0.622	0.838	0.864	0.069	0.719	0.550	0.634
CogVideoX-5B	0.712	0.602	0.710	0.647	0.127	0.658	0.390	0.524
Closed-source								
Wan 2.5	0.935	0.835	0.826	0.981	0.122	0.885	0.600	0.742
Hailuo v2	0.952	0.848	0.796	0.970	0.133	0.891	0.548	0.719
Veo 3	0.955	0.802	0.831	0.968	0.132	0.829	0.604	0.716
Seedance 1.0	0.936	0.814	0.829	0.964	0.166	0.904	0.614	0.759
Robotics-specific								
Cosmos 2.5	0.868	0.730	0.841	0.925	0.071	0.799	0.578	0.688
DreamGen(gr1)	0.858	0.682	0.823	0.885	0.079	0.770	0.542	0.656
DreamGen(droid)	0.842	0.736	0.739	0.843	0.137	0.788	0.478	0.633
UnifoLM-WMA-0	0.168	0.285	0.349	0.512	0.112	0.226	0.115	0.170
Vidar	0.440	0.305	0.700	0.855	0.025	0.372	0.357	0.364
Mean	**0.806**	**0.674**	**0.770**	**0.862**	**0.113**	**0.740**	**0.501**	**0.620**

Table 24. **Results on Single Arm Robot with Qwen**

Model	PSS	TAC	RSS	MS	MA	TC	VQ	TS
Open-source								
Wan2.2_A14B	0.815	0.802	0.755	0.942	0.263	0.809	0.568	0.688
LongCat-Video	0.750	0.712	0.812	0.945	0.298	0.736	0.660	0.698
Wan2.2_5B	0.605	0.650	0.626	0.943	0.313	0.628	0.394	0.511
Wan2.1_14B	0.700	0.616	0.638	0.849	0.282	0.659	0.379	0.519
Skyreels	0.760	0.699	0.731	0.911	0.286	0.732	0.492	0.612
LTX-Video	0.535	0.525	0.710	0.804	0.158	0.530	0.443	0.486
FramePack	0.400	0.500	0.674	0.888	0.104	0.394	0.413	0.403
CogVideoX-5B	0.460	0.460	0.553	0.815	0.256	0.460	0.288	0.374
Closed-source								
Wan 2.5	0.895	0.918	0.747	0.969	0.412	0.895	0.589	0.742
Hailuo v2	0.910	0.906	0.705	0.989	0.396	0.902	0.489	0.695
Veo 3	0.898	0.891	0.756	0.977	0.362	0.890	0.594	0.742
Seedance 1.0	0.852	0.818	0.747	0.967	0.306	0.835	0.580	0.707
Robotics-specific								
Cosmos 2.5	0.795	0.765	0.751	0.888	0.206	0.776	0.516	0.646
DreamGen(gr1)	0.812	0.728	0.782	0.932	0.263	0.770	0.550	0.660
DreamGen(droid)	0.745	0.675	0.721	0.930	0.214	0.710	0.468	0.589
UnifoLM-WMA-0	0.382	0.495	0.391	0.709	0.389	0.439	0.140	0.289
Vidar	0.220	0.200	0.736	0.929	0.068	0.210	0.479	0.344
Mean	**0.658**	**0.648**	**0.679**	**0.898**	**0.263**	**0.649**	**0.454**	**0.551**

Table 25. **Results on Quadruped Robot with Qwen**

Model	PSS	TAC	RSS	MS	MA	TC	VQ	TS
Open-source								
Wan2.2_A14B	0.860	0.698	0.746	0.888	0.196	0.779	0.561	0.670
LongCat-Video	0.870	0.685	0.747	0.923	0.179	0.778	0.554	0.666
Wan2.2_5B	0.858	0.635	0.721	0.921	0.273	0.746	0.529	0.637
Wan2.1_14B	0.845	0.665	0.674	0.850	0.303	0.755	0.453	0.604
Skyreels	0.862	0.652	0.749	0.853	0.163	0.758	0.550	0.654
LTX-Video	0.818	0.690	0.709	0.676	0.122	0.754	0.423	0.588
FramePack	0.765	0.540	0.881	0.876	0.069	0.652	0.687	0.669
CogVideoX-5B	0.705	0.575	0.639	0.618	0.220	0.640	0.349	0.494
Closed-source								
Wan 2.5	0.902	0.740	0.769	0.948	0.322	0.821	0.601	0.711
Hailuo v2	0.888	0.722	0.654	0.961	0.354	0.805	0.458	0.631
Veo 3	0.880	0.722	0.793	0.961	0.214	0.801	0.652	0.726
Seedance 1.0	0.865	0.710	0.728	0.945	0.334	0.788	0.560	0.674
Robotics-specific								
Cosmos 2.5	0.850	0.612	0.755	0.892	0.137	0.731	0.547	0.639
DreamGen(gr1)	0.852	0.672	0.701	0.788	0.164	0.762	0.460	0.611
DreamGen(droid)	0.745	0.592	0.724	0.854	0.160	0.669	0.500	0.584
UnifoLM-WMA-0	0.310	0.300	0.399	0.497	0.132	0.305	0.197	0.251
Vidar	0.495	0.445	0.575	0.749	0.074	0.470	0.290	0.380
Mean	**0.782**	**0.612**	**0.705**	**0.825**	**0.198**	**0.701**	**0.494**	**0.597**