PhysVista: Benchmarking Physical Intelligence in VLMs via a Perception-Reasoning-Assessment Loop

Xinge Peng[1] **Yiting Lu**[1] **Tianwu Zhi**[2] **Wen Wen**[3] **Jianzhao Liu**[2] **Xin Li**[1] * **Zhibo Chen**[1] *
[1]University of Science and Technology of China
[2]ByteDance
[3]City University of Hong Kong
xg.pengv@gmail.com

Project **Dataset**

Abstract

Vision-Language Models (VLMs) have shown strong multimodal reasoning capabilities, yet whether they truly capture the physical consistency underlying real-world dynamics remains unclear. Existing benchmark paradigms often suffer from fragmented evaluation, focusing on isolated cognitive stages while overlooking the inherent synergy between perception, reasoning, and physical judgment. The lack of a holistic perspective limits the ability to diagnose whether VLMs can reliably serve as evaluators of physical plausibility in generative content. To address these issues, we introduce **PhysVista**, a benchmark designed to evaluate physical intelligence in VLMs through a closed cognitive loop framework inspired by the human **seeing–reasoning–assessment** process. PhysVista restores this loop by jointly evaluating physical state perception, physical dynamics reasoning, and physical plausibility assessment. It further distinguishes **event-level** reasoning and **scale-level** reasoning to enable fine-grained analysis of physical understanding. In addition, PhysVista incorporates both real-world and AI-generated videos, allowing evaluation across diverse domains and emerging generative scenarios. Extensive experiments across a diverse set of VLMs reveal substantial limitations in physical reasoning and plausibility assessment, highlighting a persistent gap between visual recognition and genuine physical understanding, and pointing toward more principled designs for physically grounded multimodal intelligence.

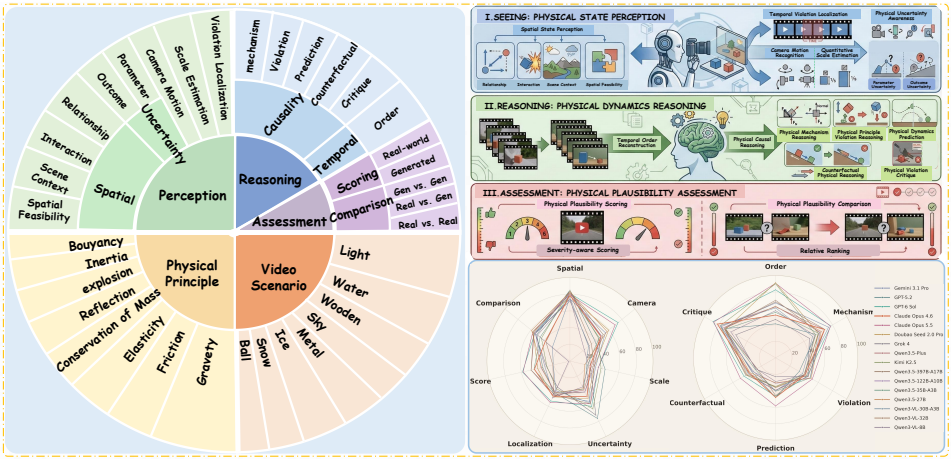


Figure 1: Overview of PhysVista. PhysVista evaluates physical intelligence in VLMs through *Seeing*, *Reasoning*, and *Assessment* across real and generated videos.

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Vision-Language Models (VLMs) [24, 46, 11, 41, 1] have achieved remarkable progress in visual understanding, reasoning, and multimodal interaction. Driven by their strong performance, some works have increasingly explored VLMs for *visual quality assessment* [54, 50, 28, 35], with recent efforts going beyond *appearance-level evaluation* to judge the *physical plausibility* of content produced by generative models [29, 31]. However, the reliability of such physical judgments remains unclear, raising a fundamental question: do VLMs truly understand the underlying physical dynamics of the world? Recent studies [30, 45, 39] reveal that modern models frequently produce physically inconsistent interpretations, suggesting that strong visual recognition does not necessarily translate into genuine physical understanding. However, the extent of this limitation remains unclear due to the lack of systematic evaluation of physical intelligence in VLMs.

Physical understanding inherently follows a closed cognitive loop. Humans interpret physical events by first perceiving observable states, then reasoning about the underlying causal mechanisms, and finally assessing whether the observed dynamics conform to physical laws. This *seeing–reasoning–assessment* loop forms the foundation of physical intelligence, enabling consistent interpretation, prediction, and validation of dynamic environments. However, existing benchmarks evaluate physical understanding only at fragmented stages of this cognitive loop. PhysBench [17] primarily focuses on perception-level evaluation. In terms of reasoning, most physical reasoning benchmarks [19, 37, 27, 48] cover only a limited range of tasks and evaluation capabilities. Moreover, they mainly evaluate reasoning at the *event-level*, focusing on whether an event occurs, rather than *scale-level* reasoning that requires distinguishing fine-grained differences in magnitude, quantity, precise position, or proportion. The most recent work, PAI-Bench [56], evaluates both perceptual and reasoning aspects of physical understanding. However, it still lacks the final component required to complete the closed loop—*assessment*. A comparison of task types is illustrated in Tab. 1. As a result, current benchmarks measure isolated manifestations of physical intelligence rather than holistic physical cognition.

Beyond limited task coverage, the evaluation paradigm itself remains problematic. PhysBench [17] suffers from language bias that introduces shortcuts, allowing models to answer questions in a video-blind manner. IntPhys2 [14] formulates evaluation as binary plausibility judgments of events, which mainly tests shallow physical intuition rather than deeper causal reasoning. Other benchmarks [30, 45, 39] focus on applying physical laws in structured problem-solving settings, where visual information serves only as auxiliary cues. This setup is fundamentally misaligned with the real-world physical understanding, which is perception-driven and must be grounded in visual observations. Consequently, current benchmarks fall short of evaluating comprehensive physical understanding in a reliable, in-depth way.

Additionally, the lack of domain diversity in data source also limits the scope of evaluation. Early benchmarks [26, 37, 49, 42, 9, 13, 53, 36] are predominantly simulation-based, with limited evaluation on real-world videos. However, the simplicity of simulated environments limits their ability to faithfully reflect a model's capability in reasoning about complex physical dynamics in real-world scenarios. Recent works [19, 56, 23] have begun incorporating real-world videos into evaluation. Meanwhile, the rapid emergence of generative video models [51, 16] and the widespread use of AI-generated videos make the evaluation of generated content increasingly important, yet this aspect remains largely unexplored. Generated videos often exhibit physical implausibilities arising from complex scene changes, object interactions, and human motions. However, the gap in VLM physical understanding between real-world and generated videos remains unclear. Moreover, such content introduces new evaluation requirements, including tasks such as Physical Violation Critique and Physical Plausibility Scoring.

To address these challenges, we introduce **PhysVista**, a benchmark for evaluating physical intelligence in Vision-Language Models through a unified closed-loop framework. Rather than emphasizing new data collection, PhysVista focuses on principled task formulation, systematic re-annotation, and unified evaluation for physical understanding. It systematically measures physical state perception, temporal and causal reasoning, and plausibility assessment, enabling holistic diagnosis beyond isolated capability testing. It also includes both real-world and AI-generated videos. Beyond task and data diversity, we further design an elaborate evaluation framework that assesses reasoning at both the event and scale levels. Our main contributions are as follows:

- **PhysVista Benchmark.** We introduce **PhysVista**, a benchmark for evaluating physical intelligence in Vision-Language Models that restores the *seeing–reasoning–assessment* cognitive

		Perception					Reasoning		Assessment	
Benchmark	**Data Source**	**Spatial**	**Localization**	**Camera**	**Scale**	**Uncertainty**	**Order**	**Causality**	**Scoring**	**Comparison**
IntPhys2 [14]	Simulation	✗	✗	✗	✗	✗	✗	◐	✗	✗
PhysBench [17]	Real+Simulation	✓	✗	◐	✗	✗	✓	◐	✗	✗
CausalVQA [19]	Real	◐	✗	✗	✗	✗	✗	◐	✗	✗
MVP Bench [27]	Real+Embodied+Simulation	◐	✗	✗	✗	✗	✗	◐	✗	✗
PAI-Bench [56]	Real+Embodied	✓	✗	◐	✗	✗	✓	◐	✗	✗
PhysVista (Ours)	Real + Generated Videos	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison with related physical understanding benchmarks. PhysVista provides unified evaluation across perception, reasoning, and assessment, and explicitly includes generated-video scenarios.

loop. It covers perception, reasoning, and plausibility evaluation using both real-world and AI-generated videos.

