MECHVERSE: EVALUATING PHYSICAL MOTION CONSISTENCY IN VIDEO GENERATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text- and image-conditioned video generation models have achieved strong visual fidelity and temporal coherence, but they often fail to generate motion governed by kinematic and geometric constraints. In these settings, object parts must remain rigid, maintain contact or coupling with neighboring components, and transfer motion consistently across connected parts. These requirements are especially explicit in articulated mechanical assemblies, where motion is constrained by rigid-link geometry, contact/coupling relations, and transmission through kinematic chains. For example, gears rotate in coordination, cam–follower systems convert rotation into translation, and linkages propagate motion through connected rigid parts. A generated video may therefore appear plausible while violating the intended mechanism, such as rotating a part that should translate, deforming a rigid component, breaking coupling between parts, or failing to move downstream components. To evaluate this gap, We introduce MechVerse, a benchmark for mechanically consistent image-to-video generation. MechVerse contains 21,156 synthetic clips from 1,357 mechanical assemblies across 141 categories, organized into three tiers of increasing kinematic complexity: independent articulation, pairwise coupling, and densely coupled multi-part mechanisms. Each clip is paired with a structured prompt describing part identities, stationary supports, moving components, motion primitives, direction, speed/extent, and inter-part dependencies. We evaluate proprietary, open-source, and fine-tuned image-to-video models using standard video metrics, instruction-following scores, and human judgments of motion correctness and kinematic coupling. Results show that current models can preserve appearance and smoothness while failing to generate mechanically admissible motion, with errors increasing as coupling complexity grows. MechVerse provides a benchmark for measuring and improving mechanism-aware video generation from image and language inputs. The project page is: https://mechverse.pages.dev/

1 INTRODUCTION

Recent image-to-video generation models can animate a single image from a text prompt with high visual fidelity and temporal smoothness (Blattmann et al., 2023; Yang et al., 2024b; Kong et al., 2024; Wan et al., 2025; Ma et al., 2025). This progress has expanded the range of scenes and object motions that can be synthesized from simple image-language inputs. However, existing benchmarks mostly focus on natural scenes, human actions, broad object motion, or general visual quality (Liu et al., 2024b; 2023; Huang et al., 2024). These benchmarks are useful for measuring perceptual quality, temporal coherence, and text-video alignment. But they do not test whether generated motion follows explicit part-level mechanical constraints.

Mechanical assemblies expose this limitation. Their motion is governed by structured dependencies between parts. When one component moves, other components often need to move through joints, contact, or linkages. If these dependencies are violated, the video may still look smooth but become mechanically incorrect. This raises a key question: can current image-to-video models preserve part-level geometric and kinematic constraints? We study this question and find that existing video generation benchmarks have three limitations:

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

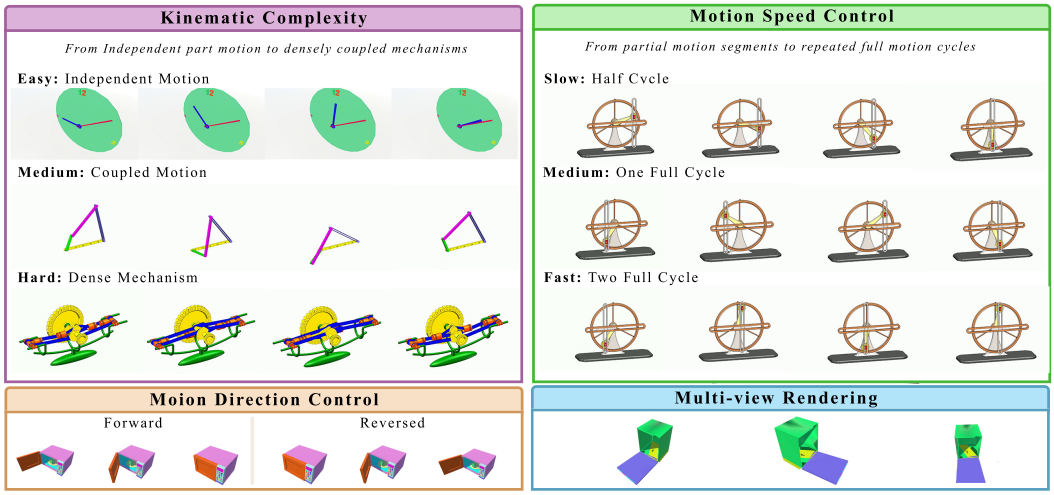


Figure 1: Overview of MechVerse and its structured motion variation design. **Top-left:** MechBench spans three levels of kinematic complexity, ranging from independent articulated motion (*Easy*) to coupled mechanisms (*Medium*) and densely interacting multi-part systems (*Hard*). **Top-right:** Motion speed is systematically varied across partial, single-cycle, and repeated-cycle motion trajectories (*Slow*, *Medium*, *Fast*). **Bottom-left:** Each motion sequence is paired with forward and reversed temporal directions, enabling semantically distinct motions such as opening/closing or clockwise/counter-clockwise rotation. **Bottom-right:** Assemblies are rendered from multiple camera viewpoints to provide viewpoint diversity while preserving identical underlying mechanical motion.

- **Limited coverage of articulated mechanisms.** Existing benchmarks cover diverse scenes, actions, and object motions (Liu et al., 2024b; 2023; Huang et al., 2024). However, they do not explicitly study mechanical assemblies, where motion is governed by coupled part-level constraints and rich geometric structure.
- **Insufficient control over motion dependencies.** Many benchmarks evaluate general motion quality or prompt alignment, but do not control whether motion is independent, pairwise coupled, or propagated through densely connected multi-part mechanisms. This makes it difficult to analyze how model performance changes with kinematic complexity.
- **Lack of structured and fine-grained motion specifications.** Existing prompts often describe motion at a semantic level, such as an object moving, rotating, or opening (Unmesh et al., 2023; Ji et al., 2020). They usually do not specify stationary supports, moving components, motion primitives, direction, speed or extent, and inter-part dependencies. Without such structure, it is difficult to test whether a generated video follows the intended mechanism.

Image-to-video generation is being applied to procedural and instructional content (Miech et al., 2019; Ben-Shabat et al., 2021), world models and synthetic interaction data for embodied agents (Yang et al., 2024a; Xiang et al., 2020), and CAD-based design and digital-twin visualization, where the usefulness of a generated clip depends on whether parts move under the correct kinematic constraints rather than only on appearance. Mechanically constrained motion is also a probe for the broader claim that video generation models implicitly learn world dynamics (OpenAI, 2024; Ha & Schmidhuber, 2018), since existing physics-oriented evaluations focus on rigid-body, fluid, and collision behavior (Bansal et al., 2024; Meng et al., 2024b; Motamed et al., 2025) and rarely consider articulated mechanisms whose constraints are deterministic and propagate globally across coupled parts. Varying dependency complexity from independent parts to densely coupled mechanisms further turns kinematic structure into a controlled axis of compositional generalization, of the kind where diagnostic benchmarks (Johnson et al., 2017; Bear et al., 2021) have historically driven progress that aggregate video-quality scores (Huang et al., 2024; Liu et al., 2024b) cannot expose. Closing this gap therefore matters application reliability, for scientific evaluation of generative world models, and for benchmark methodology.

To study this problem, we introduce **MechVerse**, a synthetic video benchmark for mechanically consistent image-to-video generation. Large-scale web video corpora provide broad appearance and motion priors but offer little control over explicit part-level dependencies. MechVerse provides controlled examples where moving parts, stationary supports, motion primitives, and inter-part dependencies are known. It is constructed from two complementary sources: the PartNet-Mobility dataset (Xiang et al., 2020), which provides kinematically annotated articulated objects spanning 46 categories, and a curated library of CAD mechanical assemblies covering linkages, cam-and-follower systems, gear trains, and complex multi-part mechanisms. All assemblies are rendered in Unity using a frame-accurate animation pipeline with normalized time stepping, enabling precise and reproducible motion control. The resulting dataset contains over 21,156 videos from 1,357 assemblies, organized into three tiers of increasing kinematic complexity: *Easy* (single- or dual-part articulation), *Medium* (two-part coupled mechanisms with 3–8 parts), and *Hard* (strongly coupled multi-part assemblies with 10–50 parts), as shown in Figure 1. Each clip is varied along three axes: motion speed (slow, mid, fast), camera viewpoint (canonical and mirrored), and motion direction (forward and reversed). Each clip is paired with an input image and a structured text prompt describing part identities, stationary and moving components, motion type, direction, speed or extent, and inter-part dependencies. This setup allows us to evaluate whether generated videos preserve both individual part motion and the dependencies between interacting parts.

Using MechVerse, we conduct the first large-scale benchmark of state-of-the-art video generation models on mechanically consistent video synthesis. We evaluate 14 models, including open-source models such as DynamiCrafter (Xing et al., 2024), ConsistI2V Ren et al. (2024), and CogVideoX (Yang et al., 2024b), as well as closed-source systems such as KlingV3, Wan 2.7, and Happy Horse. We use standard video metrics, instruction-following scores, and human judgments of motion correctness and coupling Huang et al. (2024); Li et al. (2025). Our analysis reveals three key findings: (i) perceptual video quality is a weak proxy for mechanical correctness; (ii) models often confuse motion primitives, fail to move driven components, or break coupling between interacting parts; and (iii) errors increase with kinematic complexity, especially for densely coupled multi-part mechanisms. Overall, MechVerse shows that current models can produce smooth and visually plausible videos but do not yet reliably capture structured inter-part motion from image and language inputs.

2 Related Work

Video Generation. Early text-to-video models extended image diffusion models with temporal modules (Blattmann et al., 2023; Guo et al., 2023; Chen et al., 2024; Wang et al., 2023), while Diffusion Transformer-based systems (Peebles & Xie, 2023) enabled higher-fidelity generation in models such as CogVideoX (Yang et al., 2024b), HunyuanVideo (Kong et al., 2024), Wan2.1 (Wan et al., 2025), Step-Video-T2V (Ma et al., 2025), and Sora (OpenAI, 2025). Physical consistency in generation has been studied through evaluation benchmarks (Bansal et al., 2024; Meng et al., 2024a), training-time annotation (Wang et al., 2025), and prompt refinement (Xue et al., 2025), while animation methods (Le Cleac'h et al., 2023; Montanaro et al., 2024; Liu et al., 2024a) animate scenes under simplified physical assumptions. These works address global physical realism but not structured kinematic coupling between mechanically interacting parts, which is focus of MechVerse.

