PHYPROBE: Rethinking Physical Consistency Evaluation in Generated Videos

Max Ku[123*] **Jiaojiao Fan**[1] **Zekun Hao**[1] **Francesco Ferroni**[1]
Heng Wang[1] **Wenhu Chen**[23] **Ming-Yu Liu**[1] **Prithvijit Chattopadhyay**[1]

[1]**NVIDIA** [2]**University of Waterloo** [3]**Vector Institute**

Abstract

Evaluating the physical consistency of generated videos remains a fundamental challenge. Existing approaches rely on off-the-shelf vision-language models, which can often be myopic to physical dynamics, or fine-tuned evaluators trained on human annotations, which overfit to dataset-specific cues and fail to generalize. A key challenge is that existing supervision sources provide either relative ordering or absolute scores, but not both reliably and consistently across varied settings. To this end, we introduce PHYPROBE, an evaluator that extracts features from a frozen pretrained spatio-temporal encoder and maps them to a scalar physical consistency violation score via a lightweight scoring head. PHYPROBE is trained through a unified objective combining pairwise ranking, regression on noisy scalar annotations, and anchor-based calibration over a curated set of heterogeneous supervision sources. Experiments show that PHYPROBE outperforms prior methods on most pairwise benchmarks spanning real–generated and generated–generated pairs under varying correspondence, with the largest gains in no-correspondence and generated–generated settings where existing fine-tuned evaluators degrade sharply. PHYPROBE achieves strong correlation with human judgments, with close agreement between rank-based and linear metrics, indicating that scores are both well ordered and anchored to a stable [0, 1] scale. Further, despite being trained on supervision indicative of physical consistency, without explicit general-preference labels, PHYPROBE also performs competitively on human preference benchmarks: consistent with the observation that physics violations are entangled with broader quality degradations.

1 Introduction

Recent advances in video generation [58, 17, 52, 41] have made considerable progress across several dimensions: from producing videos of high visual quality [42, 35], temporally coherent motion [49, 32], to native multimodal support [48, 14, 18]. Yet one critical dimension remains stubbornly difficult to evaluate: **physical consistency**. A generated video can exhibit sharp textures and smooth motion while depicting objects that defy gravity, pass through solid surfaces, or spontaneously change form [2, 28]. But improving physical consistency requires first being able to measure it, and reliable measurement remains an open problem [47, 5]. The difficulty here is not in detecting visual quality degradations – often, there may not be any —- but in measuring how far the underlying dynamics in the generated video depart from physical reality. Crucially, simply identifying whether a video is physically plausible or not is insufficient. Not all physical violations are equal: an object drifting slightly off a surface is qualitatively different from an object exploding into fragments mid-air. A useful evaluator for *physical consistency* must quantify the "egregiousness" of a violation: producing a continuous, consistently scaled score that reflects severity, not just presence. This frames physical consistency evaluation as a measurement problem, and measurement demands precision, stability, and scale, as illustrated in Figure 1. Measurement must operate over a suitable representation of

*Part of the work done during internship at NVIDIA Cosmos Lab

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

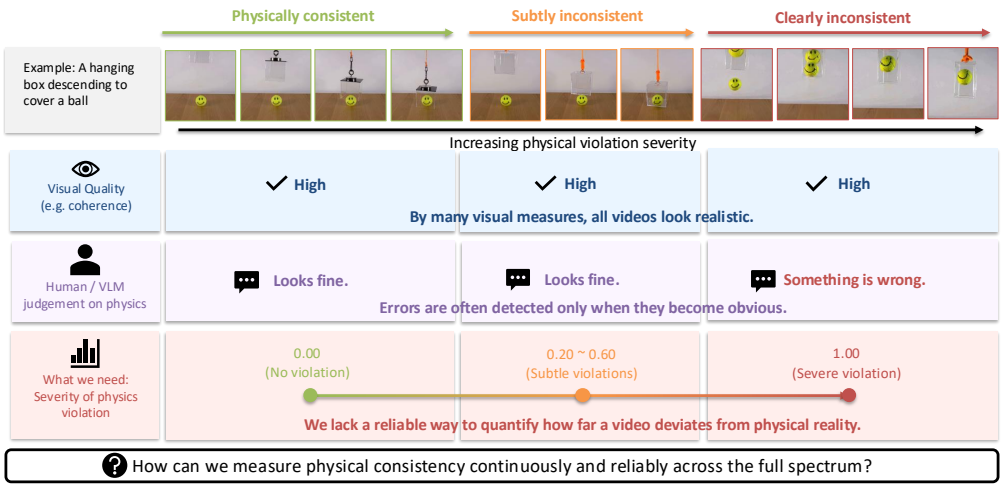


Figure 1: **Challenges in evaluating physical consistency.** Generated videos can be visually realistic across the board, while their physical consistency varies from fully consistent to severely violated. Human perception and current VLM-based evaluators can be thresholded and inconsistent, making it difficult to capture the continuous spectrum of violation severity. This motivates our goal: learning a bounded, continuous measure of physical consistency violation.

video dynamics; we argue that this representation should capture spatiotemporal structure directly rather than relying on semantic proxies such as language descriptions.

Existing approaches have made meaningful progress, but suffer from limitations in both evaluator design and supervision. Off-the-shelf VLMs [22, 26, 16] increasingly serve as evaluators through semantic reasoning, yet remain myopic to physical dynamics and require carefully crafted, context-dependent prompts to expose violations. Fine-tuned VLM evaluators [5, 6, 19, 20, 36] and detection-based approaches [54, 15, 44] improve performance through supervision on human annotations, but still operate largely on semantic proxy signals, relying on dataset-specific shortcuts rather than modeling underlying temporal dynamics. As a result, their performance often degrades under distribution shift. Even though human annotations remain the most reliable ground truth for physical consistency, they can be noisy at times: observers tend to detect only salient violations while absolute ratings drift as generative quality improves [39, 40, 38]. This motivates an evaluator that learns from such annotations without being limited by their precision. Moreover, physical inconsistencies often co-occur with other visual artifacts, causing annotators to respond to overall perceptual quality rather than physical dynamics specifically. Consequently, absolute ratings often reflect the prevailing data distribution at annotation time rather than a stable notion of physical consistency, and evaluators trained on such annotations inherit the resulting noise and drift (Details in Section 4).

Desiderata. These limitations motivate us to establish a clear desiderata of what a reliable physical consistency evaluator should achieve. (1) Reliable relative orderings: even if absolute scoring is imperfect, consistent "video A has more physical consistency violations than video B" judgments are sufficient for measurement. (2) Anchored scores on a fixed reference scale, so that absolute values remain comparable across datasets and time periods rather than drifting with the data distribution. (3) Operate on a single video at inference time, without requiring a paired reference or prompt alignment, since many physical violations are identifiable from temporal dynamics alone [59, 33].

Contributions. To this end, we introduce PHYPROBE, a video evaluator for physical consistency. Given a generated video, PHYPROBE extracts features from a frozen pretrained spatio-temporal encoder and maps them through a lightweight scoring head to a "violation score", $s(v)$, indicative of the degree of physical consistency violation. We find that the key to learning a meaningful scoring function like PHYPROBE lies in the following:

1. **Strong spatio-temporal representations.** We show that frozen pretrained spatio-temporal encoders [1, 8], trained through self-supervised or weakly-supervised objectives on large-scale video data, provide a foundation on which a lightweight scoring head can map video representations to a scalar violation score, without requiring VLMs or language-mediated reasoning.