- **Versatile Evaluation Framework.** We propose a versatile evaluation framework that measures physical state perception, temporal and causal reasoning, and plausibility assessment. It further distinguishes *event-level* and *scale-level* reasoning to enable fine-grained evaluation of physical understanding.
- **Key Insights into VLM Physical Intelligence.** We provide detailed performance analysis and uncover important observations that highlight the limitations of current VLMs, as well as inform the design of future physically grounded models.

2 Related Work

2.1 Physics-related Benchmarks for Generative Models.

While traditional metrics (*e.g.*, FVD [20]) and multi-dimensional benchmarks [25, 55, 40, 18] assess visual and fine-grained attributes, they neglect long-term physical modeling. To address this, recent VLM-based benchmarks explicitly evaluate physical dynamics and causality: PhyGenBench [31] and WorldBench [43] analyze rule decomposition, PhyWorldBench [22] tests anti-physics scenarios, and VideoVerse [47] assesses causal QA. For closed-loop decision-making, DrivingGen [57] tests driving trajectory safety, WorldArena [38] simulates environments, and 4DWorldBench [29] evaluates 3D/4D spatiotemporal consistency. However, these methods often output single scalars without structural decomposition. Relying solely on final generations without continuous state supervision also obscures long-term error propagation, highlighting the critical need for systematic, interactive evaluation frameworks.

2.2 Physics-related Benchmarks for VLMs.

Early benchmarks [26, 37, 49, 42, 9, 13, 53, 36, 14] adopt simulation-based environments to assess intuitive physics through violation-of-expectation paradigms, they often lack the visual complexity and distributional diversity of real-world or model-generated videos. PhysBench [17] emphasizes physical perception by covering object properties, spatial relations, scene dynamics, and motion understanding, but suffers from language shortcut. To overcome this, MVP Bench [27] construct controlled minimal video pairs to reduce shortcut learning and evaluate models' sensitivity to physically inconsistent events. Another line of works primarily evaluate the causal reasoning ability in real-world videos. CausalVQA [19] and CoPhyBench [48] focus on predicting outcomes and reasoning under interventions or conditional observations, probing models' ability to infer causal physical mechanisms. Other physical benchmarks [30, 45, 39] focus on structured physical problem-solving in exam-style settings, which fails to reflect open-world physical understanding. There are also some application-oriented evaluation benchmarks [56, 23] targets for assessing physical understanding particularly in embodied AI and robotics scenarios. More recently, ExeWorldBench [34] has integrated physical reasoning with 3D coding. Overall, existing benchmarks focus on isolated aspects of physical understanding while largely neglecting evaluation on AI-generated content.

3 PhysVista

3.1 Overview

As shown in Fig. 1, PhysVista is a benchmark designed to evaluate physical intelligence in Vision-Language Models through a closed-loop cognitive framework consisting of three stages: Physical State Perception (Seeing), Physical Causal Reasoning (Reasoning), and Physical Plausibility Assessment (Assessment). Unlike prior benchmarks that evaluate isolated physical capabilities, PhysVista systematically measures whether models can reconstruct observable physical states, infer causal mechanisms, and assess global physical Plausibility with scores and ranks. The benchmark incorporates both real-world and AI-generated videos containing diverse object interactions, motion dynamics, and varying degrees of physical consistency, enabling holistic evaluation of physical cognition beyond event-level detection. More detailed benchmark statistics are provided in Appendix.

3.2 Seeing: Physical State Perception

3.2.1 Task Information.

The Physical State Perception stage evaluates whether models can accurately recover observable physical states directly from visual inputs. We consider five complementary aspects: spatial state perception, temporal state perception, camera motion recognition, quantitative scale estimation, and physical uncertainty awareness.

SPATIAL STATE PERCEPTION. We decompose spatial perception into four subtypes reflecting different dimensions of spatial cognition.

- **RELATIONSHIP.** Evaluates whether models can correctly identify relative spatial configurations among objects, including direction and distance. (*e.g.*, left/right, near/far, containment, support).
- **INTERACTION.** Examines recognition of physically meaningful interactions between entities, such as contact, collision.
- **SCENE CONTEXT.** Assesses whether models can perceive environmental conditions that influence physical interpretation, including terrain type, weather, or surface properties observable from the scene.
- **SPATIAL FEASIBILITY.** Measures the ability to judge whether spatial configurations are geometrically and physically realizable under real-world constraints (*e.g.*, size compatibility or object fitting).

TEMPORAL VIOLATION LOCALIZATION. This task requires models to identify the precise time interval during which a physically inconsistent event occurs. Unlike global video-level plausibility detection in prior works, this task requires second-level fine-grained temporal grounding of specific physical violations.

CAMERA MOTION RECOGNITION. Although prior works have explored this task, they lack a systematic taxonomy of camera motions and cover only limited categories, leaving camera motion recognition largely underexplored. The model is required to classify camera motion into ten predefined categories given a video.

QUANTITATIVE SCALE ESTIMATION. While existing benchmarks rely on coarse descriptive adjectives to assess relative scale perception, we reformulate the task in a quantitative manner by requiring explicit relative ratio estimation between objects and express responses as fractional values (*e.g.*, $1/3$), enabling fine-grained and numerically grounded physical perception.

PHYSICAL UNCERTAINTY AWARENESS. We incorporate uncertainty awareness as a fundamental dimension of physical perception, which remains largely neglected in existing benchmarks. We define two complementary forms of uncertainty:

- **PARAMETER UNCERTAINTY AWARENESS.** The ability to recognize when latent physical parameters (*e.g.*, friction coefficients or reflectance properties) cannot be determined from visual evidence alone.
- **OUTCOME UNCERTAINTY AWARENESS.** Assesses whether models correctly identify scenarios in which future physical outcomes are inherently indeterminate due to incomplete information or long-term, unobservable processes.

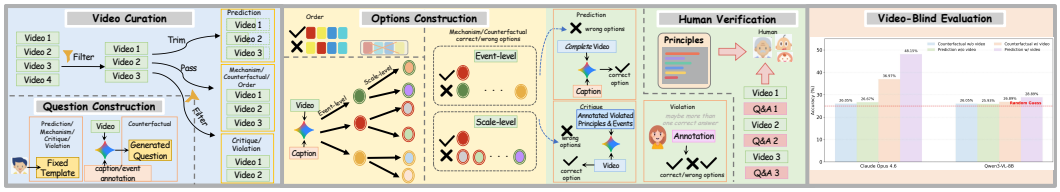


Figure 2: Data collection & evaluation of Physical Dynamics Reasoning.

Together, these tasks establish the perceptual foundation required for subsequent physical reasoning and assessment.

3.2.2 Data Collection.

Our data collection follows two stages: **video curation** and **question–answer construction**.

Stage I: Video curation. Except for **Camera Motion Recognition**, videos are sourced from the real-world dataset WISA-80K [44] and the video-generation dataset VideoPhy-2 [12], both covering diverse physical categories for physical understanding. For **Camera Motion Recognition**, we use Multi-CamVideo [10], which provides accurate camera pose trajectories. We select clips with a *single dominant* camera motion and no interfering movements to enable unambiguous motion categorization. To ensure sufficient task complexity, we additionally remove overly simple scenes that lack rich multi-object interactions or meaningful physical dynamics. All tasks are formulated as **single-answer multiple-choice** questions.

Stage II: Question–answer construction. For each task, we manually design a **single fixed template** (*e.g.*, **Spatial State Perception**, **Temporal Order Reconstruction**, and **Camera Motion Recognition**). Each template is task-specific but *content-agnostic* (no slot filling or instance-dependent variables), ensuring consistent structure and minimizing auxiliary textual cues.