Articulated Object Datasets. PartNet-Mobility (Xiang et al., 2020) provides kinematic joint annotations for over 2,000 objects and is the most widely used articulated motion resource. Shape2Motion (Wang et al., 2019), RPM-Net (Yan et al., 2020), ACD (Iliash et al., 2024), and GAPartNet (Geng et al., 2023) extend coverage to joint discovery, motion fields, and functional part semantics. Patel et al. (2025) introduced a benchmark of 693 synthetic CAD gear assemblies for coupled motion prediction from point clouds. In the video domain, HA-ViD (Zheng et al., 2023) and IKEA Video Manuals (Liu et al., 2024c) provide assembly recordings with temporal annotations. All of these resources target 3D motion estimation or action recognition rather than 2D generative video evaluation, and none provides the structured prompt-paired clips with controlled kinematic complexity that MechVerse offers.

3 MECHVERSE DATASET

MechVerse comprises 21,156 video clips derived from 1,357 unique mechanical assemblies spanning 141 categories, constructed from two complementary sources: 904 assemblies from PartNet-Mobility (Xiang et al., 2020), covering everyday articulated objects with independent part motion assigned to the *Easy* tier, and 453 CAD mechanical assemblies curated specifically for this dataset covering linkages, cam-and-follower systems, engine pistons, and complex multi-part mechanisms, assigned to the *Medium* and *Hard* tiers. The following subsections describe the complexity structure, clip variation axes, prompt design, annotation pipeline, and train/test split in detail.

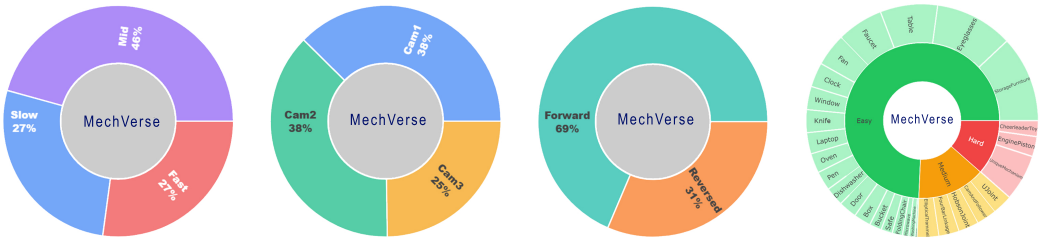


Figure 2: MechVerse dataset statistics. *Top row:* Distribution of clips by speed variant (Slow 27%, Mid 46%, Fast 27%), camera viewpoint (Cam1 25%, Cam2 38%, Cam3 37%), motion direction (Forward 69%, Reversed 31%), and category hierarchy by complexity tier.

3.1 COMPLEXITY TIERS

A single aggregate evaluation score cannot reveal *where* a generative model fails, whether it struggles with any motion at all, or only when motion must propagate across interacting parts. To enable this diagnostic analysis, MechVerse organizes assemblies into three tiers of kinematic complexity, reflecting a fundamental distinction between independent and coupled part motion. *Easy* assemblies (15,720 clips) contain one or two parts where motion is independent: a door swinging, a clock hand rotating, or a drawer sliding, consistent with the motion structure of prior articulated object datasets. *Medium* assemblies (2,412 clips) contain 3-8 parts where motion is kinematically coupled: a rotating input drives other components through linkages or contact, as in four-bar linkages, cam-and-follower mechanisms, and Hobson joints. *Hard* assemblies (3,024 clips) contain 10 or more parts, where dense kinematic coupling and frequent self-occlusion in 2D projection make motion understanding significantly more challenging. This tiered structure enables progressive evaluation of generative model behavior as interaction complexity increases, and allows failures to be localized to specific regimes of motion complexity rather than averaged away.

3.2 CLIP VARIATION

To prevent models from exploiting spurious correlations between visual appearance and motion behavior, each assembly is rendered under systematic variation along three axes: motion speed, camera viewpoint, and motion direction. Speed variants are defined by the fraction of the input joint range covered: *Slow* covers the first half of the joint range, *Mid* covers the full range, and *Fast* repeats the full range twice. For CAD assemblies, these correspond to 180°, 360°, and 720° of input rotation respectively. All clips are rendered at 16 FPS for a duration of 2 seconds, yielding 32 frames per clip. PartNet-Mobility assemblies are rendered from three camera viewpoints (Cam1, Cam2, Cam3) placed at fixed positions in the Unity virtual environment, while CAD assemblies are rendered from two viewpoints. Reversed clips are generated by reversing the frame order of forward clips, yielding semantically distinct motion directions (e.g., clockwise vs. counter-clockwise, opening vs. closing, extending vs. retracting). The dataset contains 14,508 forward and 6,648 reversed clips.

3.3 PROMPT STRUCTURE

For a video generation model to produce mechanically correct output, it must receive conditioning information that is precise, unambiguous, and complete. A vague prompt such as “a box opening”

leaves the model free to hallucinate motion that looks plausible but violates the assembly's kinematic structure. To address this, each clip in MechVerse is paired with a structured natural language prompt composed of six semantically distinct components: **(1) assembly overview**, **(2) part identification** by flat matte color, **(3) moving/rigid classification** per colored part, **(4) motion type** (rotation, translation, rotation+translation, or planar), **(5) direction of motion** (e.g., clockwise, closing, sliding left), and **(6) speed description** varying across slow, mid, and fast variants. This fixed ordering ensures every prompt is grounded in expert mechanical knowledge and fully characterizes the clip's visual and kinematic content. Figure 3 illustrates two representative examples with each component color-coded accordingly.

3.4 Fine-Grained Motion Control via Prompt Variation

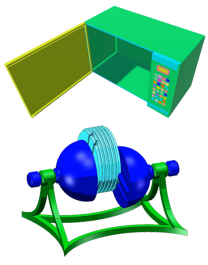
A key design principle of MechVerse is that the same input image can condition generation of multiple semantically distinct animations by varying only the prompt. Because speed, direction, and motion extent are explicitly encoded as separate components of each prompt, the dataset enables fine-grained control over motion behavior without any change to the visual input. Figure 4 illustrates this with four clips derived from the same washing machine assembly: varying the speed description produces animations where the door completes half its range (slow), one full cycle (mid), or two full cycles (fast), while varying the direction keyword switches the motion from closing to opening. This structured variation is present across all 1,357 assemblies in MechVerse and constitutes a systematic evaluation axis that existing articulated object datasets do not provide.

3.5 Prompt Annotation

Automating prompt generation via MLLMs failed systematically on Medium and Hard assemblies, coupled motion types were consistently misclassified, for example a rotating cam labeled as linear motion. Full details and failure examples are provided in Appendix A.6. We therefore developed an expert annotation pipeline in which two mechanical engineering annotators labeled each clip using a custom web tool (see Appendix A.4), recording part colors, moving/stationary classification, motion type, and direction keywords. Structured fields were then assembled into a draft prompt and reformatted into fluent prose using GPT-4o mini under a strict system prompt that prohibited hallucination and omission; the LLM served only as a formatter. Final prompts were verified in a second review application (see Appendix A.4). MechVerse releases video clips, prompts, and underlying 3D assets (OBJ files) for all assemblies.

3.6 Train/Test Split

A common failure mode in dataset design is test contamination through visual similarity, if a model has seen assemblies of the same category during training, strong performance on the test set may reflect memorization rather than generalization. To prevent this, MechVerse uses a category-level holdout split in which no assembly category present in the test set appears in the training set. The dataset is divided into 18,672 training clips and 1,004 test clips. The test set covers 8 held-out



A microwave with a yellow door and a green body has the yellow door swinging closed, rotating around a hinge on one side of the microwave. The green body of the microwave remains completely rigid and stationary throughout the animation. The door completes a full motion cycle, rotating to the end of its range.

Oldham's coupling features two blue shafts as the main driving motor and a light blue or cyan intermediate flange. The blue part rotates clockwise, and the cyan part also rotates clockwise. The blue and cyan moving parts are connected by kinematic constraints, so the motion of one drives the motion of the others. The green stationary parts remain completely rigid and stationary throughout the animation. The mechanism completes one full motion cycle.

Figure 3: Representative MechVerse prompt examples with color-coded components: assembly overview, part identification, moving/rigid classification, motion type, direction, and speed description. Top: Easy-tier example (microwave). Bottom: Hard-tier example (Oldham's coupling) where multiple parts are kinematically coupled.

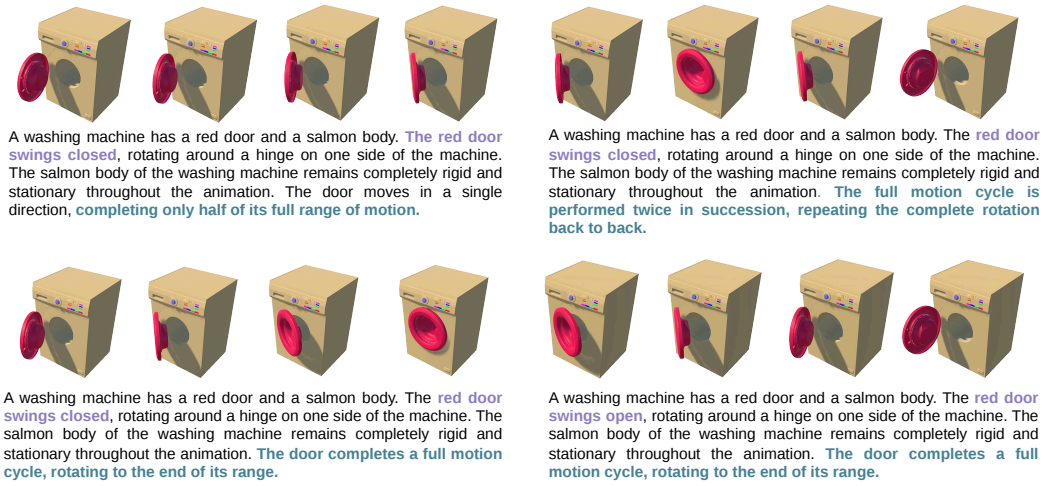


Figure 4: Fine-grained motion control in MechVerse via prompt variation. All four rows share the same input image (a washing machine with a red door and a salmon body). Varying the **direction** and **speed** components of the prompt produces four semantically distinct ground-truth clips: *Row 1*: closing, slow (half range); *Row 2*: closing, mid (full cycle); *Row 3*: opening, mid (full cycle); *Row 4*: closing, fast (two full cycles).

categories (Bucket, CamAndFollower, FoldingChair, Microwave, Refrigerator, Safe, UniqueMechanism, WashingMachine) spanning all three complexity tiers, with 476 Easy, 294 Medium, and 234 Hard clips. To ensure that test evaluation measures motion understanding rather than viewpoint generalization, all test clips are rendered from a single standardized camera viewpoint (Cam1). An additional 1,480 clips corresponding to alternative camera views of test assemblies are excluded from both splits.