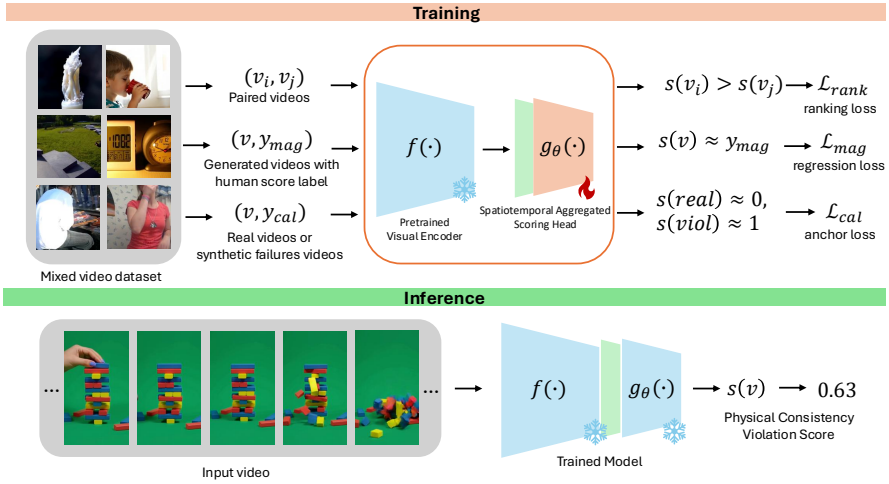


Figure 2: **Overview of PHYPROBE.** Given a video, we extract a learned physical representation using a pretrained spatiotemporal visual encoder, followed by a lightweight scoring head to predict a continuous violation score $s(v)$. Training combines three complementary supervision signals: (i) pairwise ranking on video pairs (v_i, v_j) to model relative violation, (ii) regression on human-annotated scores y_{mag}, and (iii) anchoring with calibration targets to enforce absolute scale (e.g., $s(\text{real}) = 0$, $s(\text{viol}) = 1$). At inference time, the model operates on a single video without requiring pairwise comparison, producing a continuous estimate of physical violation severity.

2. **Curating rich supervision from heterogenous sources.** We reorganize diverse supervision sources into a unified training framework: real vs. generated comparisons (with and without semantic correspondence) [61], real vs. real pairs for consistency and calibration [24, 11], synthetic failure cases [2, 28], and human-annotated plausibility judgments [6, 20]. We organize these into three forms of supervision:
   (a) Pairwise Ranking Supervision: Stating "video A is more physically violated than video B" is more noise-tolerant than assigning absolute scores, providing stable relative ordering.
   (b) Regress to Noisy Scalar Annotations: Where available, human-annotated severity scores provide approximate magnitude information that grounds the ordering in absolute values.
   (c) Anchoring Violation Scores: Real videos map to 0, clearly violated videos map to 1, stabilizing the absolute scale across datasets.
3. **Unified training objective for scoring.** The three supervision signals are combined into a single objective that decomposes the measurement problem into ranking (Bradley-Terry), magnitude (regression), and scale calibration (anchoring) losses, each addressed by the data-source best suited for it. PHYPROBE trains only 1M parameters on top of a frozen 2B-parameter encoder, and is an effective evaluator for physical consistency.

We extensively evaluate PHYPROBE across diverse settings to assess preference accuracy, human alignment, and generalization. On pairwise benchmarks spanning real versus generated and generated versus generated comparisons under varying correspondence levels [54, 6, 20, 36], PHYPROBE maintains strong performance even in no correspondence and generated versus generated settings where prior evaluators are inferior. It also shows moderate correlation on VideoPhy2 [6] and strong correlation on VideoFeedback2 [20] human judgment test sets, indicating well ordered and stably anchored scores. Despite being trained on supervision indicative of physical consistency, without explicit general-preference labels, PHYPROBE further generalizes to broader human preference benchmarks [56, 45, 46, 29], suggesting that representations for detecting physical violations capture broader quality signals.

2 Related Work

Intuitive Physics and Human Perception. Understanding physical plausibility has long been studied in cognitive science under the framework of intuitive physics, which investigates how humans perceive and reason about physical interactions [55, 3, 43, 4, 23]. Prior work suggests that human

judgments on visual abnormalities are often coarse and thresholded, with observers reliably detecting only salient violations while overlooking subtle inconsistencies [39, 40, 38]. These findings indicate that human perception does not provide a precise or stable measure of physical correctness. Rather than modeling human reasoning itself, our work focuses on how such judgments are operationalized in evaluation, motivating the need for measurement frameworks that do not directly depend on unstable human perception. This further motivates evaluation approaches that avoid directly modeling human perception, and instead rely on more stable representation-level signals of physical dynamics.

Video Physical Benchmarks and Datasets. A wide range of benchmarks have been proposed to evaluate physical plausibility in video generation, including Physics-IQ, PhyGenBench, WorldModelBench, Physion, and IntPhys2 etc. [37, 34, 27, 7, 9, 13, 51]. More recent datasets such as ImpossibleVideos [2] and BrokenVideos [28] focus specifically on failure cases in generated videos. Many of these benchmarks rely on human annotations, either through pairwise preferences or scalar ratings, which are increasingly used to train automated evaluators [5, 6, 19, 20, 36, 53]. However, such annotations are inherently unstable: physical inconsistencies often co-occur with rendering artifacts and compositional errors, causing annotators to respond to overall perceptual quality rather than physical dynamics specifically. Moreover, as generative models improve, the annotation scales may drift over time and become implicitly calibrated to the prevailing data distribution rather than a fixed notion of physical consistency. These limitations highlight the difficulty of isolating physical consistency as a distinct evaluation signal.

Video Evaluators for Physical Consistency. Recent approaches increasingly employ VLM-based evaluators trained on human annotations to assess physical plausibility in generated videos [19, 20, 36], such as VideoPhy2 [5, 6]. Other approaches fine-tune VLMs to detect or localize specific physical violations [54, 15, 44]. While effective on benchmark tasks, these methods fundamentally operate on semantic proxy signals and annotation patterns rather than directly modeling physical dynamics, making them sensitive to training data coverage and prone to degradation under distribution shift or subtle out-of-distribution violations. In contrast, our work formulates physical consistency evaluation as a representation-level measurement problem over temporal dynamics.

3 PHYPROBE

PHYPROBE is designed to measure physical consistency in generated videos, mapping a video v to a scalar violation score $s(v)$ that indicates the extent of physical consistency violation. To be effective, such an evaluator must be able to: (1) quantify violation severity through reliable relative ordering across videos and, (2) provide absolute scores on a stable, anchored scale to enable comparison across settings. We describe how PHYPROBE achieves this through its architecture, data curation pipeline, and training objective.

Table 1: **Training datasets used for PHYPROBE.** Each dataset contributes to different supervision signals: pairwise preference learning (rank), regression (mag), and anchor-based calibration (cal). During training, we sample data from different sources using dataset-level weights to balance their contributions, preventing large datasets from dominating the learning signal.

Training Dataset	Pairs	Type	Correspondence	Rank	Mag	Cal
PAI-Bench [61]	42K	R-G	Image + Text	✓		✓
VideoPhy2 [6]	19K	G-G	Action / Text	✓	✓	
VideoFeedback2 [20]	31K	G-G	None	✓	✓	
BrokenVideos [28]	14K	G-G	None	✓	✓	
ImpossibleVideos [2]	27K	R-G	None	✓		✓
TRAVL [36]	98K	R-G	None	✓		✓
Kinetics-700 [11]	140K	R-R	Action	✓		✓
Total	371K					

Architecture. Given a video v (generated or real), we extract video representations using a frozen pretrained spatio-temporal visual encoder (we primarily use Perception Encoder [8]) to obtain a feature representation $\mathbf{z} = f(v)$, where $f(\cdot)$ denotes the frozen backbone. The input videos are preprocessed as fixed-length clips with frames T, resized to a fixed spatial resolution with the aspect ratio preserved via padding, followed by backbone-specific normalization. The resulting representation $\mathbf{z}$ is mapped to a scalar violation score via a lightweight learned scoring head $g_\theta(\cdot)$, i.e., $s(v) = g_\theta(\mathbf{z})$. In principle, PHYPROBE is agnostic to the choice of pretrained backbone, and

we evaluate multiple architectures in Section 4. To obtain a video-level descriptor, we aggregate per-frame backbone features using a statistics pooling operator that concatenates the temporal mean, standard deviation, and max across timesteps. This captures overall content, temporal variability, and salient events, providing a simple yet effective summary of the feature distribution while remaining lightweight and invariant to video length. The aggregated representation is then passed to a lightweight three-layer Multi-Layer Perceptron (MLP) with GELU activations that maps it to a scalar violation score. The design is intentionally simple and parameter-efficient, allowing the pretrained backbone to handle spatiotemporal modeling while the head focuses on learning a smooth, consistently scaled scoring function. We ablate alternative aggregation strategies in Section 4.3.