Task-specific option construction. **Spatial State Perception:** options are generated from sampled frames and the original dataset captions. **Temporal Violation Localization:** human annotators label the start and end timestamps of intervals exhibiting physically implausible behavior. **Camera Motion Recognition:** we automatically assign one of ten predefined motion types using camera pose trajectories. Motions are first grouped into rotation-only (*e.g.*, pan/tilt) versus translation-with-rotation based on camera centers and viewing directions; the latter are further split into arc motions and pure translational motions by predefined geometric criteria. **Quantitative Scale Estimation:** human annotators provide explicit relative object ratios using fractional values as precise ground truth. **Physical Uncertainty Awareness:** for **Parameter Uncertainty**, we prompt Gemini 3.1 Pro [21] to generate questions about latent physical parameters; for **Outcome Uncertainty**, questions follow VideoPhy-2 indeterminate-outcome annotations, where the correct answer is deterministically *Indeterminate*.

Human verification. We conduct human verification for tasks whose ground truth is produced using VLMs (*e.g.*, **Spatial State Perception** and **Parameter Uncertainty**) to ensure data quality. Additionally, we also employ human annotators to rewrite samples exhibiting strong stylistic patterns, thereby reducing potential language bias.

3.3 Reasoning: Physical Dynamics Reasoning

3.3.1 Task Information.

The Physical Dynamics Reasoning stage evaluates whether models can reason about underlying physical processes beyond directly observable appearances. Unlike perception-level tasks that focus on state recognition, this stage requires inferring temporal dependencies and several causal mechanisms with physical principles. We decompose this stage into two complementary reasoning dimensions: temporal reasoning and causal reasoning, instantiated by Temporal Order Reconstruction and five causal reasoning tasks, respectively.

Temporal Order Reconstruction. Given temporally shuffled video frames sampled at irregular intervals, the model reconstructs a coherent physical event sequence. This task evaluates physical temporal reasoning by testing whether models can infer physically grounded temporal

dependencies in real-world videos while remaining robust to minor inconsistencies in generated content.

PHYSICAL CAUSAL REASONING. Beyond temporal dependencies, this stage evaluates comprehensive physical causal reasoning through five complementary subtasks:

- **PHYSICAL MECHANISM REASONING.** Requires inferring the causal physical mechanism governing the observed event, rather than providing a superficial description of its appearance.
- **PHYSICAL PRINCIPLE VIOLATION REASONING.** Requires determining all instances of violations of fundamental physical principles (*e.g.*, conservation laws, gravity) present in the video.
- **PHYSICAL DYNAMICS PREDICTION.** Assesses whether models can predict subsequent physical states or outcomes based on current observations and inferred dynamics, requiring event-level specificity and fine-grained scale-level distinctions (*e.g.*, stopping before versus exactly at a reference point).
- **COUNTERFACTUAL PHYSICAL REASONING.** Evaluates the ability to reason about hypothetical scenarios by modifying a key physical condition (*e.g.*, surface friction or applied force) and predicting the resulting outcome with event-level and scale-level precision.
- **PHYSICAL VIOLATION CRITIQUE.** Requires models to identify physically implausible events, analyze the underlying violation of real-world physical constraints, and provide a principled critique by suggesting modifications to the key physical factors that would render the scenario physically plausible.

3.3.2 Data Collection.

As illustrated in Fig. 2, our **Physical Causal Reasoning** data collection contains two stages: **video curation** and **question–answer construction**.

Stage I: Video curation. We first filter out static or low-dynamic clips and retain videos with rich physical motion and multi-object interactions. We then apply task-specific filtering to ensure unambiguous evaluation: (i) for **Physical Violation Critique**, we exclude videos containing multiple simultaneous principle violations; (ii) for **Physical Principle Violation Reasoning**, we remove intermediate or ambiguous violations to keep the violated principle clear and consistent; (iii) for **Physical Dynamics Prediction**, we only provide the early segment as input and withhold the final outcome event. All tasks are formulated as single-answer multiple-choice questions, except **Physical Principle Violation Reasoning**.

Stage II: Question–answer construction. For all tasks except **Counterfactual Physical Reasoning**, we use a **fixed, content-agnostic template** that is invariant across instances to minimize linguistic cues and video-blind answering. For **Counterfactual Physical Reasoning**, questions are centered on the primary physical event. We prompt Gemini 3.1 Pro to identify the main event using sampled frames and captions or event annotations from WISA-80K [44] and VideoPhy-2 [12], respectively. Counterfactual conditions are introduced *implicitly* by modifying event scenarios (*e.g.*, replacing an icy surface with a cobbled road) rather than explicitly stating parameter changes.

Option design principles. Answer generation is task-specific but follows four shared principles: (1) all options are physically plausible and consistent with basic intuition; (2) all options align with the video context; (3) the correct option has clearly superior causal explanatory power with minimal ambiguity; and (4) incorrect options are comparable to the correct one in either event-level structure or scale-level granularity.

Task-specific option construction. For **Temporal Order Reconstruction**, four frames are sampled at unequal temporal intervals, and distractors are constructed by swapping adjacent frames only. For **Physical Mechanism Reasoning** and **Counterfactual Physical Reasoning**, we adopt a hierarchical three-stage procedure: *(1) Event-level generation:* produce and rank four coarse-grained base options, selecting the top-ranked as the candidate correct answer; *(2) Scale-level refinement:* refine each base option into semantically similar variants with fine-grained distinctions and rank them within each base group; *(3) Selection:* use (a) *coarse-grained selection* by taking the top-ranked option from each base group to evaluate event-level reasoning, and (b) *fine-grained selection* by sampling multiple variants within the same base group (optionally alongside other groups) to test scale sensitivity. For **Physical Principle Violation Reasoning**, correct answers come from annotated violated principles,

while distractors are sampled from non-violated principles within the same context, potentially yielding multiple correct answers. For **Physical Dynamics Prediction**, distractors follow the same hierarchical strategy and the correct option is derived from the withheld subsequent segment. For **Physical Violation Critique**, we prompt Gemini 3.1 Pro to minimally edit the annotated violated principles and associated events such that the revised scenario becomes physically coherent; this annotation-grounded prompting constrains reasoning and mitigates hallucinations. Distractors include both event-level and scale-level alternatives.

Human verification. Finally, we conduct human verification to: (i) ensure data quality. (ii) Remove samples that can be answered correctly without viewing the video. The video-blind evaluation in Fig. 2 further demonstrates that the dataset is unbiased. (iii) Paraphrase samples with recognizable LLM-style phrasing to minimize potential language bias.

3.4 Assessment: Physical Plausibility Assessment

3.4.1 Task Information.

The Physical Plausibility Assessment stage measures a model's comprehensive understanding of physical plausibility by requiring both 5-level scoring for single videos and relative plausibility comparison between video pairs. Compared to perception and reasoning stages that focus on recognizing physical states or inferring causal mechanisms, the assessment stage emphasizes holistic evaluation of physical plausibility across the entire video, particularly for generated videos where subtle physical inconsistencies may arise, requiring models to calibrate their judgments to the *severity* of such inconsistencies.

PHYSICAL PLAUSIBILITY SCORING. Requires assigning a discrete plausibility rating from 1 to 5 to a single video by evaluating its overall physical coherence, with the score calibrated to the severity of physical inconsistencies rather than a binary valid/invalid decision.

PHYSICAL PLAUSIBILITY COMPARISON. Requires determining which video exhibits higher physical plausibility when comparing two samples, based on their overall physical coherence and realism. This task provides a relative judgment setting that is less sensitive to score-scale evaluation and directly tests whether the model can rank physical realism between samples.

3.4.2 Data Collection.

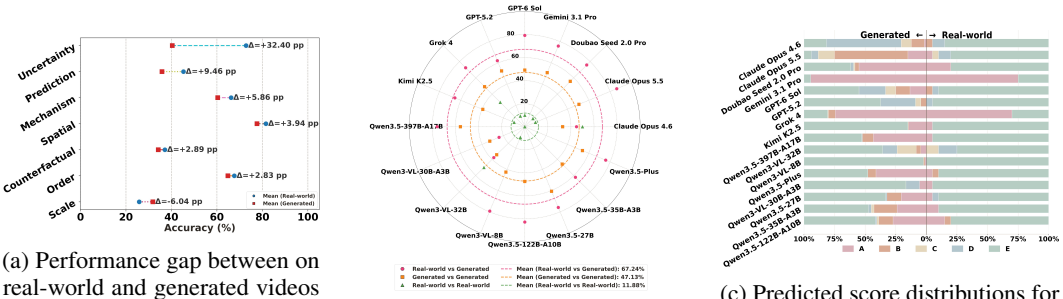
Our data collection consists of two stages: **video curation** and **question–answer construction**.