4 Experiments

Evaluation Setup. We use MechVerse to evaluate whether image-to-video models can generate motion that is both visually coherent and consistent with the geometric and kinematic constraints of mechanical assemblies. The evaluation is performed on a test set that contains assemblies with varying levels of complexity (easy, medium, and hard), ensuring coverage of diverse interaction scenarios. Each model takes an input image and a textual description that includes object-level motion and interaction constraints and generates a video. We evaluate the generated outputs directly using default inference settings for each model, without additional modifications.No additional post-processing or normalization is applied.

Evaluation Models. We evaluate a broad set of recent image-to-video systems, including proprietary models, open-source models, and fine-tuned variants. The proprietary models include Wan 2.7 Wan AI (2026), Seedance 2 ByteDance Seed Team (2026), Kling 3 Kling AI (2026), Happy Horse 1.0 HappyHorse (2026), and Seedance 1.5 Pro ByteDance Seed Team (2025). The open-source models include Wan 2.2 Wan et al. (2025), CogVideoX Yang et al. (2024b), CogVideoX 1.5 Yang et al. (2024b), HunyuanVideo Kong et al. (2024), VideoCrafter Chen et al. (2024), DynamiCrafter Xing et al. (2024), ConsistI2V Ren et al. (2024), LTX-Video HaCohen et al. (2024), and LTX-2 HaCohen et al. (2026). We also evaluate fine-tuned variants of Wan 2.2 and CogVideoX trained on the MechVerse training split. These models represent a diverse set of recent approaches for image-to-video generation, spanning diffusion-based video generation, transformer-based video diffusion, and image-conditioned generation paradigms Blattmann et al. (2023); Chen et al. (2024); Girdhar et al. (2024); Wan et al. (2025); Kong et al. (2024). All models are evaluated under a consistent setup using the same input image and prompt pairs.

Metrics. We use VBench-I2VHuang et al. (2024) to measure standard image-to-video quality, including temporal consistency, motion smoothness, visual fidelity, and consistency with the input im-

Table 1: Evaluation results on **MechVerse**. VBench I2V metrics (scores in %): **Subj.** = Subject Consistency; **BG Cons.** = Background Consistency; **Motion** = Motion Smoothness; **T.Flick.** = Temporal Flickering **Dyn.** = Dynamic Degree; **Aesth.** = Aesthetic Quality; **Imaging** = Imaging Quality; **I2V Subj.** = I2V Subject fidelity. WorldModelBench metrics: **Instr.** = Instruction-following score; **Physics** = Physical law adherence; **Common** = Common-sense reasoning; **Sum** = aggregate of the three. Gold / Blue / Green = 1st / 2nd / 3rd best per column.

	VBench I2V								WorldModelBench			
Model	Subj.	BG Cons.	Motion	T.Flick.	Dyn.	Aesth.	Imaging	I2V Subj.	Instr.	Physics	Common	Sum
Proprietary Models												
Wan2.7	93.65	**96.07**	99.22	99.20	34.36	51.04	**64.09**	96.98	**2.00**	3.70	1.90	**7.59**
Happy Horse	**94.95**	96.03	99.43	**99.41**	18.73	50.83	58.68	**98.41**	1.95	3.70	**1.91**	7.56
Kling3	94.91	95.28	99.48	99.29	37.55	**51.65**	56.07	98.15	**2.02**	3.71	**1.93**	**7.66**
Open-Source Models												
CogVideoX-1.0	89.78	94.29	98.83	98.60	29.58	48.88	54.37	93.70	1.33	3.13	1.32	5.79
CogVideoX-1.5	87.37	92.63	99.04	98.72	50.20	45.09	57.28	93.93	1.34	3.17	1.16	5.68
DynamiCrafter	93.43	95.15	99.31	98.66	25.90	47.57	57.98	95.25	1.21	**3.75**	1.71	6.68
HunyuanVideo-1.5	**95.61**	**96.78**	**99.49**	**99.35**	12.85	**52.41**	56.89	**98.48**	1.88	**3.76**	**1.93**	7.56
LTX-2	90.66	93.80	99.18	99.01	44.52	51.08	55.20	94.89	1.67	3.27	1.37	6.31
LTX-Video	93.04	95.63	**99.51**	99.16	50.00	50.96	60.26	95.63	1.44	3.56	1.57	6.57
VideoCrafter	93.30	94.91	97.53	96.34	**65.24**	47.59	**70.00**	88.85	1.24	3.64	1.71	6.58
Wan2.2	89.64	92.57	98.45	98.10	44.02	47.89	58.59	95.13	1.63	3.36	1.57	6.56
ConsistI2V	89.85	92.52	98.80	98.06	19.96	42.45	58.77	91.60	1.12	3.47	1.06	5.65
Fine-tuned Models												
CogVideoX (FT)	93.13	94.90	99.18	99.13	14.54	47.72	50.04	94.59	1.47	3.54	1.64	6.64
Wan2.2 (FT)	90.44	93.36	98.54	98.13	**51.89**	49.50	55.62	93.71	1.43	3.31	1.41	6.15

age. This covers metrics such assubject consistency, background consistency, motion smoothness, temporal flickering, dynamic degree, aesthetic quality, imaging quality, and I2V subject fidelity. We further include WorldModelBench Li et al. (2025) metrics to evaluate whether the generated videos follow the instruction, remain physically plausible, and satisfy basic common-sense constraint (frame-wise and temporal quality). We exclude the fluid dynamics sub-category from the physics score as our dataset contains only rigid-body mechanical assemblies. We adopt this benchmark as our prompts describe mechanical assembly animations where models must follow explicit motion instructions and respect physical laws and constraints.

Training Details. We fine-tune both models using LoRA with rank 32 for 2 epochs on the full training set. We use a batch size of 1 for all fine-tuning experiments. All experiments are conducted on an NVIDIA 4-way GH200 GPU cluster with an aarch64 system architecture.

5 Results

5.1 Quantitative Results

We report quantitative results in Table 1. In general, all models obtain strong VBench-I2V scores, with most models achieving motion smoothness above 98 and temporal flickering close to 99. This suggests that current models can generate visually stable videos. However, they still struggle on WorldModelBench, where the task requires following motion instructions and preserving physically plausible motion. We find that Kling3 is the strongest closed-source model with the highest aggregate score of 7.66, followed by Wan2.7 and Horse with 7.59 and 7.56. Closed-source models are also stronger on instruction following, while open-source models perform noticeably worse on this metric.

Interestingly, open-source models remain competitive on perceptual video quality. HunyuanVideo-1.5 achieves the best scores on several VBench metrics, and VideoCrafter achieves the highest imaging quality. These findings show that open-source models can match or exceed closed-source models on standard perceptual metrics, but are less reliable when evaluation focuses on instruction following and physical plausibility.

We further break down results along three axes of our dataset: assembly complexity, motion speed, and direction. **Assembly complexity.** As shown in Appendix A.9, VBench metrics slightly drop from Easy to Hard, but the differences are small. Motion smoothness and temporal flickering re-

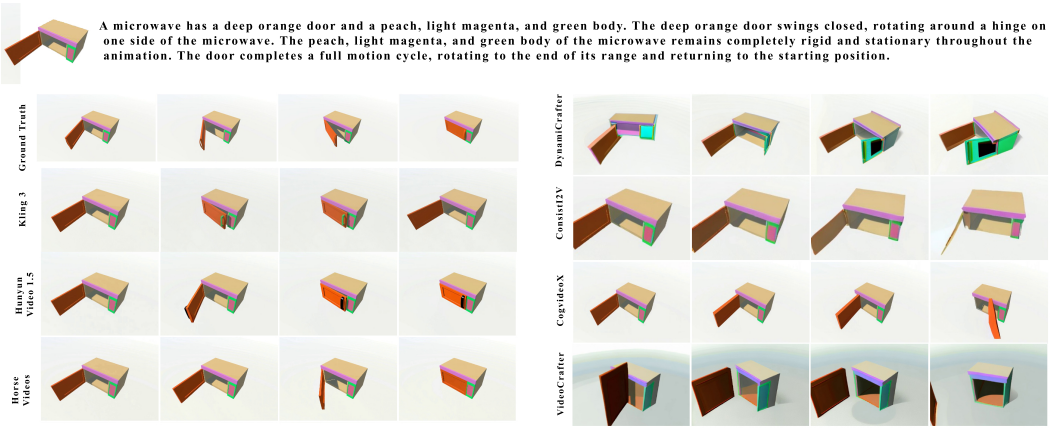


Figure 5: Qualitative comparison of generated videos on a Easy-tier MechVerse assembly.

main high across complexity levels, so VBench does not strongly separate Easy, Medium, and Hard cases.On WorldModelBench, however, we observe the opposite trend where scores actually go up with complexity. Scores often increase with complexity for example HunyuanVideo-1.5 goes from 7.26 on Easy to 7.67 on Medium and 8.03 on Hard, Happy Horse from 7.14 to 7.80 and 8.12, and DynamiCrafter from 6.34 to 6.86 and 7.12. Overall, complexity exposes a mismatch between automatic video-quality metrics and mechanism-level correctness. **Motion type.** As shown in Appendix A.10, we further split the results by forward and reversed motion directions. In general, the gap is small, but reversed motions perform slightly better for most models in WorldModelBench. For example, HunyuanVideo-1.5 increases from 7.44 on forward motion to 7.75 on reversed motion. However, we see the opposite trend in VBench metrics. Across all models, reversed clips consistently perform lower on subject consistency, background consistency, and I2V subject fidelity compared to their forward motion. **Effect of motion speed.** Appendix A.10 shows that speed separates models the least in WorldModelBench, with most scores changing by less than 0.2 across slow, medium, and fast clips. Proprietary models improve slightly on fast clips, while most open-source models remain flat or slightly degrade. For VBench metric, we observe opposite trend where Faster videos are harder for all models. Overall Finetuning results are mixed. **CogVideoX** improves on nearly every metric, with its WorldModelBench aggregate rising from **5.79 to 6.64**. **Wan2.2**, in contrast, gains in perceptual quality and dynamic degree but loses ground on physical-law adherence and common-sense reasoning, with its aggregate dropping from **6.56 to 6.15**. Mechanism-aware fine-tuning therefore helps weaker base models substantially, but for stronger models perceptual gains can come at the cost of physical reasoning.