Data Curation. We train PHYPROBE on a curated dataset of approximately 371K video pairs drawn from seven diverse sources. Each data source spans a range of pair types (real vs. generated comparisons [61], real vs. real pairs for consistency and calibration [24, 11], synthetic failure cases [28, 2], and human-annotated judgments [6, 20]) under varying degrees of semantic correspondence as shown in Table 1. This heterogeneity is deliberate: by exposing the model to comparisons with and without shared prompts or content, we discourage reliance on semantic shortcuts and encourage sensitivity to physical consistency itself. The videos cover a broad range of physical violation types: behavioral abnormalities such as object interpenetration, unrealistic fluid dynamics, implausible deformation, and erratic motion violating expected force and collision dynamics; time abnormalities including temporal jumps, motion reversal, and violations of causal ordering; objectness abnormalities such as objects suddenly appearing or disappearing, unexpected morphing, and inconsistent spatial attributes across frames; and perception abnormalities in the form of visual breakdown, local blurring, and style inconsistencies introduced by generative artifacts.

Learning Objectives. We categorize the training supervision from these data sources into three complementary forms, each addressing a distinct aspect of the scoring problem.

1. **Ranking Supervision.** We construct video pairs from both aligned sources (shared prompts or content) and unaligned sources (no shared prompt), providing relative ordering signals under varying semantic correspondence. Including both prevents the model from being over-reliant on shared semantic correspondence. To recover a scalar score from pairwise preferences, we adopt the Bradley-Terry formulation [10], which treats $s(\cdot)$ as a latent severity score and models the probability that v_i is more violated than v_j as $\sigma(s(v_i) - s(v_j))$, where σ is the logistic sigmoid. Given a pair (v_i, v_j) from the ranking supervision dataset $\mathcal{D}_{\text{rank}}$, we define:
$$\mathcal{L}_{\text{rank}} = -\mathbb{E}_{(v_i,v_j,y_{ij})\sim\mathcal{D}_{\text{rank}}} \left[y_{ij} \log \sigma(s(v_i) - s(v_j)) + (1 - y_{ij}) \log \sigma(s(v_j) - s(v_i)) \right], \quad (1)$$
where $y_{ij} \in \{0, 0.5, 1\}$ indicates the relative preference between v_i and v_j: 1 if v_i is more violated, 0 if v_j is, and 0.5 for real–real pairs that carry no preference, so the ranking term pushes $s(v_i)$ and $s(v_j)$ toward equality.
2. **Regression for Violation Magnitude.** Noisy scalar annotations provide approximate violation severity. Different datasets use varying scales and definitions, so for datasets with discrete or ordinal ratings (e.g., VideoFeedback2 [20], VideoPhy2 [6]), we apply monotonic transformations to a unified range while preserving relative ordering. When explicit scores are unavailable, we derive approximate magnitudes from dataset-specific signals (e.g, spatial extent and duration of annotated errors in BrokenVideos [28]), treating them as noisy measurements. To model the severity of violations, given noisy scalar annotations from $\mathcal{D}_{\text{noisy}}$ we incorporate a regression objective using the Huber loss [21]:
$$\text{Huber}(s, y) = \begin{cases} \frac{1}{2}(s - y)^2, & \text{if } |s - y| \le \delta \\ \delta(|s - y| - \frac{1}{2}\delta), & \text{otherwise} \end{cases} \quad (2)$$
$$\mathcal{L}_{\text{mag}} = \mathbb{E}_{v\sim\mathcal{D}_{\text{noisy}}} \left[\text{Huber}(s(v), y_v) \right], \quad (3)$$
3. **Anchor Loss for Scale Calibration.** Anchor samples represent approximate endpoints of the violation spectrum: they are selected to represent approximate endpoints, including high-confidence real videos [24] (low violation; anchor toward $s \approx 0$) and synthetic failure cases [2] (high violation; anchor toward $s \approx 1$). Real-real pairs reinforce the near-zero baseline, and real-generated comparisons help establish relative scale across domains. These provide weak but stable reference points for resolving scale ambiguity across heterogeneous sources (a curated $\mathcal{D}_{\text{anchor}}$ dataset with binary labels). To calibrate the score range, we introduce an anchor loss:
$$\mathcal{L}_{\text{cal}} = \mathbb{E}_{v\sim\mathcal{D}_{\text{anchor}}} \left[\|s(v) - \mathbf{I}(\mathbf{v})\|_2^2 \right], \quad (4)$$
where $\mathbf{I}(v) = \begin{cases} 0 & \text{if } v \in \mathcal{D}_{\text{anchor}}^{\text{plausible}} \\ 1 & \text{if } v \in \mathcal{D}_{\text{anchor}}^{\text{violated}} \end{cases}$ based on the curated $\mathcal{D}_{\text{anchor}}$ samples.

The overall objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \lambda_{\text{mag}}\mathcal{L}_{\text{mag}} + \lambda_{\text{cal}}\mathcal{L}_{\text{cal}}, \tag{5}$$

In practice, the $[0, 1]$ violation interval should be viewed as a calibration target rather than a strict architectural constraint. While the regression and anchor losses encourage scores to concentrate near this range during training, the scoring head itself remains unconstrained and does not explicitly enforce bounded outputs. To obtain a stable and interpretable violation scale, we report $s(v) = \text{clip}(\tilde{s}(v), 0, 1)$ at inference time. Empirically, we observe that the learned score distributions already remain close to the target interval even without explicit constraints as shown in Table 4.

4 Experiment Results

4.1 Experimental Setup

Evaluation Benchmarks. We evaluate PHYPROBE predominantly on pairwise preference accuracy on multiple benchmarks spanning different levels of semantic correspondence and pair composition. The role and constitution of each evaluation dataset is summarized in Table 13 in Appendix B. For pairwise physical consistency, we use four benchmarks: ImplausibleBench [36] (image and text semantic correspondence with real-generated test pairs), VideoPhy2 [6] (text correspondence with generated-generated pairs), PhyDetEx [54] (no semantic correspondence, with real-generated pairs), and VideoFeedback2 [20] (no correspondence, with generated-generated pairs). Since absolute annotation scales vary across datasets, we convert scalar-annotated benchmarks into a pairwise format by ranking samples within each dataset, enabling consistent comparison alongside natively paired benchmarks. We report pairwise accuracy for ranking performance. To measure alignment with human-ratings (i.e. human correlation), we use VideoPhy2 and VideoFeedback2, both of which provide scalar physical plausibility annotations. We compute both Spearman's rank correlation (ρ) to assess ordering consistency and Pearson correlation (r) to evaluate linear alignment. We note that neither metric measures absolute calibration; we assess scale stability via the anchor ablation (Table 4). To further measure PHYPROBE's *alignment with human preference*, we evaluate on MonetBench [56], RewardBench [29], and Rapidata [45, 46], and which assess broader video quality and preference alignment beyond physical consistency.

Baselines and Prior Methods. We compare against five baselines spanning different evaluation paradigms: VideoPhy-2-AutoEval [6], a fine-tuned VLM evaluator for physical plausibility; VideoScore2 [20], a reward model trained on human preferences but also support physical and commonsense evaluation; V-JEPA Surprisal Score [1], a self-supervised prediction-error metric based on V-JEPA2; finally Gemini-3.1-Flash-Lite [16] and Gemini-3.1-Pro, off-the-shelf VLMs prompted zero-shot.