Stage I: Video curation. We draw generated videos from VideoPhy-2 [12] and real-world videos from WISA-80K [44]. VideoPhy-2 provides physical plausibility annotations with discrete scores ranging from 1 to 5. To ensure balanced coverage of different realism levels, we uniformly sample generated videos across all five score categories. In addition, real-world videos from WISA-80K are incorporated as physically consistent samples and assigned the highest plausibility score of 5.

Stage II: Question–answer construction. We adopt fixed templates of question and answer for both scoring and comparison tasks, and all answers are formatted as multiple-choice options. For the Physical Plausibility Comparison task, we construct three types of video pairs: real-world vs. generated, real-world vs. real-world, and generated vs. generated. In the generated–generated setting, which constitutes the most fine-grained comparison among three types, pairs are constructed to control for semantic similarity, and most pairs differ by no more than two plausibility levels, encouraging models to distinguish subtle differences in physical coherence.

4 Experiments

Following the design of PhysVista, experiments are organized into three stages corresponding to physical cognition: Physical State Perception (Seeing), Physical Causal Reasoning (Reasoning), and Physical Plausibility Assessment (Assessment). This evaluation protocol enables comprehensive diagnosis of model capabilities from observable state perception to causal inference and global physical judgment.



(a) Performance gap between on real-world and generated videos across tasks.

(b) Accuracy on three types of plausibility comparison.

(c) Predicted score distributions for plausibility scoring.

Figure 3: Model behavior analyses on PhysVista.

4.1 Overall Benchmark Results

As shown in Tab. 2, current VLMs perform relatively better on perception tasks than reasoning and assessment tasks, particularly *Spatial State Perception*, yet remain limited on fine-grained perceptual capabilities such as *Quantitative Scale Estimation* and *Physical Uncertainty Awareness*. Tab. 3 further indicates that *Physical Dynamics Reasoning* is the most challenging stage overall, with especially low accuracy on *Physical Dynamics Prediction* and *Counterfactual Physical Reasoning*. In contrast, results on *Physical Plausibility Assessment* show that models struggle with calibrated, fine-grained plausibility judgment and relative ranking, as evidenced by the low scores on both *Physical Plausibility Scoring* and *Physical Plausibility Comparison* in Tab. 2.

4.2 Cognitive Gap Analysis

Our benchmark is designed to reflect a *seeing–reasoning–assessment* loop. The results reveal a clear cognitive gap across the three stages of physical understanding, manifested as substantial *within-model inconsistencies in capability rankings*: models that perform strongly in observable state perception may rank much lower in causal reasoning, while those that excel at physical mechanism reasoning may perform less strongly in physical plausibility assessment.

Perception → Reasoning gap. While GPT-5.2 [32] and Kimi K2.5 [41] rank third and forth on *State Perception* (Tab. 2), respectively, their relative rankings drop substantially on *Dynamics Reasoning* (Tab. 3). This cross-stage ranking inconsistency suggests that strong performance in observable state perception does not necessarily translate into equally strong causal reasoning. Since reasoning about complex dynamic scenarios depends on both perceptual understanding and physical reasoning, these ranking shifts further suggest that the observed limitations cannot be attributed to perceptual capability alone.

Reasoning → Assessment gap. Assessment tasks require global plausibility judgments that are sensitive to severity, rather than mechanism explanations for a certain event. While Gemini 3.1 Pro [21] and Doubao Seed 2.0 Pro [15] achieve the third and forth rank on *Physical Dynamics Reasoning* (Tab. 3), their performance decreases when evaluated on *Physical Plausibility Assessment* (Tab. 2), demonstrating a clear capability gap between dynamics reasoning and plausibility assessment. This finding highlights a fundamental limitation of current VLMs: although they can sometimes correctly infer the causal mechanisms of violations, they struggle to calibrate physical plausibility along fine-grained levels.

4.3 Subtask Difficulty and Failure Modes in Reasoning

Tab. 3 highlights strong heterogeneity in reasoning difficulty.

Temporal reasoning is easier than causal reasoning. As for *Temporal Order Reconstruction*, models generally perform better on this task than on causal reasoning tasks, suggesting that temporal ordering can often be solved using basic physical commonsense with temporally coherent, whereas causal tasks require more complex reasoning with explicit physical modeling.

Prediction and Counterfactual reasoning remain difficult. *Physical Dynamics Prediction* and *Counterfactual Physical Reasoning* rely more heavily on temporal information analysis than other causal reasoning tasks, as models must infer how the observed dynamics evolve over time rather than

Table 2: Overall performance on Physical State Perception and Physical Plausibility Assessment tasks. Localization uses mean IoU, scoring uses SRCC/PLCC, and other tasks use accuracy. Avg. Acc. is task-count weighted. Best and second-best results are highlighted in brown and light brown, respectively.

Model	Spatial	Camera	Scale	Uncertainty	Avg. Acc.	Localization	Score	Comparison
GPT-6 Sol [33]	81.00%	74.00%	35.00%	60.00%	68.59%	53.03%	0.6141 / 0.5586	54.00%
GPT-5.2 [32]	85.00%	38.50%	43.33%	76.11%	64.06%	42.09%	0.3328/0.3423	51.00%
Claude Opus 5.5 [8]	79.50%	69.50%	35.00%	56.67%	65.78%	54.87%	0.5702 / 0.5882	57.00%
Claude Opus 4.6 [7]	79.00%	47.50%	25.00%	47.22%	55.15%	46.63%	0.4482/0.4320	38.00%
Gemini 3.1 pro [21]	74.50%	54.50%	31.67%	53.33%	58.28%	45.34%	0.3087/0.3087	53.00%
Doubao Seed 2.0 Pro [15]	77.50%	59.50%	30.00%	50.00%	59.69%	47.48%	0.4018/0.3949	57.00%
Grok 4 [52]	72.00%	29.50%	25.00%	71.67%	54.22%	34.61%	0.2853/0.2952	54.00%
Qwen3.5-Plus [6]	82.00%	48.00%	26.67%	38.33%	53.91%	45.79%	0.4809/0.4790	55.00%
Kimi K2.5 [41]	84.50%	44.00%	38.33%	63.33%	61.56%	36.92%	0.2852/0.2849	44.00%
Qwen3.5-397B-A17B [5]	83.50%	51.00%	28.33%	41.11%	56.25%	43.92%	0.4817/0.4788	56.00%
Qwen3.5-122B-A10B [2]	81.00%	48.50%	30.00%	41.11%	54.84%	41.05%	0.3593/0.3560	53.00%
Qwen3.5-35B-A3B [4]	79.50%	45.00%	21.67%	37.22%	51.41%	42.25%	0.4462/0.4440	46.00%
Qwen3.5-27B [3]	82.50%	57.00%	18.33%	36.67%	55.63%	42.21%	0.3505/0.3521	59.00%
Qwen3-VL-30B-A3B [11]	76.00%	28.50%	20.00%	55.56%	50.16%	39.09%	0.1945/0.2189	27.00%
Qwen3-VL-32B [11]	81.00%	33.00%	30.00%	56.67%	54.38%	44.61%	0.2105/0.2125	37.00%
Qwen3-VL-8B [11]	75.00%	22.00%	23.33%	34.44%	42.19%	48.83%	0.0240/0.0071	50.00%

Table 3: Performance on Physical Dynamics Reasoning. Violation is evaluated by Micro-F1, others are evaluated by accuracy. Avg. Acc. is task-count weighted. Best and second-best results are highlighted in blue and light blue, respectively.