5.2 Human Evaluation

We conduct human evaluation on 43 assemblies sampled from the test set, stratified by complexity level and motion category. Overall the total output is rated by six evaluators on 12 dimensions covering **motion correctness**, **visual realism**, and **prompt adherence**, using a 1–5 Likert scale (Figure 6). Overall, human scores remain modest across models, showing that mechanism-aware video generation is still challenging. HunyuanVideo 1.5 achieves the highest overall score among the evaluated models (2.91), followed by Horse (2.89), Wan2.7 (2.78), and fine-tuned Wan2.2 (2.65). The largest gaps appear on motion-related axes such as correct part motion, direction following, temporal completion, and kinematic coupling. While several models preserve visual appearance, they often fail to generate physically consistent part motion. Performance also drops from Easy to Medium and Hard examples. This degradation is clearer in human evaluation than in VBench, where scores remain relatively consistent. We observe reliable inter-rater agreement, with Krippendorff's α of 0.72 for motion correctness, 0.63 for visual realism, and 0.77 for prompt adherence. Agreement is higher for prompt adherence and motion correctness than for visual realism. This suggest that evaluators can more consistently judge whether a video follows the instruction and preserves the intended mechanism. Together, these results show that human evaluation captures mechanism-level failures that are often missed by standard perceptual metrics.

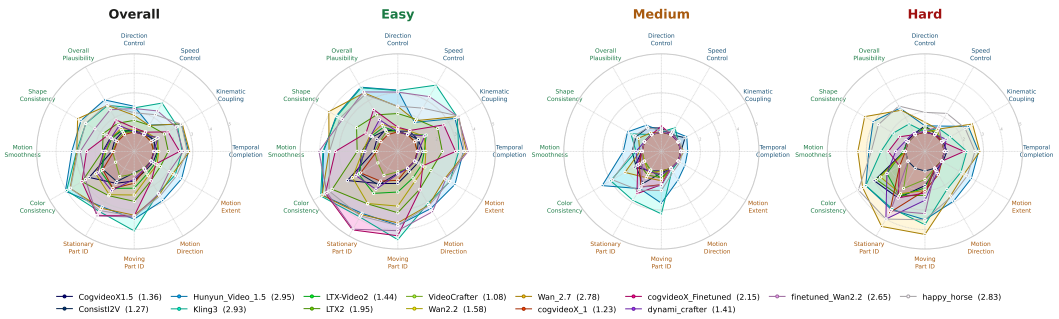


Figure 6: Human evaluation results across 14 models on 12 dimensions, grouped into three evaluation axes: **Motion Correctness**, **Visual Realism**, and **Prompt Adherence**.

6 Analysis

We list the analysis of our results below:

- **Automatic perceptual metrics are weak proxies for mechanism-aware generation.** Most models obtain high VBench-I2V scores, indicating strong appearance preservation and temporal stability. However, these metrics do not assess whether the generated motion satisfies the mechanical constraint in the prompt. As a result, visually smooth videos may still contain wrong part motion, deformation, reversed direction, or broken coupling.
- **MLLM judges struggle with fine-grained mechanical failures.** WorldModelBench provides better separation between models than VBench-I2V. However, its trends do not always align with human evaluation as assembly complexity increases. Many mechanical failures are local and subtle, such as incorrect motion of small parts, missing dependent motion, or motion that breaks over time.
- **Human evaluation reveals the main bottleneck.** Human scores show a clear gap between appearance-related and motion-related criteria. Models preserve color, shape, and overall plausibility more reliably than direction control, speed control, temporal completion, and kinematic coupling. This suggests that the main bottleneck is mechanism-level motion reasoning rather than visual fidelity.
- **Fine-tuning gives limited improvement.** Fine-tuned models improve over their zero-shot baselines on some WorldModelBench scores. This indicates that training on MechBench can improve mechanical prompt following. However, persistent errors in part motion, direction following, and coupling suggest that existing video models need more explicit representations of parts, joints, and dependency structures.

7 Conclusion and Limitations

Conclusion. We introduced MechVerse, a benchmark for evaluating image-to-video generation in structured mechanical interaction scenarios. Unlike standard video generation benchmarks that primarily focus on perceptual quality, temporal coherence, and prompt alignment, MechVerse evaluates whether generated videos preserve motion direction, dependency relationships, and interaction dynamics between mechanical components. Our evaluation across open-source and closed-source models shows that current image-to-video systems can often generate visually plausible and temporally smooth videos, but still struggle to produce functionally correct interactions. These results highlight a gap between visual realism and interaction-aware motion generation.

Limitations. MechVerse focuses on mechanical assemblies and does not cover all forms of physical interaction, such as deformable objects, fluids, human-object interactions, or open-world scenes. Since the benchmark is based on structured assemblies, the results may not fully reflect performance in natural videos. Our evaluation also relies on existing perceptual metrics, instruction-following scores, and human judgments, which may not capture every aspect of physical correctness. Future work can extend MechVerse to broader interaction categories, richer 3D annotations, and more detailed evaluations of force, contact, and long-horizon motion propagation.

AI USE STATEMENT

In this work, we used GPT-4o mini to format prompts in the dataset annotation pipeline, as described in the Prompt Annotation subsection. GPT-4o mini did not generate semantic content: all motion descriptions were authored by expert annotators. The model was used only to transform annotator-provided descriptions into a standardized prompt format.

We have not used generative AI tools to generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims or proofs; propose or refine hypotheses; design or provide feedback on research methodology or experiments; implement methods; clean or reformat datasets beyond the prompt-formatting procedure described above; perform qualitative or thematic data analysis; interpret results; or translate research materials. The remaining required disclosure tasks are not applicable to this work.

Additionally, we used ChatGPT to edit portions of the manuscript for grammar, clarity, and readability. The authors independently verified the dataset content, experimental protocol, results, technical claims, and citations. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves two expert annotators for dataset labeling and six evaluators for human evaluation. All participants were compensated at a rate of $20 per hour. Annotation instructions, the labeling interface, and the data-collection procedure are described in Appendix A.4. The human-evaluation dimensions and Likert-scale protocol are described in Section section 4.We collected and used only the information necessary for annotation and evaluation. Participant responses were analyzed in aggregate and are not reported in a manner intended to identify individual participants. We informed participants about the nature of the annotation or evaluation task, the expected time commitment, compensation, and how their responses would be used. We acknowledge that subjective judgments in human evaluation may be affected by annotator expertise, individual preferences,and the limited number of evaluators; we mitigate this risk by providing standardized instructions and applying the same evaluation protocol across all compared methods.

The released dataset includes video clips, and prompts. We release these materials in accordance with applicable licenses and data usage terms. We also recognize that generative video methods may present risks of misuse, including deceptive or misleading synthetic media. Our release is intended for research and benchmarking purposes, and users should evaluate relevant privacy, consent, licensing, and misuse considerations before downstream deployment.

REPRODUCIBILITY STATEMENT

To support reproducibility, we publicly release MechVerse on Hugging Face, including video clips, and prompts. The project page, available at `https://mechverse.pages.dev`, provides access to the dataset and accompanying documentation. All evaluated models are run under their default inference settings and are evaluated using the same input prompt–asset pairs, ensuring a consistent comparison protocol across methods. We evaluate generated outputs using metrics and models provided by VBench-I2V and WorldModelBench, which are publicly available with documented implementations. Additional details of the dataset construction, evaluation protocol, input pairs, and model settings are provided in the main paper and supplementary material.

REFERENCES

Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The IKEA ASM dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

ByteDance Seed Team. Seedance 1.5 pro. `https://seed.bytedance.com/en/seedance1_5_pro`, 2025. Accessed: 2026-05-07.

ByteDance Seed Team. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

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

Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7081–7091, 2023.

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Factorizing text-to-video generation by explicit image conditioning. In *European Conference on Computer Vision*, pp. 205–224. Springer, 2024.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

HappyHorse. Happyhorse 1.0. `https://happyhorse.app/`, 2026. Accessed: 2026-05-07.

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

Denys Iliash, Hanxiao Jiang, Yiming Zhang, Manolis Savva, and Angel X Chang. S2o: Static to openable enhancement for articulated 3d objects. *arXiv preprint arXiv:2409.18896*, 2024.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10236–10247, 2020.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Kling AI. Klingai 3.0 series. `https://kling.ai/`, 2026. Accessed: 2026-05-07.

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

Simon Le Cleac'h, Hong-Xing Yu, Michelle Guo, Taylor Howell, Ruohan Gao, Jiajun Wu, Zachary Manchester, and Mac Schwager. Differentiable physics simulation of dynamics-augmented neural objects. *IEEE Robotics and Automation Letters*, 8(5):2780–2787, 2023.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024a.

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

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

Yunong Liu, Cristobal Eyzaguirre, Manling Li, Shubh Khanna, Juan Carlos Niebles, Vineeth Ravi, Saumitra Mishra, Weiyu Liu, and Jiajun Wu. Ikea manuals at work: 4d grounding of assembly instructions on internet videos. 2024c.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024a.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024b.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. *Advances in Neural Information Processing Systems*, 37:123155–123181, 2024.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.

OpenAI. Video generation models as world simulators. `https://openai.com/research/video-generation-models-as-world-simulators`, 2024. Technical report.

OpenAI. Sora. `https://openai.com/sora/`, 2025. Accessed: 2025-10-07.

Mayank Patel, Rahul Jain, Asim Unmesh, and Karthik Ramani. Dynamo: Dependency-aware deep learning framework for articulated assembly motion prediction. *arXiv preprint arXiv:2509.12430*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

Asim Unmesh, Rahul Jain, Jingyu Shi, VK Chaithanya Manam, Hyung-Gun Chi, Subramanian Chidambaram, Alexander Quinn, and Karthik Ramani. Interacting objects: A dataset of object-object interactions for richer dynamic scene representations. *IEEE Robotics and Automation Letters*, 9(1):451–458, 2023.