4.2 Core Results

▷ **PHYPROBE is a strong physical consistency evaluator.** PHYPROBE achieves the highest pairwise accuracy on three out of four physical benchmarks as shown in Table 2, reaching 98.0–98.9% on R-G pairs and 66.1–75.2% on harder G-G pairs. Unlike baselines, whose performance varies considerably across correspondence settings, PHYPROBE remains strong with and without semantic correspondence. It also achieves the strongest correlation with human judgments on VF2 (0.596/0.604, ρ/r) is competitive with VideoPhy2-AutoEval on VP2. Overall, PHYPROBE delivers what a physical consistency evaluator should: bounded, consistently scaled scores that reliably order videos and reflect the severity of physical violations. This can also be seen in Figure 3 and 4 – PHYPROBE consistently aligns with physical plausibility, assigning low scores to physically consistent motion and high scores to clear violations such as "object explosion" or "inconsistent dynamics", even when human annotations are ambiguous or inconsistent. We feature more qualitative results in Figure 6 and 7 in Appendix A.1. With the same V-JEPA2-H/16 encoder, PHYPROBE outperforms V-JEPA Surprise by 13.9–48.1 points on all four benchmarks (Table 7).

▷ **PHYPROBE balances alignment across annotation scales.** From Table 2, we can see VP2-AutoEval (row 1) achieves the highest VP2 correlation ($\rho = 0.359$) but does not generalize well to VF2. Conversely, VideoScore2 (row 3) shows strong VF2 correlation ($\rho = 0.462$) but degrades on VP2 ($\rho = 0.197$). Both methods appear calibrated to their respective annotation distributions rather than capturing transferable measurements of physical consistency. In contrast, PHYPROBE achieves moderate correlation on VP2 ($\rho = 0.293$) and the strongest correlation on VF2 ($\rho = 0.596$) simultaneously, without collapsing on either. This balance reflects co-training across datasets and

Table 2: **Core results on physical video benchmarks. Left Columns:** pairwise accuracy (%) disentangled by pair type (real (R) vs generated (G)) and correspondence type (image+text, text, none). Benchmark^p indicates the benchmark is reordered in pairwise manner without ties. We additionally report a weighted average across all pairwise benchmarks based on the number of pairs in each benchmark. **Right Columns:** correlation with human judgments on VideoPhy2 (VP2) and VideoFeedback2 (VF2). We report Spearman's rank correlation (ρ) and Pearson correlation (r) to assess both ranking consistency and linear alignment. We additionally report Fisher-z averaged correlations across datasets, weighting each benchmark equally. We further studied the prior method behaviors in Appendix D.

		Pairwise Accuracy (%) ↑					**Human Correlation ↑**					
		Img+Txt	**Text**	**None**		**Avg.**	**VP2**		**VF2**		**Avg.**	
Method	**Backbone**	ImplB^p (R–G)	VP2^p (G–G)	PhyDEx^p (R–G)	VF2^p (G–G)		ρ	r	ρ	r	ρ	r
1 VP2-AutoEval	VideoCon (7B)	28.0	28.5	30.5	36.3	32.1	**0.359**	**0.362**	0.235	0.240	0.298	0.302
2 V-JEPA Surprise	V-JEPA2-H/16 (600M)	33.3	46.3	49.0	45.5	46.6	0.065	0.065	0.121	0.133	0.093	0.099
3 VideoScore2	Qwen2.5-VL (7B)	68.7	49.7	60.6	65.0	59.9	0.197	0.160	0.462	0.430	0.336	0.301
4 Gemini-3.1-Flash-Lite	N/A	91.3	61.8	89.0	74.5	77.3	0.302	0.303	0.419	0.409	0.362	0.357
5 Gemini-3.1-Pro	N/A	95.2	**70.5**	86.2	73.7	78.1	0.277	0.240	0.402	0.383	0.341	0.313
6 **PHYPROBE**	PE-Core-G14-448 (2B)	**98.0**	66.1	**98.9**	**75.2**	**82.4**	0.293	0.302	**0.596**	**0.604**	**0.458**	**0.466**

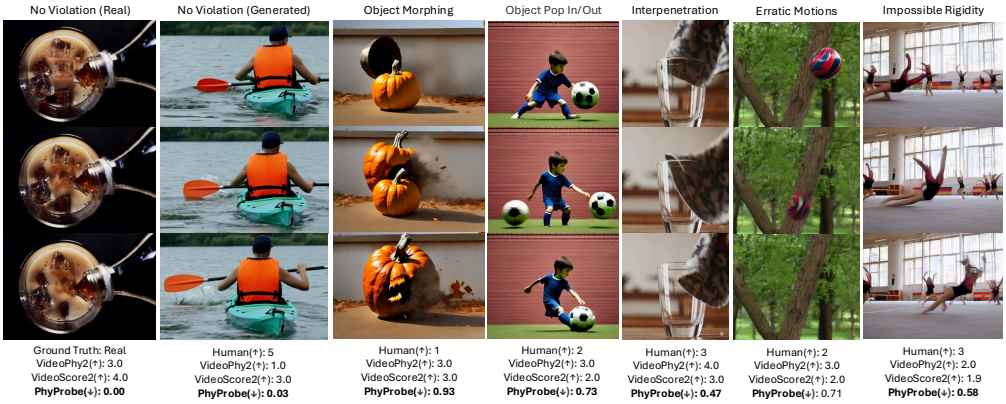


Figure 3: **Qualitative comparison of physical plausibility assessment across videos.** Each column shows representative frames, along with human annotations and scores from PHYPROBE and prior physical evaluators [6, 20]. Arrows indicate whether the score is higher the better or vice versa.

partial transfer between their supervision signals, although the leave-one-dataset-out analysis reveals meaningful dataset-specific dependence (Table 9). As further shown in Figure 5, prior methods produce predictions concentrated around discrete annotation levels, whereas PHYPROBE yields more smoothly separated score distributions with broader intra-class variation. This behavior suggests that PHYPROBE captures a continuous notion of violation severity beyond discrete human rating bins.

▷ **PHYPROBE transfers to general preference benchmarks without explicit preference-reward supervision.** Although PHYPROBE is trained on supervision indicative of physical consistency, without explicit general-preference labels, it achieves surprisingly strong performance on human preference benchmarks as shown in Table 3, outperforming dedicated reward models on MonetBench (72.7% vs. VisionReward [56] 68.2% and VideoReward [29] 56.1%, w/ tie), leading on Rapidata-I2V (62.7%), and remaining competitive on RewardBench (65.3% Overall w/o tie). We hypothesize this transfer occurs because physics violations frequently co-occur with broader quality degradations in generated videos, making physical consistency a significant component of perceptual preference.

4.3 Ablations

▷ **Effect of PHYPROBE loss components.** We ablate the effect of each loss term by removing it from PHYPROBE training as shown in Table 4. Removing $\mathcal{L}_{\text{rank}}$ degrades pairwise ordering, particularly on VP2 and PhyDetEx. Without explicit preference supervision, the model lacks direct signal for relative comparison between videos of similar quality. Removing $\mathcal{L}_{\text{mag}}$ leaves pairwise accuracy