Model	Order	Mechanism	Violation	Prediction	Counterfactual	Critique	Avg. Acc.
GPT-6 Sol [33]	82.67%	77.60%	44.48%	46.67%	44.54%	88.00%	62.68%
GPT-5.2 [32]	41.33%	56.77%	33.22%	48.15%	33.61%	71.00%	49.86%
Claude Opus 5.5 [8]	90.67%	77.08%	51.60%	58.52%	49.58%	87.00%	67.89%
Claude Opus 4.6 [7]	52.00%	68.75%	41.47%	48.15%	36.97%	82.00%	57.61%
Gemini 3.1 pro [21]	66.67%	74.48%	40.94%	40.00%	38.66%	76.00%	60.20%
Doubao Seed 2.0 Pro [15]	91.33%	59.90%	42.09%	35.56%	40.34%	70.00%	60.06%
Grok 4 [52]	46.67%	47.40%	45.59%	22.96%	27.73%	56.00%	40.37%
Qwen3.5-Plus [6]	62.00%	68.75%	45.71%	39.26%	41.18%	78.00%	58.19%
Kimi K2.5 [41]	90.00%	56.25%	40.17%	34.81%	42.02%	78.00%	60.06%
Qwen3.5-397B-A17B [5]	68.00%	69.27%	43.40%	36.30%	37.82%	73.00%	57.76%
Qwen3.5-122B-A10B [2]	62.67%	62.50%	48.85%	31.11%	32.77%	73.00%	52.87%
Qwen3.5-35B-A3B [4]	55.33%	59.90%	48.64%	40.74%	28.57%	70.00%	51.29%
Qwen3.5-27B [3]	62.67%	64.06%	49.82%	30.37%	36.13%	72.00%	53.59%
Qwen3-VL-30B-A3B [11]	60.67%	52.08%	37.98%	40.00%	31.93%	78.00%	51.87%
Qwen3-VL-32B [11]	70.00%	61.46%	40.20%	45.93%	26.05%	71.00%	55.60%
Qwen3-VL-8B [11]	57.33%	54.69%	41.31%	28.89%	26.89%	63.00%	46.70%

only identifying causal mechanisms. This makes it difficult for current models to simultaneously maintain event-level correctness and fine-grained scale-level precision. Although both tasks involve temporal reasoning and causal understanding, they differ in their reasoning requirements: *Physical Dynamics Prediction* can often be solved through temporal extrapolation combined with causal reasoning about the observed motion, whereas *Counterfactual Physical Reasoning* additionally requires reasoning under hypothetical interventions and implicitly simulating the resulting dynamics. This difference in reasoning demands likely accounts for the performance discrepancy between the two tasks within the same model.

4.4 Real-world vs. Generated Content Analysis

To analyze model behavior across data sources, we compare performance on real-world and generated videos across multiple tasks (Fig. 3a). Overall, models perform better on real-world videos, revealing a systematic generalization gap.

Generated content amplifies subtle failure modes. Across most tasks, accuracy on real-world videos is higher than on generated videos. While basic spatial perception and temporal reasoning remain relatively stable, showing only small decreases on generated content, the performance gap becomes much larger for *Physical Uncertainty Awareness* and *Physical Dynamics Prediction*. These tasks require abstract perception or fine-grained physical judgment, making them particularly sensitive to subtle physical inconsistencies that often appear in generated videos. This suggests that current models rely heavily on visually and physically consistent patterns commonly observed in real-world data. When subtle violations appear, models often fail to internally reconcile these inconsistencies during the reasoning process, revealing a lack of deeper physical reasoning capability.

Not all tasks become harder on generated videos. An interesting exception appears in *Quantitative Scale Estimation*, where performance on generated videos slightly exceeds that on real-world videos. One possible explanation is that generated videos may present cleaner geometric layouts or more controlled object configurations, making relative scale relationships easier to infer compared to real-world scenes with cluttered backgrounds and measurement ambiguity.

4.5 Assessment Task Analysis

Physical Plausibility Assessment consists of two tasks: *Physical Plausibility Scoring* (single-video, 1–5 rating) and *Physical Plausibility Comparison* (pairwise ranking). We analyze models' scoring calibration and distribution patterns in the scoring task, as well as their decision behavior across different pair types in the comparison task.

Physical Plausibility Scoring: Calibration vs. Score Collapse Fig. 3c visualizes the distribution of predicted scores for generated and real-world videos. We observe the critical failure mode: **1) Score collapse in plausibility scoring.** We observe a consistent *score collapse* phenomenon in several models (*e.g.*, Qwen3-VL-8B and Kimi K2.5), where a disproportionately large fraction of predictions concentrate on the highest score level for *both* generated and real-world videos. This phenomenon suggests that many models are unable to effectively differentiate between varying levels of physical plausibility, but instead default to a dominant score when clear physical violations are not confidently identified. Consequently, correct predictions on real-world videos may arise merely because these samples are labeled with the highest score—the default prediction for many models—rather than reflecting accurate plausibility estimation. Notably, the collapse is model-dependent: some model shift the default score to the lowest level (*e.g.*, Gemini 3.1 Pro and Grok 4), performance on real-world videos deteriorates dramatically. This difference likely stems from variations in their inherent priors toward certain score tokens. **2) Competent scoring requires distributional separation.** Models that exhibit genuine scoring ability show distinct distributions between generated and real-world videos: real-world predictions concentrate on high plausibility consistently, while generated predictions spread across lower levels. This separation indicates sensitivity to graded realism and better calibration to violation severity.

Physical Plausibility Comparison: Varying Difficulty Across Pair Types The comparison task includes three pair types: *Real-world vs. Generated*, *Generated vs. Generated*, and *Real-world vs. Real-world*, with results shown in Fig. 3b. **1) Real-world vs. Generated is easiest.** This setting exhibits clear physical contrast, leading to the highest accuracy and mainly testing the ability to reject obvious implausibility. **2) Generated vs. Generated is substantially harder.** Accuracy drops when comparing two generated videos with similar semantics, requiring fine-grained evaluation of physical coherence and ranking of plausibility among imperfect samples. **3) Real-world vs. Real-world exposes over-confident preference.** When both videos are physically valid, models still tend to force a choice, leading to low accuracy and indicating limited ability to recognize equivalence. This highlights the need for models to explicitly handle the "equal" case.

5 Conclusion

In this work, we introduce PhysVista, a benchmark for evaluating physical intelligence in Vision-Language Models through a closed cognitive loop of Physical State Perception, Physical Dynamics Reasoning, and Physical Plausibility Assessment. Unlike prior benchmarks that focus on isolated abilities, PhysVista provides a unified evaluation across real-world and AI-generated videos. Extensive experiments show that although modern VLMs exhibit strong semantic perception, their physical intelligence remains incomplete across the full cognitive loop. Nevertheless, PhysVista still has limitations in evaluation granularity and the depth and breadth of its question coverage, leaving room for more fine-grained and comprehensive evaluations of physical intelligence. We hope this diagnostic framework will facilitate future research on physically grounded vision-language systems.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 62371434 and U25B2010, the Fundamental Research Funds for the Central Universities (No. WK2100250064), and ZGCA Project C20250302.

References

[1] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[2] Alibaba. Qwen3.5-122b-a10b. `https://qwen.ai/apiplatform`, 2026. Accessed: 2026-03-01.

[3] Alibaba. Qwen3.5-27b. `https://qwen.ai/apiplatform`, 2026. Accessed: 2026-03-01.

[4] Alibaba. Qwen3.5-35b-a3b. `https://qwen.ai/apiplatform`, 2026. Accessed: 2026-03-01.

[5] Alibaba. Qwen3.5-397b-a17b. `https://qwen.ai/apiplatform`, 2026. Accessed: 2026-03-01.

[6] Alibaba. Qwen3.5-plus. `https://qwen.ai/apiplatform`, 2026. Accessed: 2026-03-01.

[7] Anthropic. Claude opus 4.6. `https://www.anthropic.com/news/claude-opus-4-6`, 2026. Accessed: 2026-03-01.

[8] Anthropic. Claude opus 5.5. `https://www.anthropic.com/claude-opus-5-5`, Sept. 2026. Released September 22, 2026.

[9] T. Ates, M. Ateşoğlu, Ç. Yiğit, I. Kesen, M. Kobas, E. Erdem, A. Erdem, T. Goksun, and D. Yuret. Craft: A benchmark for causal reasoning about forces and interactions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2602–2627, 2022.

[10] J. Bai, M. Xia, X. Fu, X. Wang, L. Mu, J. Cao, Z. Liu, H. Hu, X. Bai, P. Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025.

[11] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[12] H. Bansal, C. Peng, Y. Bitton, R. Goldenberg, A. Grover, and K.-W. Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

[13] F. Baradel, N. Neverova, J. Mille, G. Mori, and C. Wolf. Cophy: Counterfactual learning of physical dynamics. *arXiv preprint arXiv:1909.12000*, 2019.

[14] F. Bordes, Q. Garrido, J. T. Kao, A. Williams, M. Rabbat, and E. Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

[15] ByteDance. Seed2.0 pro. `https://seed.bytedance.com/en/seed2`, 2026. Accessed: 2026-03-01.