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

Wan AI. Wan 2.7. `https://wan.video/`, 2026. Accessed: 2026-05-07.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8876–8884, 2019.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18826–18836, 2025.

Zihao Yan, Ruizhen Hu, Xingguang Yan, Luanmin Chen, Oliver Van Kaick, Hao Zhang, and Hui Huang. Rpm-net: recurrent prediction of motion and parts from point cloud. *arXiv preprint arXiv:2006.14865*, 2020.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations (ICLR)*, 2024a.

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

Hao Zheng, Regina Lee, and Yuqian Lu. Ha-vid: A human assembly video dataset for comprehensive assembly knowledge understanding. *Advances in Neural Information Processing Systems*, 36:67069–67081, 2023.

A MechVerse: Extended Dataset Details

A.1 Dataset Statistics

Table 2 provides a full breakdown of MechVerse clip counts by category, complexity tier, speed, camera, and motion direction across the full dataset, training split, and test split.

Table 2: MechVerse dataset statistics across splits.

	Full	**Train**	**Test**
Total clips	21,156	18,672	1,004
Total assemblies	1,357	1,153	204
Source			
PartNet-Mobility	15,720	14,292	476
CAD (MechVerse)	5,436	4,380	528
Complexity			
Easy	15,720	14,292	476
Medium	2,412	1,824	294
Hard	3,024	2,556	234
Speed			
Slow	5,742	5,033	295
Mid	9,672	8,606	414
Fast	5,742	5,033	295
Direction			
Forward	14,508	12,909	621
Reversed	6,648	5,763	383
Camera			
Cam1	7,958	6,954	1,004
Cam2	7,958	6,954	–
Cam3	5,240	4,764	–

A.2 Category List

MechVerse covers 141 standard categories. Categories sourced from PartNet-Mobility include: Box, Bucket, Dishwasher, Door, Eyeglasses, Fan, Faucet, FoldingChair, Knife, Laptop, Microwave, Oven, Pen, Refrigerator, Safe, StorageFurniture, Table, WashingMachine, and Window. Categories curated as CAD assemblies include: CamAndFollower, CheerleaderToy, EllipticalTrammel, EnginePiston, FourBarLinkage, HobsonJoint, UJoint, and UniqueMechanism. The UniqueMechanism category aggregates 114 individual assembly types for which only one or a small number of instances exist; when expanded, MechVerse covers 141 distinct mechanical categories.

A.3 Rendering Pipeline

PartNet-Mobility assemblies were imported into Unity as pre-processed OBJ files with kinematic joint parameters extracted from the dataset's `mobility_v2.json` and `result.json` metadata files. Parts were assigned flat matte colors from a fixed palette using a per-joint material assignment strategy, with one consistent color per joint across all OBJs belonging to that joint, and a randomized palette starting index per assembly to avoid color bias across the dataset. Animation was driven by a frame-accurate normalized time stepping system using `SetNormalizedTime()`, which steps through exact normalized time positions per frame without relying on Unity's real-time clock, avoiding frame capture stalls observed with `Time.captureFramerate`. Three cameras were placed at fixed positions in the virtual environment to capture Cam1, Cam2, and Cam3 viewpoints. CAD assemblies were animated using the kinematic solvers built into the respective CAD software, with joint linkages explicitly defined, and rendered from two viewpoints at the same frame specification.

All clips are 32 frames at 16 FPS (2 seconds). Reversed clips were generated by reversing the frame order of the corresponding forward clip in post-processing.

A.4 Annotation Pipeline

Figure 7 shows the custom web annotation tool used in Stage 1 of the annotation process. For each clip, annotators selected all visible part colors from a fixed palette, classified each color as moving or stationary, selected the motion type for each moving part (Rotation, Translation, Rotation+Translation, Planar), and specified a direction keyword from a controlled vocabulary (Clockwise, Counter-clockwise, Opening, Closing, Extending, Retracting, Folding, Unfolding, Sliding Left, Sliding Right, Sliding Up, Sliding Down, Tilting Left, Tilting Right). A category-level description and an optional quick note capturing assembly-specific details were also recorded. Speed-specific text fields were pre-populated from category-level templates and editable by the annotator. The structured annotation fields were assembled into a draft prompt via a deterministic semantic mapping algorithm and then passed to the GPT-4o mini API for reformatting into fluent prose under a strict system prompt that required all annotated fields to be preserved verbatim and prohibited any addition or omission of content. Figure 8 shows the prompt review web application used in Stage 2, where annotators reviewed each clip alongside its generated prompt and submitted corrections where the prompt did not accurately reflect the observed motion.

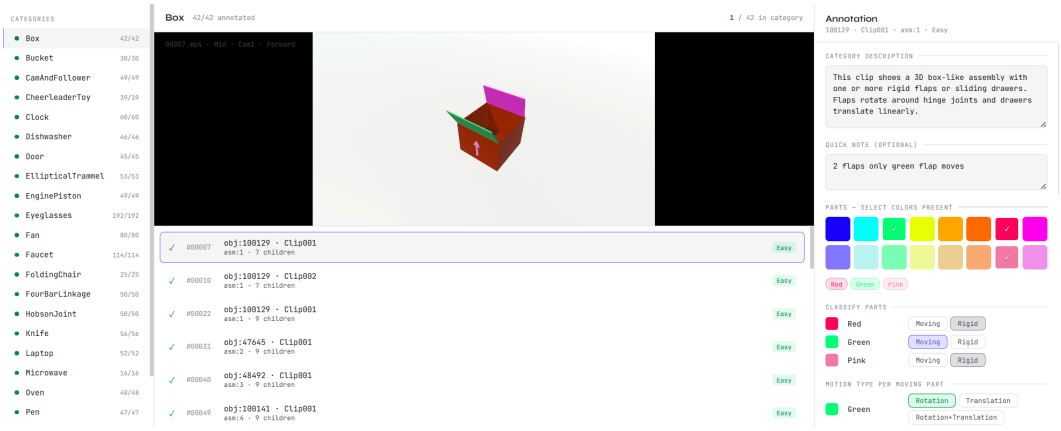


Figure 7: Stage 1 annotation web application. Annotators select visible part colors, classify each as moving or stationary, specify motion type and direction per moving part, and review category-level and speed-specific prompt text.

A.5 Sample Prompt Structure

The following shows a representative entry from the final dataset JSON for two speed variants of the same assembly clip, illustrating how prompt content varies across speed while all other annotation fields remain consistent.

```
Slow variant:
"A bucket with a blue handle and a pink body has the blue handle tilting to
the left, rotating around a horizontal axis attached to the sides of the
bucket. The pink body remains completely rigid and stationary. The handle
moves in a single direction, completing only half of its full range of motion."

Mid variant:
"A bucket with a blue handle and a pink body features a blue handle that tilts
to the left, rotating around a horizontal axis attached to the sides of the
bucket. The pink body remains completely rigid and stationary. The handle
completes a full motion cycle, sweeping from one side to the other."
```

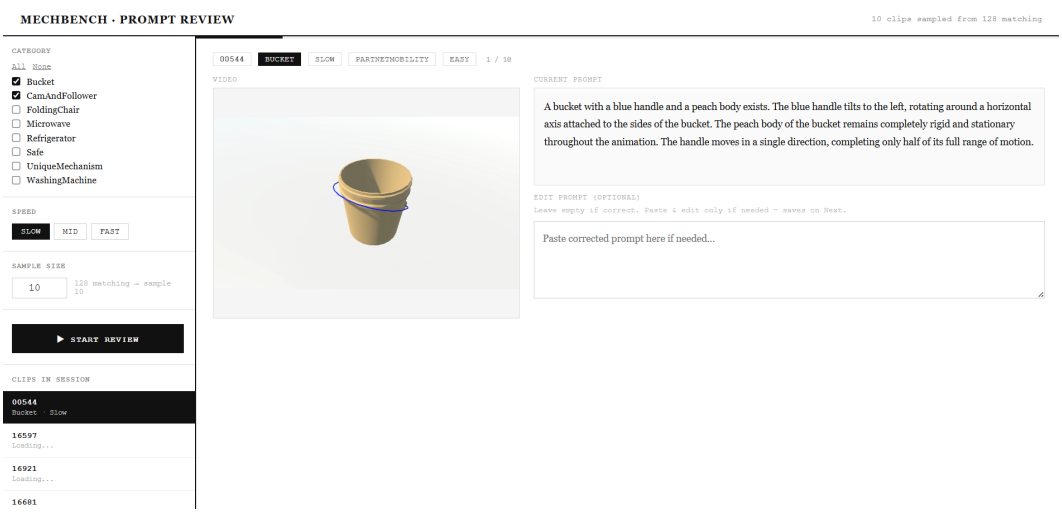


Figure 8: Stage 2 prompt review web application. Each clip is displayed alongside its GPT-4o mini generated prompt. Annotators verify accuracy and submit corrections where needed.

A.6 Preliminary MLLM-Based Annotation Attempts

Prior to developing the expert manual annotation pipeline described in Section 3, we explored whether multimodal large language models (MLLMs) could be used to automatically annotate clips with structured motion descriptions. We evaluated an MLLM-based pipeline across four objectives: (a) part counting, (b) detection of moving parts, (c) detection of stationary parts, and (d) motion type classification. We further explored three approaches for converting the structured output into natural language prompts suitable for video generation conditioning.

Part Counting. A single middle frame from each animation was sent to the MLLM along with a prompt requesting a structured JSON response containing the total part count and per-part color descriptions. This objective worked reliably for simple assemblies with clearly distinguishable part colors.

Moving and Stationary Part Detection. Three temporally-spaced frames (early, middle, late) were sent together, and the model was asked to compare them and identify which parts moved and which remained stationary. For simple assemblies, this produced broadly correct outputs.

Motion Type Classification. Motion type classification (rotation, translation, or oscillating) was performed concurrently with part detection. The model returned a motion type field for each identified moving part along with a free-text description of the observed motion.

Prompt Generation Approaches. Beyond structured annotation, we explored three strategies for generating natural language conditioning prompts:

Approach 1: Image + Video → Prompt Early and late frames were provided and the model was prompted to generate a conditioning prompt describing the observed motion.

Approach 2: Caption → Prompt Only the structured annotation fields (part count, moving parts, stationary parts, motion types) were provided as text, no images. The model generated a prompt purely from textual annotations.

Approach 3: Image + Video + Caption → Prompt Both the frame pair and structured annotations were provided together, combining visual and textual information.