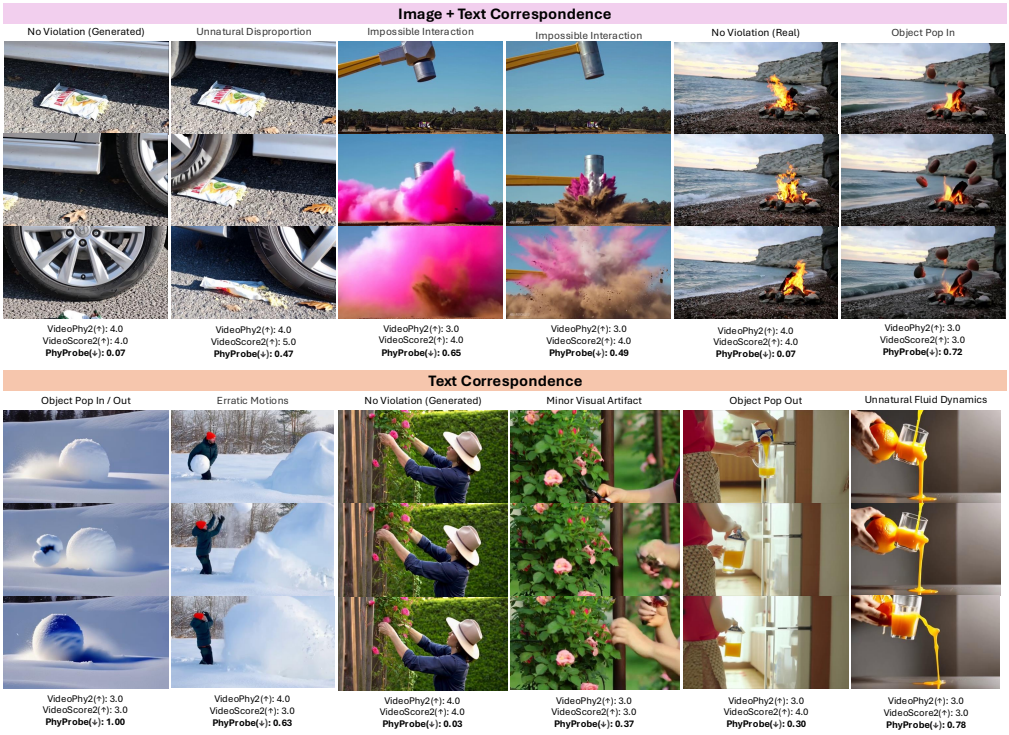


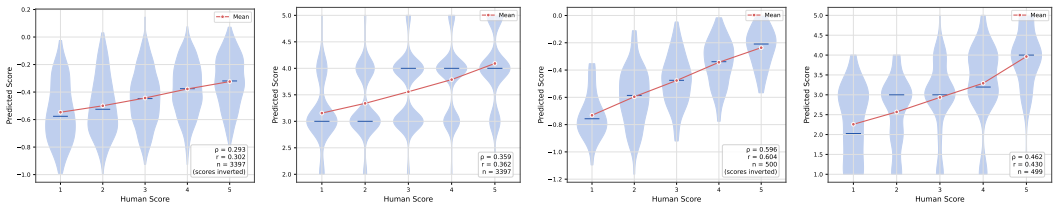
Figure 4: **Qualitative results across correspondence levels.** PHYPROBE assigns low violation scores to videos with coherent object motion and interaction dynamics, while assigning higher scores to videos exhibiting implausible behaviors, effective across various correspondence level.

Table 3: **PHYPROBE aligns with broader human preference beyond physical consistency.** Pairwise accuracy (%) on human preference (quality) benchmarks [56, 45, 46, 29] under different tie-handling protocols. Despite being trained without explicit general-preference labels, PHYPROBE achieves competitive performance across all datasets, categories (overall, MQ, VQ), and evaluation protocols. PHYPROBE uses a model-side tie threshold at the 20th percentile.

Method	**MonetBench**		**Rapidata-I2V**	**VideoGen-RewardBench**					
	w/ tie	w/o tie	w/o tie only	Overall		MQ		VQ	
				w/ tie	w/o tie	w/ tie	w/o tie	w/ tie	w/o tie
1 Random	33.2	49.9	49.7	33.2	49.7	33.2	49.5	33.3	49.8
2 V-JEPA Surprise	52.4	53.8	45.1	45.2	44.2	48.8	47.1	48.7	47.6
3 VideoScore2	42.3	41.1	45.8	57.4	58.9	54.3	60.6	55.3	60.1
4 VideoPhy-2-AutoEval	22.5	16.5	9.5	24.6	19.5	37.9	20.5	34.4	20.4
5 VisionReward	68.2	**72.2**	47.1	64.6	67.6	54.6	61.3	54.8	59.1
6 VideoReward	56.1	58.2	51.5	**69.7**	**73.6**	60.3	**75.2**	**63.6**	**75.9**
7 **PHYPROBE**	**72.7**	71.0	**62.7**	64.1	65.3	**62.3**	65.9	61.4	63.2

largely intact but causes the score range to explode: without regression targets, PHYPROBE learns to order videos correctly but has no incentive to map violation scores to a bounded, interpretable scale. Removing $\mathcal{L}_{\text{cal}}$ also causing score range to explode and collapses real–generated pairwise accuracy: without anchor points defining what constitutes a clearly consistent or clearly violated video, the model cannot assign appropriate scores to obvious cases, even though it may still rank ambiguous pairs well. The full objective provides the best trade-off: trades some correlation for bounded range and strong pairwise accuracy, and alignment with human ratings.

▷ **Effect of Backbone Representations.** We evaluate PHYPROBE with different frozen pretrained video encoders, as shown in Table 5. We compare state-of-the-art video tokenizers (Wan-2.2 VAE [50]), latent world model architectures (V-JEPA2 [1]), and weakly supervised pretrained



(a) (b) (c) (d)

Figure 5: Distribution of evaluator predictions conditioned on human rating levels on VideoPhy2 (VP2) and VideoFeedback2 (VF2). From left to right: (a) PHYPROBEon VP2, (b) VP2-AutoEval on VP2, (c) PHYPROBEon VF2, and (d) VideoScore2 on VF2. Each violin shows the distribution of predicted scores for videos assigned the same human rating, with the red line indicating the mean. For visualization, PHYPROBEviolation scores are inverted so that higher values correspond to greater physical plausibility, matching the direction of the baseline evaluators. Compared with prior methods, PHYPROBEcharacterizes physical inconsistencies with greater granularity than the five-point human scale, providing a fluid assessment of error intensity beyond rigid categorical labels.

Table 4: **Importance of individual loss components in PHYPROBE.** We compare PHYPROBE with different loss components using the same backbone (PE-Core-G14-448). The scoring range reflects the effective output range of each variant.

Variant	Pairwise Accuracy (%)				Human Correlation				Score Range
	None		Txt	Img+Txt	VP2		VF2		
	PhyDEx^{p} (R–G)	VF2^{p} (G–G)	VP2^{p} (G–G)	ImplB^{p} (R–G)	ρ	r	ρ	r	
PHYPROBE ($-\mathcal{L}_{\text{rank}}$)	94.7	75.6	61.9	99.2	0.311	0.316	0.614	0.609	$[-0.07, 0.88]$
PHYPROBE ($-\mathcal{L}_{\text{mag}}$)	98.7	75.3	62.3	94.1	0.315	0.316	0.584	0.583	$[-1.89, 3.45]$
PHYPROBE ($-\mathcal{L}_{\text{cal}}$)	68.7	**79.3**	64.4	84.9	0.363	0.364	0.664	0.656	$[-0.40, 4.13]$
PHYPROBE	**98.9**	75.2	**66.1**	**98.0**	0.293	0.302	0.596	0.604	$[-0.00, 1.15]$

visual encoders (Perception Encoder [8]). We observe substantial performance differences across representation types, indicating that physical consistency evaluation is highly dependent on the quality of the underlying spatiotemporal representation space. In particular, weakly supervised pretrained visual encoders (PE) achieve the strongest and most consistent performance across both pairwise evaluation and human correlation benchmarks. We hypothesize that their large-scale pretraining and robust visual representation learning pipeline better preserve object structure, motion continuity, and interaction dynamics. V-JEPA2 also performs competitively, suggesting that predictive spatiotemporal representations are naturally aligned with modeling physical dynamics, whereas latent representations optimized primarily for reconstruction exhibit weaker transfer.

▷ **Effect of Spatiotemporal Aggregation.** We compare different spatiotemporal aggregation strategies for mapping frame-level features to a scalar score in Table 6. Within each frame, patch tokens are spatially aggregated via mean pooling to produce a single frame-level descriptor. Across frames, a statistics-based aggregation (mean, standard deviation, and max pooling concatenated) provides a strong performance-efficiency trade-off. Despite its simplicity, this approach captures key spatiotemporal variations indicative of physical inconsistencies, while avoiding the optimization challenges and computational overhead of attention-based aggregation. We therefore used statistical aggregation head (mean+std+max) in our final model, as it achieves the best trade-off between accuracy and model complexity.

▷ **Further ablations (Appendix A.2).** PHYPROBE's advantage is robust to the minimum score gap used to construct evaluation pairs (Table 8); its learned signal transfers partially across supervision sources while retaining dataset-specific dependence (Table 9); and violation-flooring normalization improves pairwise accuracy over linear normalization on all four benchmarks (Table 10).