[16] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7310–7320, 2024.

[17] W. Chow, J. Mao, B. Li, D. Seita, V. Guizilini, and Y. Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

[18] H. Duan, H.-X. Yu, S. Chen, L. Fei-Fei, and J. Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27713–27724, 2025.

[19] A. Foss, C. Evans, S. Mitts, K. Sinha, A. Rizvi, and J. T. Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025.

[20] S. Ge, A. Mahapatra, G. Parmar, J.-Y. Zhu, and J.-B. Huang. On the content bias in fréchet video distance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7277–7288, 2024.

[21] Google. Gemini 3.1 pro. `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview`, 2026. Accessed: 2026-03-01.

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

[23] A. Gundawar, S. Sagar, and R. Senanayake. Pac bench: Do foundation models understand prerequisites for executing manipulation policies? *arXiv preprint arXiv:2506.23725*, 2025.

[24] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

[26] S. Jassim, M. Holubar, A. Richter, C. Wolff, X. Ohmer, and E. Bruni. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint arXiv:2311.09048*, 2023.

[27] B. Krojer, M. Komeili, C. Ross, Q. Garrido, K. Sinha, N. Ballas, and M. Assran. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs. *arXiv preprint arXiv:2506.09987*, 2025.

[28] Y. Lu, F. Guan, Y. Gao, Y. Zhong, X. Peng, J. Yuan, Y. Liu, B. Zhang, X. Li, Z. Chen, et al. Omniquality-r: Advancing reward models through all-encompassing quality assessment. *arXiv preprint arXiv:2510.10609*, 2025.

[29] Y. Lu, W. Luo, P. Tu, H. Li, H. Zhu, Z. Yu, X. Wang, X. Chen, X. Peng, X. Li, et al. 4dworld-bench: A comprehensive evaluation framework for 3d/4d world generation models. *arXiv preprint arXiv:2511.19836*, 2025.

[30] C.-W. Mak, G. Zhu, B. Zhang, H. Li, X. Chi, K. Zhang, Y. Wu, Y. He, C.-K. Fan, W. Lu, et al. Physicsmind: Sim and real mechanics benchmarking for physical reasoning and prediction in foundational vlms and world models. *arXiv preprint arXiv:2601.16007*, 2026.

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

[32] OpenAI. Gpt-5.2. `https://developers.openai.com/api/docs`, 2026. Accessed: 2026-03-01.

[33] OpenAI. Gpt-6 sol. `https://developers.openai.com/api/docs/models/gpt-6-sol`, 2026. OpenAI API model documentation.

[34] X. Peng, Z. Fang, R. Wang, Y. Lv, H. Li, G. Zhang, X. Ma, Y. Lu, L. Chen, Z. Che, et al. Exeworldbench: Benchmarking code-based executable 3d world modeling. *arXiv preprint arXiv:2610.15166*, 2026.

[35] X. Peng, Y. Lu, X. Li, and Z. Chen. Iqa-spider: Unifying multi-granularity image quality assessment with reasoning, grounding and referring. *arXiv preprint arXiv:2605.24553*, 2026.

[36] N. F. Rajani, R. Zhang, Y. C. Tan, S. Zheng, J. Weiss, A. Vyas, A. Gupta, C. Xiong, R. Socher, and D. Radev. Esprit: Explaining solutions to physical reasoning tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7906–7917, 2020.

[37] R. Riochet, M. Y. Castro, M. Bernard, A. Lerer, R. Fergus, V. Izard, and E. Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018.

[38] Y. Shang, Z. Li, Y. Ma, W. Su, X. Jin, Z. Wang, L. Jin, X. Zhang, Y. Tang, H. Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

[39] H. Shen, T. Wu, Q. Han, Y. Hsieh, J. Wang, Y. Zhang, Y. Cheng, Z. Hao, Y. Ni, X. Wang, et al. Phyx: Does your model have the" wits" for physical reasoning? *arXiv preprint arXiv:2505.15929*, 2025.

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

[41] K. Team, T. Bai, Y. Bai, Y. Bao, S. Cai, Y. Cao, Y. Charles, H. Che, C. Chen, G. Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[42] H.-Y. Tung, M. Ding, Z. Chen, D. Bear, C. Gan, J. Tenenbaum, D. Yamins, J. Fan, and K. Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Advances in Neural Information Processing Systems*, 36:67048–67068, 2023.

[43] R. Upadhyay, H. Zhang, J. Solomon, A. Agrawal, P. Boreddy, S. S. Narayana, Y. Ba, A. Wong, C. M. de Melo, and A. Kadambi. Worldbench: Disambiguating physics for diagnostic evaluation of world models. *arXiv preprint arXiv:2601.21282*, 2026.

[44] J. Wang, A. Ma, K. Cao, J. Zheng, Z. Zhang, J. Feng, S. Liu, Y. Ma, B. Cheng, D. Leng, Y. Yin, and X. Liang. Wisa: World simulator assistant for physics-aware text-to-video generation, 2025.

[45] L. Wang, E. Su, J. Liu, P. Li, J. Xiao, W. Zhang, X. Chen, Y. Meng, L. BAI, W. Ouyang, et al. Physunibench: A multi-modal physics reasoning benchmark at undergraduate level.

[46] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[47] Z. Wang, X. Wei, B. Li, Z. Guo, J. Zhang, H. Wei, K. Wang, and L. Zhang. Videoverse: How far is your t2v generator from a world model? *arXiv preprint arXiv:2510.08398*, 2025.

[48] F. Wei, K. Xu, Y. Zhang, P. Sun, J. Xiao, and A. Yao. Cophybench: Benchmarking physical reasoning from conditional video observation.

[49] L. Weihs, A. Yuile, R. Baillargeon, C. Fisher, G. Marcus, R. Mottaghi, and A. Kembhavi. Benchmarking progress to infant-level physical reasoning in ai. *Transactions on Machine Learning Research*, 2022.

[50] W. Wen, T. Zhi, K. Fan, Y. Li, X. Peng, Y. Zhang, Y. Liao, J. Li, et al. Self-evolving vision-language models for image quality assessment via voting and ranking. In *International Conference on Learning Representations*, volume 2026, pages 39444–39457, 2026.

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

[52] xAI. Grok 4. `https://x.ai/api`, 2026. Accessed: 2026-03-01.

[53] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

[54] Z. Zhang, H. Wu, E. Zhang, G. Zhai, and W. Lin. Q-Bench++: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10404–10418, 2024.

[55] D. Zheng, Z. Huang, H. Liu, K. Zou, Y. He, F. Zhang, L. Gu, Y. Zhang, J. He, W.-S. Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

[56] F. Zhou, J. Huang, J. Li, D. Ramanan, and H. Shi. Pai-bench: A comprehensive benchmark for physical ai. *arXiv preprint arXiv:2512.01989*, 2025.

[57] Y. Zhou, H. Shao, L. Wang, Z. Zong, H. Li, and S. L. Waslander. Drivinggen: A comprehensive benchmark for generative video world models in autonomous driving. *arXiv preprint arXiv:2601.01528*, 2026.

A Technical appendices and supplementary material

A.1 Definition of Camera Motions

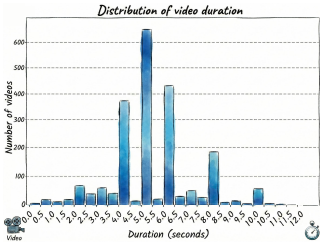
We incorporate 10 camera motion types in Camera Motion Recognition task, the definition of each type is illustrated in Tab. 4.

Table 4: Definition of Camera Motions.

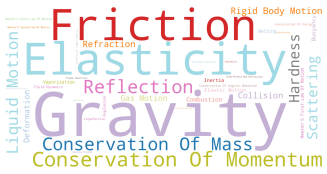
Camera Motion	Definition
Arc Left/Right	Horizontal circular movement around a subject, keeping it centered.
Pan Left/Right	Horizontal rotation of the camera on a fixed axis.
Tilt Up/Down	Vertical rotation of the camera on a fixed axis.
Translate Up/Down	Vertical displacement while adjusting pitch to maintain focus.
Static	The camera remains stationary with no movement or rotation.
Zoom Out	Decreasing the lens focal length to widen the field of view.

A.2 Benchmark Statistics

The statistics of video duration is illustrated in Fig. 4a. The word cloud of physical principles from the video is demonstrated in Fig. 4b. Fig. 4c shows a word cloud of the video captions, reflecting the content present in the scenes. We also shown the number of each tasks in Tab. 5 and Tab. 6.