Figure 9 shows results for an Easy-tier Box assembly. All three approaches produce broadly correct and coherent prompts, the model correctly identifies the oscillating lid, the stationary body, and the hinge-based rotation. This suggests that for simple assemblies with independent part motion and visually distinct colors, MLLM-based annotation is a viable approach.



(a) Part Count: 3
red/maroon — main cubic block
purple/magenta — top rectangular piece
green — cylindrical rod/shaft

(b) Moving Parts
magenta/pink — rectangular lid/cover
(c) Stationary Parts
dark red/maroon — main cube body
green — cylindrical rod/pin

(d) Motion Types
magenta/pink — oscillating
rotates back and forth between open and closed positions, pivoting around the hinge point

Approach 1: Image + Video → Prompt

```
This three-part assembly consists of a base cube that remains completely stationary, while the top lid piece executes a smooth rotational opening motion by pivoting upward around its back edge, and the green rod component simultaneously performs a linear extraction motion by sliding outward from the cube in a diagonal upward direction, creating a coordinated opening and ejection sequence.
```

Approach 2: Caption → Prompt

```
This three-part mechanical assembly should animate with the purple rectangular lid piece oscillating in a rotational motion, pivoting back and forth around its hinge point to alternately open and close over the main body, while the dark red cubic main body and the green cylindrical rod remain completely stationary throughout the entire motion sequence.
```

Approach 3: Image + Video + Caption → Prompt

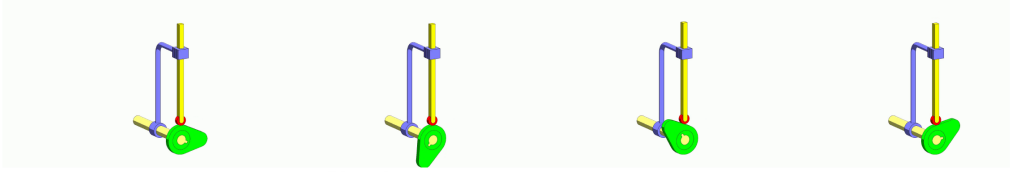
```
This three-part mechanical assembly consists of a dark red cubic main body, a green cylindrical rod that acts as a hinge pin, and a purple rectangular lid or cover piece. The main red cube and green cylindrical rod remain completely stationary throughout the animation, serving as the fixed base and pivot point respectively. The purple rectangular lid performs an oscillating rotational motion, pivoting back and forth around the green cylindrical hinge pin, alternating between an open position where it's tilted upward and away from the cube, and a closed position where it sits flat on top of the red cube, creating a box-opening and closing animation cycle.
```

Figure 9: MLLM-based annotation results for an Easy-tier assembly (Box). Four temporally-spaced frames are shown alongside structured part analysis (a–d) and generated prompts from all three approaches. The model correctly identifies the oscillating magenta lid, the stationary body and rod, and produces coherent conditioning prompts across all three approaches.

However, Figure 10 reveals a systematic failure on a Medium-tier CamAndFollower assembly. The green cam component, which performs continuous rotational motion driven by kinematic coupling, is misclassified as linear motion across all three approaches. This error propagates directly into the generated prompts, which describe the cam as moving vertically downward rather than rotating. We observed this class of failure consistently across Medium and Hard tier assemblies, where coupled and non-obvious motion types were regularly misidentified regardless of which prompt generation approach was used. These findings motivated the expert manual annotation pipeline described in Section 3, in which two annotators with mechanical engineering backgrounds annotated all clips using a controlled vocabulary and structured web interface, ensuring that motion type labels reflect true kinematic behavior rather than superficial visual appearance.

A.7 Human Evaluation — Per-Model Radar Plots

Figure 11 shows the per-model radar plots for all 14 evaluated models across the 12 human evaluation dimensions. Each subplot corresponds to one model, with the mean score across all dimensions shown in the title. The filled area represents the model's score profile; the dotted ring marks the midpoint score of 3.0. Models are sorted by overall mean score. The plots reveal distinct failure signatures across models: most open-source models score consistently low on Kinematic Coupling and Direction Following, while proprietary models such as Kling3 and Horse show broader coverage of the evaluation dimensions. Notably, even the highest-scoring models remain well below the midpoint on several motion-related dimensions, underscoring the difficulty of mechanically-consistent video synthesis.



(a) Part Count: 5
yellow — vertical rod/shaft
blue — connecting bracket/clamp
yellow — horizontal rod/pin
green — curved hook/lever component
yellow — cylindrical bearing/bushing

(b) Moving Parts
red — red piston rod
green — green piston head/cylinder

(c) Stationary Parts
yellow — vertical guide rod
blue — frame/housing structure
yellow — horizontal connecting rod/pin

(d) Motion Types
red — linear
moves vertically downward along the yellow guide rod
green — linear
moves vertically downward following the red piston rod

Approach 1: Image + Video → Prompt

```
This mechanical assembly contains five distinct parts where the blue frame structure, yellow rod, and blue pivot joint
remain completely stationary throughout the animation. The green curved lever arm executes a smooth rotational motion,
pivoting clockwise around the blue joint connection point, while the small yellow cylindrical component that connects the
lever to the blue joint also rotates in coordination with the lever's movement, maintaining its attachment relationship as
the lever swings through its arc of motion.
```

Approach 2: Caption → Prompt

```
This mechanical assembly contains five parts where two components exhibit linear motion while three remain stationary. The
red piston rod moves vertically downward along the yellow guide rod in a smooth linear translation, while the green piston
head/cylinder follows the same downward linear motion in coordination with the red piston rod. The yellow vertical guide
rod, blue frame/housing structure, and yellow horizontal connecting rod/pin all remain completely stationary throughout the
animation, serving as fixed reference points and structural supports for the moving piston assembly.
```

Approach 3: Image + Video + Caption → Prompt

```
This mechanical assembly contains five distinct parts where the yellow vertical rod, blue connecting bracket, and yellow
horizontal pin remain completely stationary throughout the animation. The red piston rod performs smooth linear motion,
moving vertically downward along the yellow guide rod in a controlled descent. Simultaneously, the green curved hook
component exhibits synchronized linear motion, also moving vertically downward in parallel with the red piston rod,
maintaining their relative positioning as they descend together through the assembly's operating cycle.
```

Figure 10: MLLM-based annotation failure on a Medium-tier assembly (CamAndFollower). The green cam component is misclassified as exhibiting linear motion (column d) across all three prompt generation approaches, when it is in fact rotating continuously under kinematic coupling. This systematic failure on coupled mechanisms was observed consistently across Medium and Hard tier assemblies and motivated the expert manual annotation pipeline used in MechVerse.

A.8 Qualitative Results by Complexity Tier

Figures 12 and 13 extend the Easy-tier qualitative comparison from the main paper to Medium and Hard assemblies.

On the Medium-tier CamAndFollower assembly (Figure 12), HunyuanVideo-1.5 is the only open-source model that produces motion resembling the ground truth. No model correctly generates the coupled cam-follower interaction, while several attempt to rotate the green cam, the followers do not respond with the expected linear translation, violating their kinematic constraints. Kling3 hallucinates a fourth follower, exhibits color flickering, and rotates the followers instead of translating them. Models on the right panel perform uniformly poorly; Wan2.7 hallucinates an entirely different mechanism.

On the Hard-tier assembly (Figure 13), the top three models produce visually appealing animations that superficially resemble the ground truth. However, the kinematic constraints in the base mechanism are not respected, the rotating figures move plausibly while the underlying linkages fail to propagate motion correctly. This explains why aggregate human scores can appear relatively high on Hard assemblies (Figure 6) even when kinematic coupling scores remain low: models appear to hallucinate motion patterns from visually similar real-world content rather than reasoning from the mechanical structure. Models on the right panel generate poor or incoherent animations with no resemblance to the intended motion.

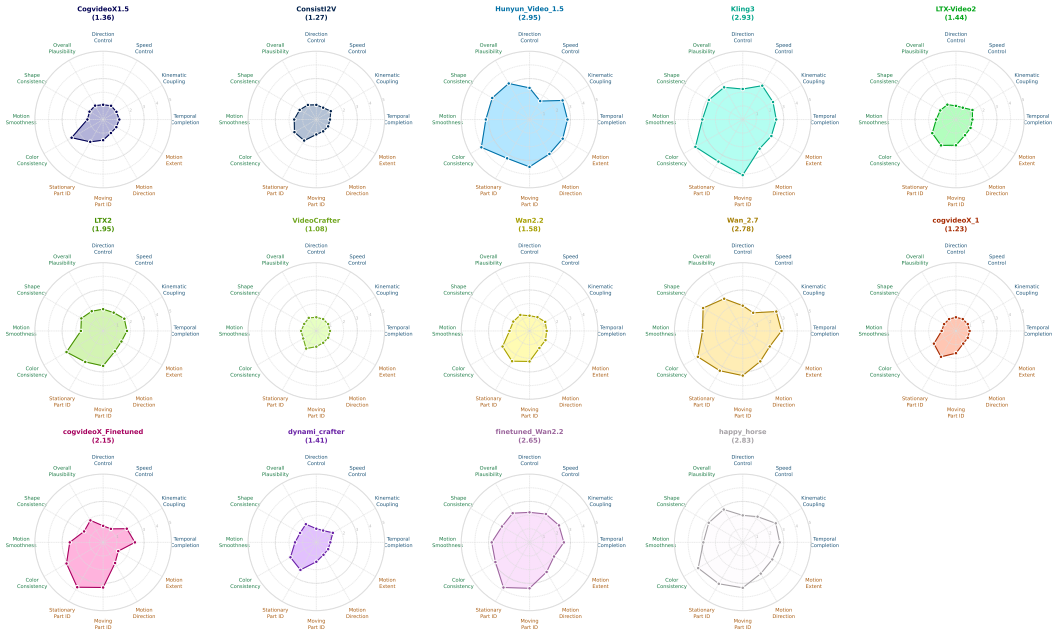


Figure 11: Per-model human evaluation radar plots across all 15 evaluated models. Each subplot shows one model's mean Likert scores (1–5) across 12 evaluation dimensions: Stationary Rigidity (Q1), Correct Part Motion (Q2), Motion Direction (Q3), Motion Extent (Q4), Overall Plausibility (Q5), Shape Consistency (Q6), Motion Smoothness (Q7), Color Consistency (Q8), Kinematic Coupling (Q9), Temporal Completion (Q10), Speed Following (Q11), and Direction Following (Q12). The dotted ring marks the midpoint score of 3.0. Mean score across all dimensions is shown in parentheses below each model name.