Table 5: **Effect of backbone representations on PHYPROBE**. We train and evaluate PHYPROBE with different pretrained video encoders, reporting pairwise accuracy (%) and correlation with human judgments. Pairwise accuracy is decomposed by correspondence type (image+text, text, none) and pair type (real (R) vs generated (G)). Benchmark^p indicates the benchmark is reordered in pairwise manner without ties. Correlation metrics (Spearman ρ, Pearson r) measure alignment with human judgments on VideoPhy2 (VP2) and VideoFeedback2 (VF2).

		Pairwise Accuracy (%) ↑				**Human Correlation ↑**			
		Img+Txt	**Text**	**None**		**VP2**		**VF2**	
PHYPROBE Backbone	**Params**	ImplB^p (R–G)	VP2^p (G–G)	PhyDEx^p (R–G)	VF2^p (G–G)	ρ	r	ρ	r
1 Wan2.2 VAE	300M	71.7	58.3	67.1	74.3	0.186	0.194	0.562	0.581
2 V-JEPA2-VIT-L/16	300M	76.1	60.6	88.8	74.2	0.228	0.206	0.564	0.580
3 V-JEPA2-VIT-H/16	600M	81.4	60.2	70.7	76.3	0.301	0.303	0.623	0.630
4 V-JEPA2-VIT-G/16	1B	88.6	60.9	93.0	72.6	0.279	0.288	0.566	0.580
5 PE-Spatial-G14-448	2B	98.7	61.3	99.5	73.0	0.244	0.253	0.566	0.573
6 PE-Core-G14-448	2B	98.0	66.1	98.9	75.2	0.293	0.302	0.596	0.604

Table 6: Ablation of projection head designs. We compare different spatiotemporal aggregation strategies in terms of performance and efficiency, under the same PECore backbone. Rel. denotes relative trainable parameter count normalized to the Mean (MLP) head.

Head	**Trainable Params**	**Input Dim**	**Rel. (×)**	**VideoPhy2**		**VideoFeedback2**	
				Acc	ρ	Acc	ρ
Mean (MLP)	0.36M	$[B, D]$	1×	61.0	0.277	73.2	0.578
Attention Pooling	6.92M	$[B, T, D]$	19.2×	63.7	0.292	75.0	0.612
Stat (mean+std+max)	1.02M	$[B, 3D]$	2.8×	66.1	0.293	75.2	0.596

5 Limitations

Current spatiotemporal representations model long-range causal dependencies imperfectly, so violations that emerge only over extended horizons may be under-encoded [57]; conversely, very brief or instantaneous violations (e.g., abrupt state changes) may not persist long enough to register in the aggregated representation, and unusual but physically valid motions in real videos can receive elevated scores. Failure cases are shown in Figure 8 (Appendix A.1). PHYPROBE is also trained to approximate human-perceived plausibility, the operational definition used by existing benchmarks; since those judgments correlate with other perceptual factors such as overall quality, our model may partially capture them as well. A controlled evaluation varying physical consistency while holding quality, scene, prompt, and generator fixed would give stronger evidence of disentanglement, but to our knowledge no benchmark with reliable human annotations enables this, a limitation shared by current physical-consistency benchmarks and evaluators.

6 Conclusion

We introduced PHYPROBE, an evaluator for physical consistency in generated videos. Through a unified training objective combining pairwise ranking, regression on noisy scalar annotations, and anchor-based calibration across heterogeneous data sources, PHYPROBE produces continuous, stably anchored violation scores that capture the severity of physical inconsistencies. Empirically, PHYPROBE achieves strong performance across diverse evaluation settings, maintaining robust pairwise ordering even in challenging regimes where prior approaches degrade, while also aligning well with human judgments. Notably, PHYPROBE generalizes to broader human-preference benchmarks despite being trained on supervision selected to indicate physical consistency rather than general preference. We hope this work encourages future research towards the development of more reliable and scalable metrics for assessing the physical fidelity of video generation systems.

Acknowledgments and Disclosures

Funding and Support. This work was supported in part by NVIDIA and the Vector Institute. We thank NVIDIA and the Vector Institute for providing computational resources and infrastructure that supported this research.

Competing Interests. Several authors are affiliated with NVIDIA, which provided research support for this work. The authors declare no other competing interests.

References

[1] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, Mojtaba, Komeili, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus, S. Arnaud, A. Gejji, A. Martin, F. R. Hogan, D. Dugas, P. Bojanowski, V. Khalidov, P. Labatut, F. Massa, M. Szafraniec, K. Krishnakumar, Y. Li, X. Ma, S. Chandar, F. Meier, Y. LeCun, M. Rabbat, and N. Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

[2] Z. Bai, H. Ci, and M. Z. Shou. Impossible videos, 2025. URL https://arxiv.org/abs/2503.14378.

[3] R. Baillargeon, E. S. Spelke, and S. Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(85)90008-3. URL https://www.sciencedirect.com/science/article/pii/0010027785900083.

[4] H. Balaban, K. A. Smith, J. B. Tenenbaum, and T. D. Ullman. Electrophysiology reveals that intuitive physics guides visual tracking and working memory. *Open Mind*, 8:1425–1446, 11 2024. ISSN 2470-2986. doi: 10.1162/opmi_a_00174. URL https://doi.org/10.1162/opmi_a_00174.

[5] H. Bansal, Z. Lin, T. Xie, Z. Zong, M. Yarom, Y. Bitton, C. Jiang, Y. Sun, K.-W. Chang, and A. Grover. Videophy: Evaluating physical commonsense for video generation, 2024. URL https://arxiv.org/abs/2406.03520.

[6] 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, 2025. URL https://arxiv.org/abs/2503.06800.

[7] D. M. Bear, E. Wang, D. Mrowca, F. J. Binder, H.-Y. F. Tung, R. T. Pramod, C. Holdaway, S. Tao, K. Smith, F.-Y. Sun, L. Fei-Fei, N. Kanwisher, J. B. Tenenbaum, D. L. K. Yamins, and J. E. Fan. Physion: Evaluating physical prediction from vision in humans and machines, 2022. URL https://arxiv.org/abs/2106.08261.

[8] D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. Rasheed, J. Wang, M. Monteiro, H. Xu, S. Dong, N. Ravi, D. Li, P. Dollár, and C. Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network, 2025. URL https://arxiv.org/abs/2504.13181.

[9] F. Bordes, Q. Garrido, J. T. Kao, A. Williams, M. Rabbat, and E. Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025. URL https://arxiv.org/abs/2506.09849.

[10] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952. URL https://api.semanticscholar.org/CorpusID:125209808.

[11] J. Carreira, E. Noland, C. Hillier, and A. Zisserman. A short note on the kinetics-700 human action dataset, 2022. URL https://arxiv.org/abs/1907.06987.

[12] 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, 2024. URL https://arxiv.org/abs/2401.09047.

[13] Y. Chen, X. Zhu, and T. Li. A physical coherence benchmark for evaluating video generation models via optical flow-guided frame prediction, 2025. URL https://arxiv.org/abs/2502.05503.

[14] G. DeepMind. Veo 3 technical report. Technical report, Google, 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf. Accessed: 2026-04-29.

[15] Y. Gao, Y. Ding, H. Su, J. Li, Y. Zhao, L. Luo, Z. Chen, L. Wang, X. Wang, Y. Wang, X. Ma, and Y.-G. Jiang. David-xr1: Detecting ai-generated videos with explainable reasoning, 2025. URL https://arxiv.org/abs/2506.14827.

[16] Google AI for Developers. Gemini 3 developer guide. https://ai.google.dev/gemini-api/docs/gemini-3, 2026. Accessed: 2026-01-15.

[17] Y. HaCohen, N. Chiprut, B. Brazowski, D. Shalem, D. Moshe, E. Richardson, E. Levin, G. Shiran, N. Zabari, O. Gordon, P. Panet, S. Weissbuch, V. Kulikov, Y. Bitterman, Z. Melumian, and O. Bibi. Ltx-video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103.