(a) Distribution of video durations.



(b) Word cloud of physical principles evaluated in the benchmark.



(c) Word cloud of video captions summarizing the covered scenarios.

Figure 4: Data statistics of PhysVista.

Table 5: Number of samples for each task in Physical State Perception.

Spatial	Localization	Camera	Scale	Uncertainty
200	200	200	60	180

Table 6: Number of samples for each task in Physical Dynamics Reasoning and Physical Plausibility Assessment.

Order	Mechanism	Violation	Prediction	Counterfactual	Critique	Score	Comparison
150	192	224	135	119	100	200	100

A.3 Human Annotation, Verification, and Evaluation

In data generation stage, we incorporate six human annotators for data annotation and verification.

In data evaluation stage, we recruit five independent human evaluators to rate the 40% of the benchmark dataset on a scale from 1 to 5 in terms of linguistic neutrality, question-answer precision, and physical reasoning validity. The results are presented in Fig. 5, Fig. 6 and Fig. 7.

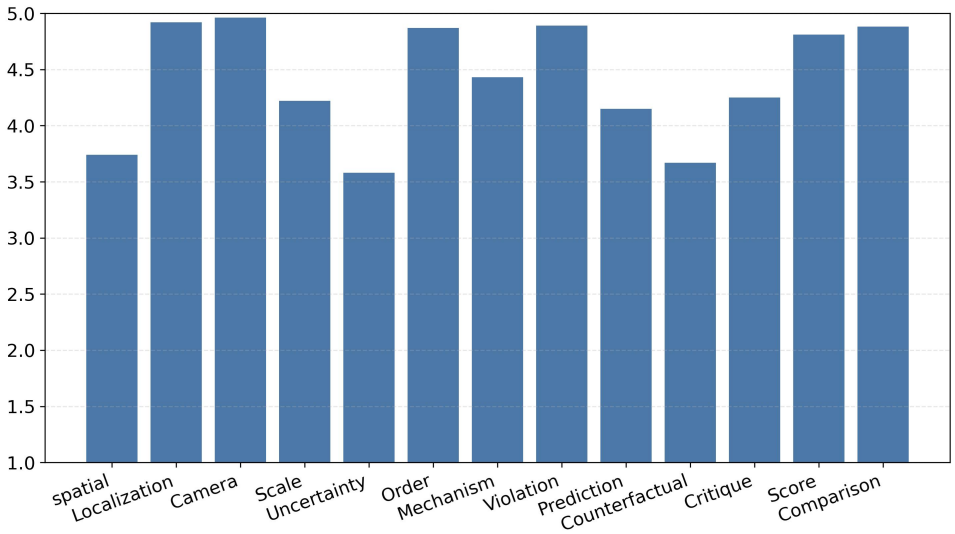


Figure 5: The results of linguistic neutrality evaluation.

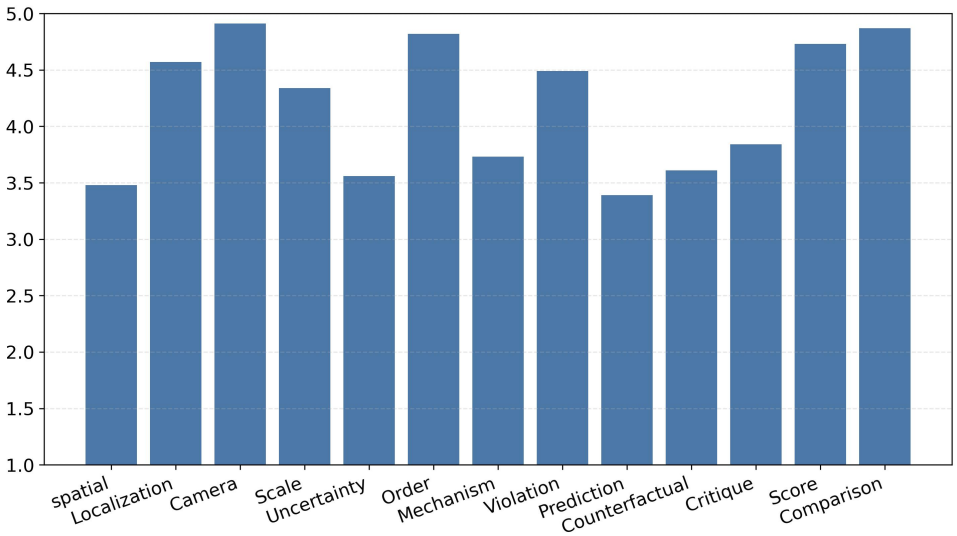


Figure 6: The results of question-answer precision evaluation.

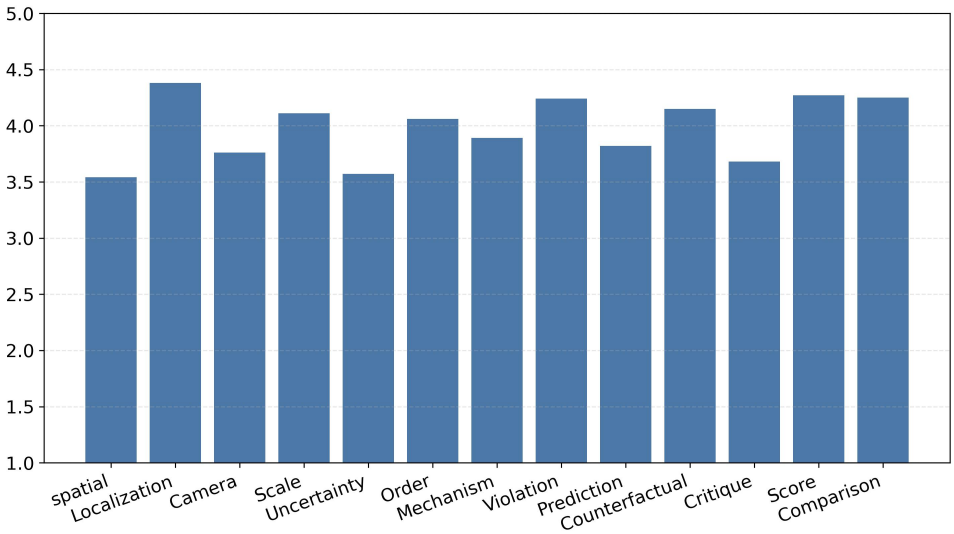


Figure 7: The results of physical reasoning validity evaluation.

A.4 More Results

For the Physical Principle Violation Reasoning task, we additionally report exact-match accuracy, option-level Micro-F1, and principle-type Macro-F1 in Tab. 7. Option-level Micro-F1 aggregates true positives, false positives, and false negatives over all A/B/C/D option predictions, measuring the overall multi-label prediction performance. Principle-type Macro-F1 first maps each selected option to its corresponding physical-principle category and then averages F1 scores across principle types, reflecting whether models perform consistently across different physical laws. This task requires mapping observed dynamics to principle-level constraints and often involves multiple simultaneous violations. Most models achieve relatively low performance on this task, suggesting that many VLMs lack reliable principle-grounded abstractions even when violations are visually salient.

Table 7: Performance on Physical Principle Violation Reasoning task. Best and second-best results are highlighted in red and light red, respectively.

Model	Exact-match Accuracy	Macro-F1	Micro-F1
GPT-6 Sol [33]	20.54%	46.60%	44.48%
GPT-5.2 [32]	14.73%	35.93%	33.22%
Claude Opus 5.5 [8]	15.63%	60.41%	51.60%
Claude Opus 4.6 [7]	12.05%	48.41%	41.47%
Gemini 3.1 pro [21]	15.62%	42.77%	40.94%
Doubao Seed 2.0 Pro [15]	14.29%	49.12%	42.09%
Grok 4 [52]	11.61%	50.03%	45.59%
Qwen3.5-Plus [6]	12.05%	47.21%	45.71%
Kimi K2.5 [41]	10.27%	40.98%	40.17%
Qwen3.5-397B-A17B [5]	7.14%	45.56%	43.40%
Qwen3.5-122B-A10B [2]	10.71%	53.79%	48.85%
Qwen3.5-35B-A3B [4]	10.71%	54.26%	48.64%
Qwen3.5-27B [3]	8.48%	58.59%	49.82%
Qwen3-VL-30B-A3B [11]	5.36%	34.90%	37.98%
Qwen3-VL-32B [11]	16.07%	41.16%	40.20%
Qwen3-VL-8B [11]	20.98%	31.66%	41.31%

Spatial State Perception: Relationship

As the yo-yo swings, which fixed object stays consistently to its left side in the camera view?