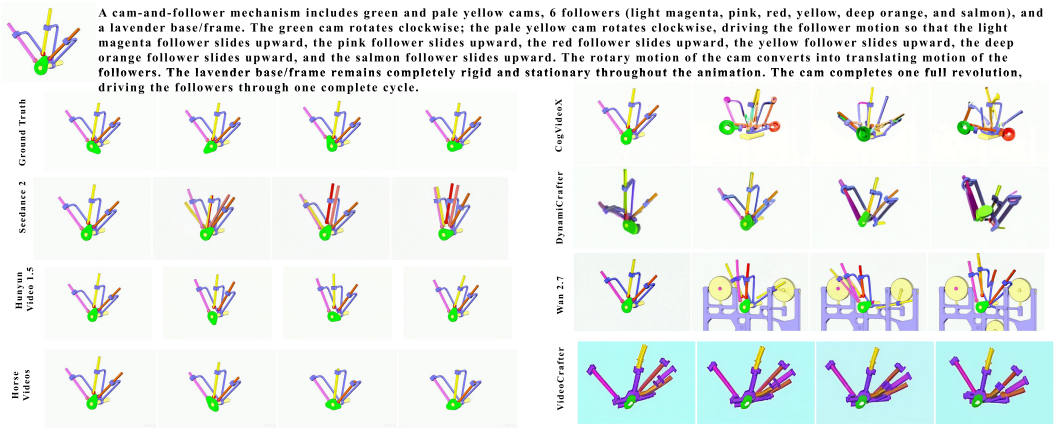


Figure 12: Qualitative comparison on a Medium-tier assembly (CamAndFollower). No model correctly reproduces the coupled cam-follower interaction. HunyuanVideo-1.5 produces the closest approximation among open-source models.

A.9 Quantitative Results by Complexity Tier

Figures 14 and 15 show VBench-I2V and WorldModelBench scores stratified by kinematic complexity tier. On VBench metrics, scores are largely stable across Easy, Medium, and Hard assemblies — motion smoothness and temporal flickering remain high throughout, confirming that perceptual quality does not degrade meaningfully with complexity. On WorldModelBench, we observe the opposite: aggregate scores tend to increase from Easy to Hard. For example, Kling3 goes from 7.41 on

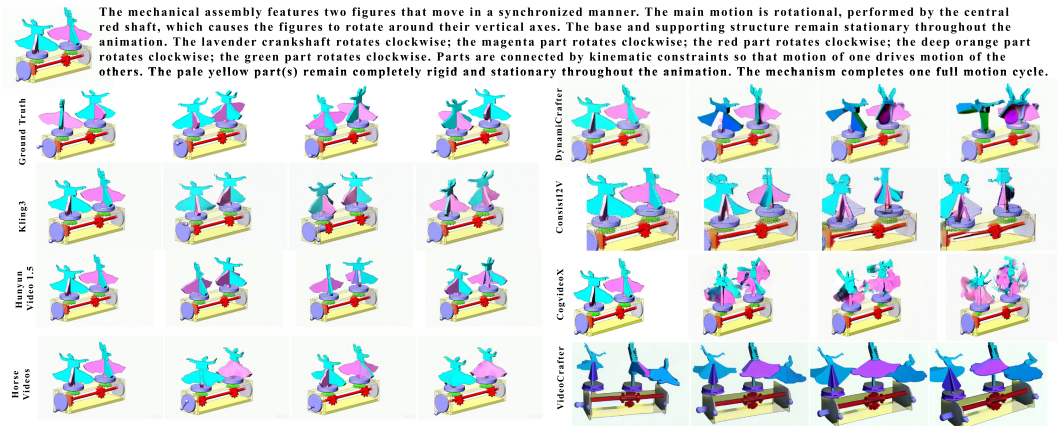


Figure 13: Qualitative comparison on a Hard-tier assembly. Top models produce visually plausible animations but fail to propagate motion through the underlying kinematic structure, illustrating why perceptual scores can remain high while mechanism-level correctness is low.

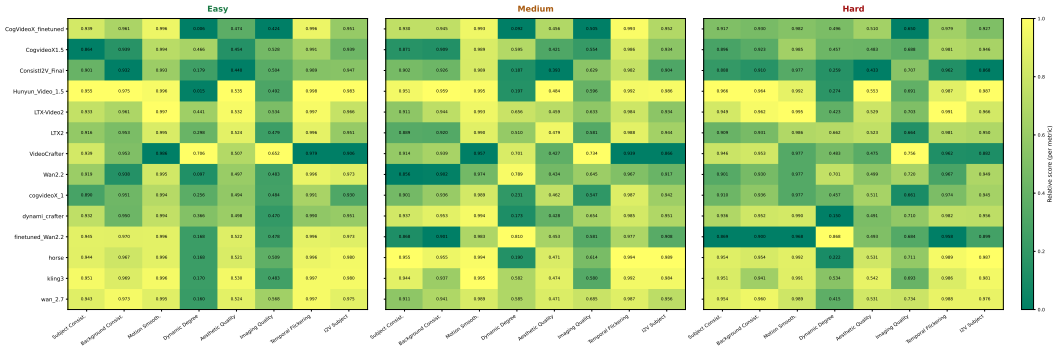


Figure 14: VBench-I2V scores stratified by kinematic complexity tier (Easy / Medium / Hard). Perceptual metrics remain largely stable across tiers, indicating that VBench does not capture the degradation in mechanical correctness as coupling complexity increases.

Easy to 7.80 on Medium and 8.00 on Hard, and Horse from 7.15 to 7.80 and 8.12. This counterintuitive trend reflects a known limitation of MLLM-based judges on complex assemblies: as coupling complexity increases, videos that appear globally plausible receive higher scores even when part-level kinematic constraints are violated. Human evaluation (Section **??**) reveals the opposite trend, with motion correctness and kinematic coupling scores falling on Medium and Hard assemblies.

A.10 Quantitative Results by Motion Speed and Direction

Figures 16 and 17 show scores split by motion speed (Slow / Mid / Fast) and direction (Reversed). On WorldModelBench, speed has the smallest effect of all axes, most models change by less than 0.2 across speed variants. Proprietary models improve slightly on fast clips while most open-source models remain flat or degrade marginally. On VBench, the opposite holds: faster clips are harder for all models, with dynamic degree and subject consistency both declining at higher speeds. For direction, reversed clips score slightly higher on WorldModelBench for most models (e.g., HunyuanVideo-1.5 goes from 7.59 on forward to 7.75 on reversed) but lower on VBench consistency metrics, suggesting that reversed motion is easier to rate as plausible but harder to render with visual fidelity.

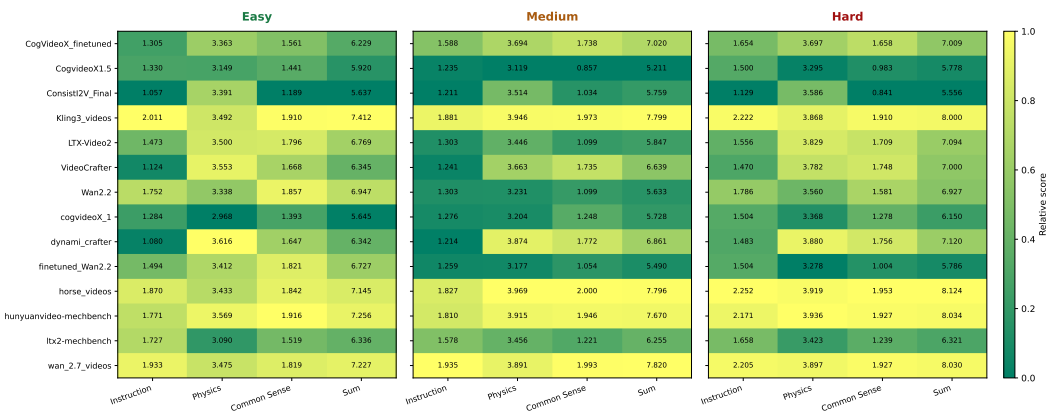


Figure 15: WorldModelBench scores stratified by kinematic complexity tier (Easy / Medium / Hard). Aggregate scores increase with complexity, reflecting the tendency of MLLM judges to rate visually plausible but mechanically incorrect videos favourably on complex assemblies.

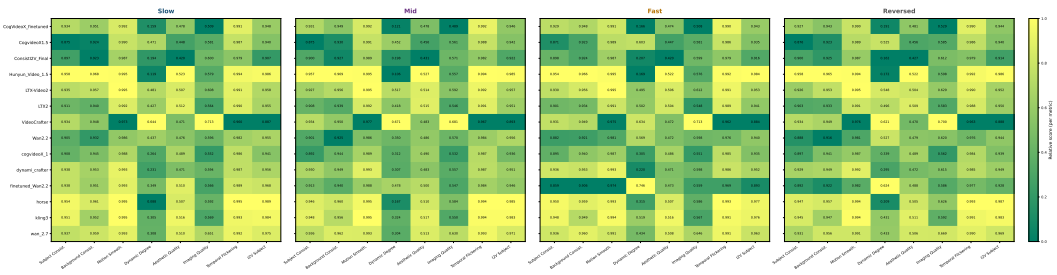


Figure 16: VBench-I2V scores stratified by motion speed (Slow / Mid / Fast) and direction (Reversed). Faster clips degrade perceptual quality metrics across most models.

A.11 Statistical Robustness of Human Evaluation and Correlation with Automated Metrics

Confidence intervals and trend significance. To verify that the complexity-driven degradation reported in Section **??** is statistically robust rather than a small-sample artifact, we compute 95% confidence intervals on the mean per-tier human score using a cluster bootstrap that resamples assemblies with replacement over 3,000 iterations, treating the assembly as the natural unit of repeated measurement since every model is rated on the same 43 assemblies, and we test for a monotonic trend across complexity tiers using the Jonckheere–Terpstra test, which is appropriate for ordered-group comparisons on a bounded Likert scale. Table 3 reports the resulting per-tier means and intervals; pooled across all models and raters, the trend is strongly negative (Jonckheere–Terpstra $z = -8.51$, $p < 0.0001$), and applied to each model individually, 9 of 14 models show a significant negative trend ($p < 0.01$; Table 4), while the remaining five (CogVideoX1.5, LTX-Video2, VideoCrafter, CogVideoX-1.0, DynamiCrafter) show no significant tier effect ($p > 0.1$), indicating uniformly weak performance across complexity levels rather than complexity-sensitive degradation for these models. We note that the tier-level trend is not strictly monotonic: Hard-tier assemblies score significantly higher than Medium-tier assemblies, evidenced by non-overlapping 95% confidence intervals in Table 3, an effect we attribute to the small Hard-tier sample ($n = 8$ assemblies) and to compositional differences in which categories were sampled into each tier; the dominant and robust effect remains the large, highly significant drop from Easy to both Medium and Hard tiers.