[18] Y. HaCohen, B. Brazowski, N. Chiprut, Y. Bitterman, A. Kvochko, A. Berkowitz, D. Shalem, D. Lifschitz, D. Moshe, E. Porat, E. Richardson, G. Shiran, I. Chachy, J. Chetboun, M. Finkelson, M. Kupchick, N. Zabari, N. Guetta, N. Kotler, O. Bibi, O. Gordon, P. Panet, R. Benita, S. Armon, V. Kulikov, Y. Inger, Y. Shiftan, Z. Melumian, and Z. Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

[19] X. He, D. Jiang, G. Zhang, M. Ku, A. Soni, S. Siu, H. Chen, A. Chandra, Z. Jiang, A. Arulraj, K. Wang, Q. D. Do, Y. Ni, B. Lyu, Y. Narsupalli, R. Fan, Z. Lyu, Y. Lin, and W. Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation, 2024. URL https://arxiv.org/abs/2406.15252.

[20] X. He, D. Jiang, P. Nie, M. Liu, Z. Jiang, M. Su, W. Ma, J. Lin, C. Ye, Y. Lu, K. Wu, B. Schneider, Q. D. Do, Z. Li, Y. Jia, Y. Zhang, G. Cheng, H. Wang, W. Zhou, Q. Lin, Y. Zhang, G. Zhang, W. Huang, and W. Chen. Videoscore2: Think before you score in generative video evaluation, 2025. URL https://arxiv.org/abs/2509.22799.

[21] P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35: 492–518, 1964. URL https://api.semanticscholar.org/CorpusID:121252793.

[22] D. Jiang, M. Ku, T. Li, Y. Ni, S. Sun, R. Fan, and W. Chen. Genai arena: An open evaluation platform for generative models, 2024. URL https://arxiv.org/abs/2406.04485.

[23] S. Joseph, Q. Garrido, R. Balestriero, M. Kowal, T. Fel, S. Bakhtiari, B. Richards, and M. Rabbat. Interpreting physics in video world models, 2026. URL https://arxiv.org/abs/2602.07050.

[24] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset, 2017. URL https://arxiv.org/abs/1705.06950.

[25] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang, K. Wu, Q. Lin, J. Yuan, Y. Long, A. Wang, A. Wang, C. Li, D. Huang, F. Yang, H. Tan, H. Wang, J. Song, J. Bai, J. Wu, J. Xue, J. Wang, K. Wang, M. Liu, P. Li, S. Li, W. Wang, W. Yu, X. Deng, Y. Li, Y. Chen, Y. Cui, Y. Peng, Z. Yu, Z. He, Z. Xu, Z. Zhou, Z. Xu, Y. Tao, Q. Lu, S. Liu, D. Zhou, H. Wang, Y. Yang, D. Wang, Y. Liu, J. Jiang, and C. Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

[26] B. Li, Z. Lin, D. Pathak, J. Li, Y. Fei, K. Wu, T. Ling, X. Xia, P. Zhang, G. Neubig, and D. Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation, 2024. URL https://arxiv.org/abs/2406.13743.

[27] D. Li, Y. Fang, Y. Chen, S. Yang, S. Cao, J. Wong, M. Luo, X. Wang, H. Yin, J. E. Gonzalez, I. Stoica, S. Han, and Y. Lu. Worldmodelbench: Judging video generation models as world models, 2025. URL https://arxiv.org/abs/2502.20694.

[28] J. Lin, W. Peng, B. Zi, Y. Gao, X. Qi, X. Ma, and Y.-G. Jiang. Brokenvideos: A benchmark dataset for fine-grained artifact localization in ai-generated videos, 2025. URL https://arxiv.org/abs/2506.20103.

[29] J. Liu, G. Liu, J. Liang, Z. Yuan, X. Liu, M. Zheng, X. Wu, Q. Wang, M. Xia, X. Wang, X. Liu, F. Yang, P. Wan, D. Zhang, K. Gai, Y. Yang, and W. Ouyang. Improving video generation with human feedback, 2025. URL https://arxiv.org/abs/2501.13918.

[30] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[31] Luma AI. Ray2: Luma ray2 video model, 2025. URL https://lumalabs.ai/ray2. Accessed: 2026-04-29.

[32] Luma AI. Ray3: Advanced reasoning video model, 2025. URL https://lumalabs.ai/ray. Accessed: 2026-04-29.

[33] F. Margoni, L. Surian, and R. Baillargeon. The violation-of-expectation paradigm: A conceptual overview. *Psychological Review*, 131(3):716–748, 2024. doi: 10.1037/rev0000450.

[34] 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, 2024. URL https://arxiv.org/abs/2410.05363.

[35] MiniMax. Hailuo AI - Text to Video & Image to Video generator, 2026. URL https://hailuoai.video/. Accessed: 2026-05-03.

[36] S. Motamed, M. Chen, L. V. Gool, and I. Laina. Travl: A recipe for making video-language models better judges of physics implausibility, 2025. URL https://arxiv.org/abs/2510.07550.

[37] S. Motamed, L. Culp, K. Swersky, P. Jaini, and R. Geirhos. Do generative video models understand physical principles?, 2025. URL https://arxiv.org/abs/2501.09038.

[38] S. J. Nightingale and H. Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, Feb 2022. doi: 10.1073/pnas.2120481119.

[39] S. J. Nightingale, K. A. Wade, and D. G. Watson. Can people identify original and manipulated photos of real-world scenes? *Cognitive Research: Principles and Implications*, 2(1):30, Jul 2017. doi: 10.1186/s41235-017-0067-2.

[40] S. J. Nightingale, K. A. Wade, H. Farid, and D. G. Watson. Can people detect errors in shadows and reflections? *Attention, Perception, & Psychophysics*, 81(8):2917–2943, Nov 2019. doi: 10.3758/s13414-019-01773-w.

[41] NVIDIA, :, N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding, D. Dworakowski, J. Fan, M. Fenzi, F. Ferroni, S. Fidler, D. Fox, S. Ge, Y. Ge, J. Gu, S. Gururani, E. He, J. Huang, J. Huffman, P. Jannaty, J. Jin, S. W. Kim, G. Klár, G. Lam, S. Lan, L. Leal-Taixe, A. Li, Z. Li, C.-H. Lin, T.-Y. Lin, H. Ling, M.-Y. Liu, X. Liu, A. Luo, Q. Ma, H. Mao, K. Mo, A. Mousavian, S. Nah, S. Niverty, D. Page, D. Paschalidou, Z. Patel, L. Pavao, M. Ramezanali, F. Reda, X. Ren, V. R. N. Sabavat, E. Schmerling, S. Shi, B. Stefaniak, S. Tang, L. Tchapmi, P. Tredak, W.-C. Tseng, J. Varghese, H. Wang, H. Wang, H. Wang, T.-C. Wang, F. Wei, X. Wei, J. Z. Wu, J. Xu, W. Yang, L. Yen-Chen, X. Zeng, Y. Zeng, J. Zhang, Q. Zhang, Y. Zhang, Q. Zhao, and A. Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

[42] OpenAI. Video generation models as world simulators, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/. Accessed: 2026-04-29.

[43] J. Perez and L. Feigenson. Violations of expectation trigger infants to search for explanations. *Cognition*, 218:104942, 2022. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2021.104942. URL https://www.sciencedirect.com/science/article/pii/S0010027721003656.

[44] Physion Labs. Galileo-0: Toward a scalable world critic for world models. https://physionlabs.ai/blog/galileo-0, Apr. 2026. Accessed: 2026-05-04.

[45] Rapidata. Image-to-video human preference: Hailuo-02 vs marey. https://huggingface.co/datasets/Rapidata/image-to-video-human-preference-hailuo-02-marey, 2025. 6k human preference annotations from 2k annotators on image-to-video generation.

[46] Rapidata. Image-to-video human preference: Seedance-1-pro benchmark. https://huggingface.co/datasets/Rapidata/image-to-video-human-preference-seedance-1-pro, 2025. Human preference dataset with pairwise video comparisons for evaluating image-to-video models.

[47] 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, 2020. URL https://arxiv.org/abs/1803.07616.