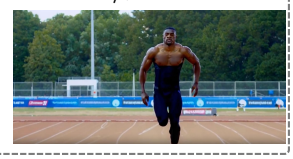
A. A wooden tree trunk
B. A park bench
C. A stone pathway
D. A metal fence



Spatial State Perception: Interaction

During the jump, which part of the athlete's body first crosses over the white track line on the ground?

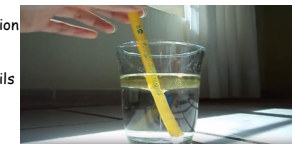
A. Head
B. Chest
C. Leading hand
D. Feet



Spatial State Perception: Scene Context

How does the submerged part of the pencil appear relative to the part above water?

A. Perfectly straight continuation
B. Bent at the water surface
C. Shorter but straight
D. Split into two separate pencils



Spatial State Perception: Spatial Feasibility

How are the black beads arranged when the hands finish adjusting them?

A. In one straight line
B. In two separated circles
C. In a compact circular cluster
D. Scattered randomly on the surface



Figure 8: Spatial State Perception Tasks.

Temporal Violation Localization

The given video frames are uniformly sampled from a video with a duration of {video_duration}. Identify the time interval where the physical implausibility occurs in the video. Output the answer in MM:SS-MM:SS format only.

00:03-00:04



Figure 9: Temporal Violation Localization Task.

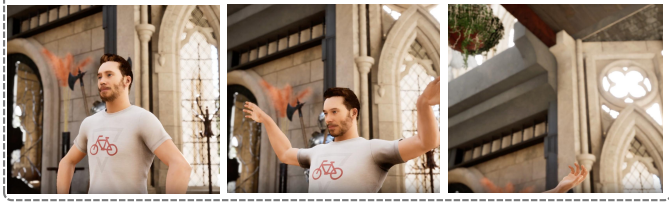
A.5 Task Illustrations

Our benchmark includes tasks forming a closed loop: Physical State Perception, Physical Dynamics Reasoning, and Physical Plausibility Assessment. Examples for these tasks are illustrated in Fig. 8, Fig. 9, Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, and Fig. 16.

Camera Motion Recognition

Which camera motion best describes the video?

A. Arc Left (with rotation)
B. Arc Right (with rotation)
C. Pan Left
D. Pan Right
E. Static
F. Tilt Down
G. Tilt Up
H. Translate Down (with rotation)
I. Translate Up (with rotation)
J. Zoom Out



Quantitative Scale Estimation

What is the length of the black square compared to the length of the red square in the video?

A. 3/4
B. 4/5
C. 7/9
D. 5/6

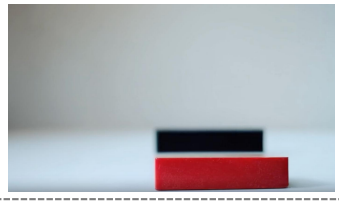
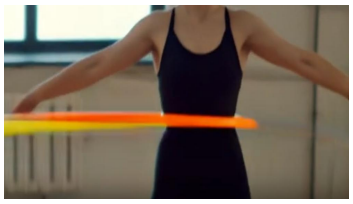


Figure 10: Camera Motion Recognition and Quantitative Scale Estimation Task.

Physical Uncertainty Awareness: Parameter Uncertainty Awareness

Is the coefficient of friction between hoop and clothing less than 0.3?
A. Yes
B. Indeterminate (not enough evidence)
C. No



Physical Uncertainty Awareness: Outcome Uncertainty Awareness

Does the man move upwards and downwards on the ladder?
A. Yes
B. No
C. Indeterminate (not enough evidence)



Figure 11: Physical Uncertainty Awareness Task.

Temporal Order Reconstruction

Arrange the shuffled video frames in the correct chronological order.
A. <frame 1> <frame 3> <frame 4> <frame 2>
B. <frame 1> <frame 4> <frame 2> <frame 3>
C. <frame 4> <frame 1> <frame 3> <frame 2>
D. <frame 1> <frame 4> <frame 3> <frame 2>



Physical Violation Critique

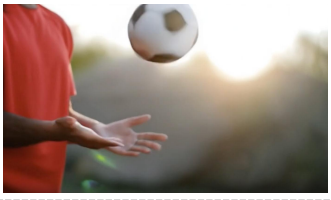
Which edit best fixes the physical implausibility with minimal change?
A. Replace the axe with a powered hydraulic splitter.
B. Make the wood log much drier and pre-cracked.
C. Increase the axe's downward speed just before impact.
D. Increase the axe's downward speed exactly at impact.



Figure 12: Temporal Order Reconstruction and Physical Violation Critique Task.

Physical Mechanism Reasoning

What is the best physical explanation for the video?
A. A cushion hidden below the frame pushes the ball upward whenever it falls below the players elbow.
B. Each time the ball drops to the players palm, the hand lifts it back up before it reaches the waistline.
C. A thin elastic cord tied to the ball pulls it upward whenever it stretches below the middle of the red shirt.
D. A second player, just outside the frame, taps the ball upward each time it falls past the players wrist.



Physical Principle Violation Reasoning

Which physical principles does the video violate?
A. Conservation of Momentum.
B. Friction.
C. Gravity.
D. Newton's First Law of Motion.



Figure 13: Physical Mechanism Reasoning and Physical Principle Violation Reasoning Task.

Physical Dynamics Prediction

What will happen in the unseen subsequent part of the video?
A. The surfer loses speed along the middle of the wave and sinks into the water before reaching the bright reflection on the right.
B. The surfer carves back toward the beach, sending a tall spray arc that reaches above his head and nearly touches the sunlit crest.
C. The surfer cuts deeper into the pocket and disappears behind the green curl without any spray arching over the top.
D. The surfer rides straight down the wave face and sinks in the white foam near the two swimmers closest to the shore.





Counterfactual Physical Reasoning

If the kayaker started paddling twice as frequently, what would happen to the kayak?
A. The kayak still reaches the wooden dock edge in the same time, stopping short of the concrete wall as before.
B. The kayak only gains speed for a moment, then matches its earlier pace and reaches the wooden dock edge no sooner than before.
C. The kayak reaches the wooden dock edge in about half the time, closing the gap to the concrete wall much faster.
D. The kayak takes longer to reach the wooden dock edge, staying near the yellow bow position seen in the first frame.

Figure 14: Physical Dynamics Prediction and Counterfactual Physical Reasoning Task.

Physical Plausibility Scoring: Generated Video

How well does the video follow physical commonsense? Select a score from 1 (very poor) to 5 (excellent).
A. 1
B. 2
C. 3
D. 4
E. 5



Physical Plausibility Scoring: Real-world Video

How well does the video follow physical commonsense? Select a score from 1 (very poor) to 5 (excellent).
A. 1
B. 2
C. 3
D. 4
E. 5



Figure 15: Physical Plausibility Scoring Task.

Physical Plausibility Comparison: Real-world vs. Generated

Compare Video 1 and Video 2, which one demonstrates higher physical commonsense quality?
A. Video 1
B. Video 2
C. Equal





Physical Plausibility Comparison: Generated vs. Generated

Compare Video 1 and Video 2, which one demonstrates higher physical commonsense quality?
A. Video 1
B. Video 2
C. Equal





Physical Plausibility Comparison: Real-world vs. Real-world

Compare Video 1 and Video 2, which one demonstrates higher physical commonsense quality?
A. Video 1
B. Video 2
C. Equal





Figure 16: Physical Plausibility Comparison Task.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our core contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included them in the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide complete details of the experiment in Section 4. Code and data are also released to facilitate reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have released all of the data and code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided the experimental details in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The error bars are commonly not reported in this domain.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have included the necessary information in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification:We have discussed the broader impact of our work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited references for the used models in the Section of experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have released the documentation with new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We have provided the information in Appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: Our study does not involve potential risks to participants, and all necessary information was disclosed to them. Given the nature of our research, IRB approval was not required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We have used LLM for evaluation and data generation.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.