Pairwise model comparisons. A Friedman test across all 14 models, using the 43 fully-crossed assembly blocks rated by our primary annotator, confirms that models differ significantly in aggregate performance ($\chi^2 = 279.7$, $p < 0.0001$), indicating the human study has genuine discriminative power across the full model set. However, pairwise Wilcoxon signed-rank tests among the top-

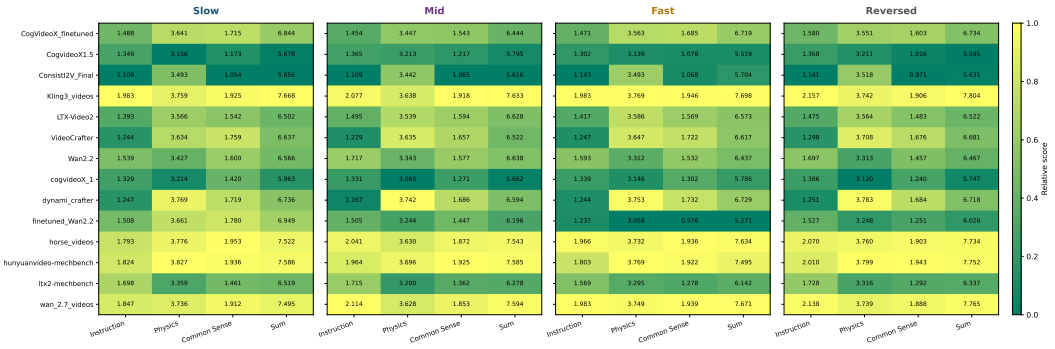


Figure 17: WorldModelBench scores stratified by motion speed (Slow / Mid / Fast) and direction (Reversed). Speed has the smallest effect on WorldModelBench scores relative to complexity or direction.

Table 3: Mean human evaluation score per complexity tier with 95% cluster-bootstrap confidence intervals (resampling assemblies, 3,000 iterations).

Tier	N assemblies	Mean	95% CI
Easy	25	2.44	[2.22, 2.66]
Medium	10	1.26	[1.19, 1.34]
Hard	8	1.84	[1.71, 1.98]

ranked models, Holm-corrected for multiple comparisons, show that HunyuanVideo-1.5, Kling3, and Happy Horse are not statistically distinguishable from one another (Table 5), indicating that while the gap between top-tier and bottom-tier models is large and robust, fine-grained ranking among the top three models should not be interpreted as statistically meaningful given the current sample size.

Inter-rater reliability with confidence intervals. We refine the single-point Krippendorff's α values reported in Section **??** in two ways: first, we recompute α separately for the subset of models with near-complete second-rater coverage (six models with $\geq 40/43$ assemblies double-rated, versus two models with sparse coverage of 2 and 17 assemblies respectively), and second, we place a cluster bootstrap 95% confidence interval on the overall-score α for this reliable subset, resampling assemblies with replacement over 300 iterations. Table 6 shows that agreement is consistently higher once incomplete-coverage models are excluded, and the bootstrap interval on the overall score, $\alpha = [0.87, 0.96]$, confirms strong and reliable agreement for the fully double-rated subset.

Correlation between human judgment and automated metrics. To substantiate our claim that VBench-I2V and WorldModelBench are weak proxies for mechanism-level correctness, we correlate per-clip automated metric scores against human ratings for the subset of clips covered by both, joined via the Mid/Forward clip used as the primary stimulus in each rated assembly, computing Spearman's ρ given the ordinal nature of the Likert scale, with 95% confidence intervals from a cluster bootstrap resampling assemblies (1,000 iterations). At the model level, overall VBench-I2V quality and overall human score are essentially uncorrelated ($\rho = -0.12$, $p > 0.1$, $n = 14$), indicating that a model's aggregate perceptual-quality score carries no information about its mechanism correctness as judged by humans. Table 7 reports dimension-matched clip-level correlations: VBench's subject_consistency correlates moderately with human-judged shape consistency and motion_smoothness correlates moderately with human-judged motion smoothness across all tiers, indicating these two sub-metrics do capture something related to what they claim to measure, whereas dynamic_degree, VBench's measure of raw motion magnitude, is negatively correlated with human-judged kinematic coupling on Easy and Medium tiers and uncorrelated on Hard, directly demonstrating that a model producing more motion is judged by human raters as coordinating parts worse rather than better.

Table 4: Per-model Jonckheere–Terpstra trend statistics across Easy → Medium → Hard. Negative z indicates a decreasing trend.

Model	z	p	Trend
finetuned_Wan2.2	-6.31	$p < 0.0001$	decreasing
CogVideoX-Finetuned	-4.41	$p < 0.0001$	decreasing
LTX2	-4.62	$p < 0.0001$	decreasing
Kling3	-3.88	$p < 0.001$	decreasing
Wan2.7	-3.52	$p < 0.001$	decreasing
Happy Horse	-3.59	$p < 0.001$	decreasing
HunyuanVideo-1.5	-3.09	$p < 0.01$	decreasing
ConsistI2V	-3.08	$p < 0.01$	decreasing
Wan2.2	-3.25	$p < 0.01$	decreasing
DynamiCrafter	-0.39	$p > 0.1$	n.s.
CogVideoX1.5	-0.43	$p > 0.1$	n.s.
LTX-Video2	-1.14	$p > 0.1$	n.s.
CogVideoX-1.0	1.49	$p > 0.1$	n.s.
VideoCrafter	0.03	$p > 0.1$	n.s.

Table 5: Pairwise Wilcoxon signed-rank tests between top-ranked models (paired by assembly, $n = 43$), Holm-corrected across the five comparisons.

Pair	Median Diff.	p (raw)	p (Holm)
HunyuanVideo-1.5 vs. Kling3	-0.25	$p < 0.1$	$p > 0.1$
HunyuanVideo-1.5 vs. Happy Horse	-0.25	$p < 0.1$	$p > 0.1$
Kling3 vs. Wan2.7	0.17	$p > 0.1$	$p > 0.1$
Kling3 vs. Happy Horse	-0.08	$p > 0.1$	$p > 0.1$
HunyuanVideo-1.5 vs. Wan2.7	0.00	$p > 0.1$	$p > 0.1$

Correlation with WorldModelBench. At the model level, human and WorldModelBench aggregate scores correlate moderately and significantly ($\rho = 0.61$, $p < 0.1$, $n = 12$). At the clip level, stratified by tier, this correlation strengthens with complexity, from $\rho = 0.13$ on Easy to $\rho = 0.50$ on Hard (Table 8), which we emphasize measures relative ranking agreement, that is, whether WorldModelBench and human raters agree on which models perform better than which within a tier, and is distinct from the absolute calibration claim made in Section **??**, namely that WorldModelBench's mean score rises with complexity even as ground-truth difficulty increases. Both properties can hold simultaneously: WorldModelBench can preserve, or even improve, its ability to discriminate between models on harder assemblies while its absolute score scale remains systematically inflated relative to true mechanism correctness, a level-calibration bias rather than a wholesale failure to discriminate between models. Correlating WorldModelBench instead against a motion-specific human composite, the mean of stationary rigidity, correct part motion, and kinematic coupling, excluding overall plausibility and appearance dimensions, yields systematically weaker correlations than against the full composite (Table 8), suggesting WorldModelBench's judgments are more influenced by overall visual plausibility than by mechanism-specific correctness, consistent with and further specifying our central claim that automated judges conflate surface plausibility with true kinematic correctness.

Table 6: Krippendorff's α by evaluation axis, all models versus the subset with near-complete dual-rater coverage.

Axis	α (all models)	α (clean subset)
Motion Correctness	0.72	0.71
Visual Realism	0.63	0.77
Prompt Adherence	0.77	0.84

Table 7: Dimension-matched Spearman correlations between human evaluation and VBench-I2V, by complexity tier. All correlations are computed at the clip level.

Human Dimension	VBench Dimension	Tier	ρ (p)
Shape Consistency (Q6)	subject_consistency	Overall	0.39 ($p < 0.0001$)
Shape Consistency (Q6)	subject_consistency	Easy	0.33 ($p < 0.0001$)
Shape Consistency (Q6)	subject_consistency	Medium	0.40 ($p < 0.0001$)
Shape Consistency (Q6)	subject_consistency	Hard	0.46 ($p < 0.0001$)
Motion Smoothness (Q7)	motion_smoothness	Overall	0.54 ($p < 0.0001$)
Motion Smoothness (Q7)	motion_smoothness	Easy	0.47 ($p < 0.0001$)
Motion Smoothness (Q7)	motion_smoothness	Medium	0.39 ($p < 0.0001$)
Motion Smoothness (Q7)	motion_smoothness	Hard	0.36 ($p < 0.001$)
Kinematic Coupling (Q9)	dynamic_degree	Overall	-0.25 ($p < 0.0001$)
Kinematic Coupling (Q9)	dynamic_degree	Easy	-0.17 ($p < 0.01$)
Kinematic Coupling (Q9)	dynamic_degree	Medium	-0.26 ($p < 0.01$)
Kinematic Coupling (Q9)	dynamic_degree	Hard	0.02 ($p > 0.1$, n.s.)

Table 8: Spearman correlation between human evaluation and WorldModelBench aggregate score, by complexity tier, using either the full 12-dimension human composite or a motion-specific composite (stationary rigidity, correct part motion, kinematic coupling).

Tier	N	ρ (Overall)	ρ (Motion-specific)	95% CI (Overall)
Easy	300	0.13 ($p < 0.1$)	0.09 ($p > 0.1$, n.s.)	[-0.01, 0.29]
Medium	120	0.34 ($p < 0.001$)	0.28 ($p < 0.01$)	[0.20, 0.48]
Hard	96	0.50 ($p < 0.0001$)	0.35 ($p < 0.001$)	[0.39, 0.59]