[48] T. Seedance, D. Chen, L. Chen, X. Chen, Y. Chen, Z. Chen, Z. Chen, F. Cheng, T. Cheng, Y. Cheng, M. Chi, X. Chi, J. Cong, Q. Cui, F. Ding, Q. Dong, Y. Du, H. Duanmu, J. Fan, J. Fang, J. Fang, Z. Fang, C. Feng, Y. Gao, D. Gu, D. Guo, H. Guo, Q. Guo, B. Hao, H. Hao, H. He, J. He, Q. He, T. Hoang, H. Hu, R. Hu, Y. Hu, J. Huang, W. Huang, Z. Huang, Z. Huang, J. Jin, M. Jing, A. Kim, S. Lao, Y. Leng, B. Li, G. Li, H. Li, H. Li, J. Li, M. Li, X. Li, X. Li, Y. Li, Y. Li, Y. Li, C. Liang, H. Liang, J. Liang, Y. Liang, W. Liao, J. H. Lien, S. Lin, X. Lin, F. Ling, Y. Ling, F. Liu, J. Liu, J. Liu, J. Liu, S. Liu, S. Liu, W. Liu, X. Liu, Z. Liu, R. Lu, L. Lyu, J. Ma, T. Ma, X. Nie, J. Ning, J. Pan, X. Pan, R. Peng, X. Qu, Y. Ren, Y. Shen, G. Shi, L. Shi, Y. Song, F. Sun, L. Sun, R. Sun, W. Tang, B. Tao, Z. Tao, D. Wang, F. Wang, H. Wang, K. Wang, Q. Wang, R. Wang, S. Wang, S. Wang, W. Wang, X. Wang, Y. Wang, Y. Wang, Y. Wang, Y. Wang, Z. Wang, Z. Wang, G. Wei, M. Wei, D. Wu, G. Wu, H. Wu, H. Wu, J. Wu, J. Wu, R. Wu, S. Wu, X. Wu, X. Wu, Y. Wu, R. Xia, X. Xia, X. Xiao, S. Xu, B. Yang, J. Yang, R. Yang, T. Yang, Y. Yang, Z. Yang, Z. Yang, F. Ye, B. Yi, X. Yin, Y. You, L. Yuan, W. Zeng, X. Zeng, Y. Zeng, S. Zhai, Z. Zhai, B. Zhang, C. Zhang, H. Zhang, J. Zhang, M. Zhang, P. Zhang, S. Zhang, X. Zhang, X. Zhang, X. Zhang, X. Zhang, Y. Zhang, Z. Zhang, H. Zhao, H. Zhao, L. Zhao, Y. Zhao, G. Zheng, J. Zheng, X. Zheng, Z. Zheng, K. Zhu, and F. Zuo. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

[49] K. Team, J. Chen, Y. Ci, X. Du, Z. Feng, K. Gai, S. Guo, F. Han, J. He, K. He, X. Hu, X. Hu, B. Jiang, F. Kong, H. Li, J. Li, Q. Li, S. Li, X. Li, Y. Li, J. Liang, B. Liao, Y. Liao, W. Lin, Q. Liu, X. Liu, Y. Liu, Y. Liu, S. Lu, H. Mao, Y. Mao, H. Ouyang, W. Qin, W. Shi, X. Shi, L. Su, H. Sun, P. Sun, P. Wan, C. Wang, C. Wang, M. Wang, Q. Wang, R. Wang, X. Wang, X. Wang, Z. Wang, M. Wei, T. Wen, G. Wu, X. Wu, Z. Wu, D. Xie, Y. Xiong, Y. Xu, S. Yang, Z. Yang, W. Ye, Z. Yuan, S. Zhang, S. Zhang, Y. Zhang, Y. Zhang, W. Zhao, R. Zhou, Y. Zhou, G. Zhu, and Y. Zhu. Kling-omni technical report, 2025. URL https://arxiv.org/abs/2512.16776.

[50] W. M. Team. Wan22 vae: Video variational autoencoder for wan video generation. https://huggingface.co/wan-model/wan22-vae, 2024.

[51] A. Tragoudaras, C. Zhang, D. Cherniavskii, A. Vozikis, T. Nijdam, D. W. E. Prinzhorn, M. Bodracska, N. Sebe, A. Zadaianchuk, and E. Gavves. Evaluating newtonian mechanics in video generative models with real physical systems, 2026. URL https://arxiv.org/abs/2504.02918.

[52] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng, J. Wang, J. Zhang, J. Zhou, J. Wang, J. Chen, K. Zhu, K. Zhao, K. Yan, L. Huang, M. Feng, N. Zhang, P. Li, P. Wu, R. Chu, R. Feng, S. Zhang, S. Sun, T. Fang, T. Wang, T. Gui, T. Weng, T. Shen, W. Lin, W. Wang, W. Wang, W. Zhou, W. Wang, W. Shen, W. Yu, X. Shi, X. Huang, X. Xu, Y. Kou, Y. Lv, Y. Li, Y. Liu, Y. Wang, Y. Zhang, Y. Huang, Y. Li, Y. Wu, Y. Liu, Y. Pan, Y. Zheng, Y. Hong, Y. Shi, Y. Feng, Z. Jiang, Z. Han, Z.-F. Wu, and Z. Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

[53] P. Wang, W. Wang, and Q. Li. Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection, 2025. URL https://arxiv.org/abs/2511.03997.

[54] Z. Wang, K. Wang, and L. Zhang. Phydetex: Detecting and explaining the physical plausibility of t2v models, 2025. URL https://arxiv.org/abs/2512.01843.

[55] L. Weihs, A. R. Yuile, R. Baillargeon, C. Fisher, G. Marcus, R. Mottaghi, and A. Kembhavi. Benchmarking progress to infant-level physical reasoning in ai. *TMLR*, 2022.

[56] J. Xu, Y. Huang, J. Cheng, Y. Yang, J. Xu, Y. Wang, W. Duan, S. Yang, Q. Jin, S. Li, J. Teng, Z. Yang, W. Zheng, X. Liu, D. Zhang, M. Ding, X. Zhang, X. Gu, S. Huang, M. Huang, J. Tang, and Y. Dong. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation, 2026. URL https://arxiv.org/abs/2412.21059.

[57] Z. Xue, M. Luo, and K. Grauman. Seeing the arrow of time in large multimodal models, 2025. URL https://arxiv.org/abs/2506.03340.

[58] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, D. Yin, Y. Zhang, W. Wang, Y. Cheng, B. Xu, X. Gu, Y. Dong, and J. Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025. URL https://arxiv.org/abs/2408.06072.

[59] S. Yasuda, T. Yates, and I. Yildirim. Physical event representations: Observers spontaneously impose discrete temporal structure in intuitive physical scene understanding. *Journal of Vision*, 21(9):2672, 2021. doi: 10.1167/jov.21.9.2672.

[60] J. Yuan, X. Zhang, F. Friedrich, N. Beltran-Velez, M. Hall, R. Askari-Hemmat, X. Han, N. Ballas, M. Drozdzal, and A. Romero-Soriano. Inference-time physics alignment of video generative models with latent world models, 2026. URL https://arxiv.org/abs/2601.10553.

[61] F. Zhou, J. Huang, J. Li, D. Ramanan, and H. Shi. Pai-bench: A comprehensive benchmark for physical ai, 2025. URL https://arxiv.org/abs/2512.01989.

A Appendix

A.1 More Qualitative Results

Figures 6 and 7 present additional qualitative examples of PHYPROBE on both image-to-video (I2V) and text-to-video (T2V) generation benchmarks. These examples illustrate that the predicted violation scores correlate well with the severity of perceived physical inconsistencies across diverse scenes, generators, and prompting settings. In particular, Figure 7 demonstrates that PHYPROBE can compare physical consistency even when videos have no semantic correspondence, supporting its use in heterogeneous pairwise evaluation.

Figure 8 presents representative failure cases. While PHYPROBE generally captures physically implausible motion, it can occasionally assign elevated violation scores to videos containing unusual but physically valid dynamics, and may underdetect violations that occur only momentarily. These examples highlight directions for improving long-horizon temporal representations and more comprehensive physical supervision.

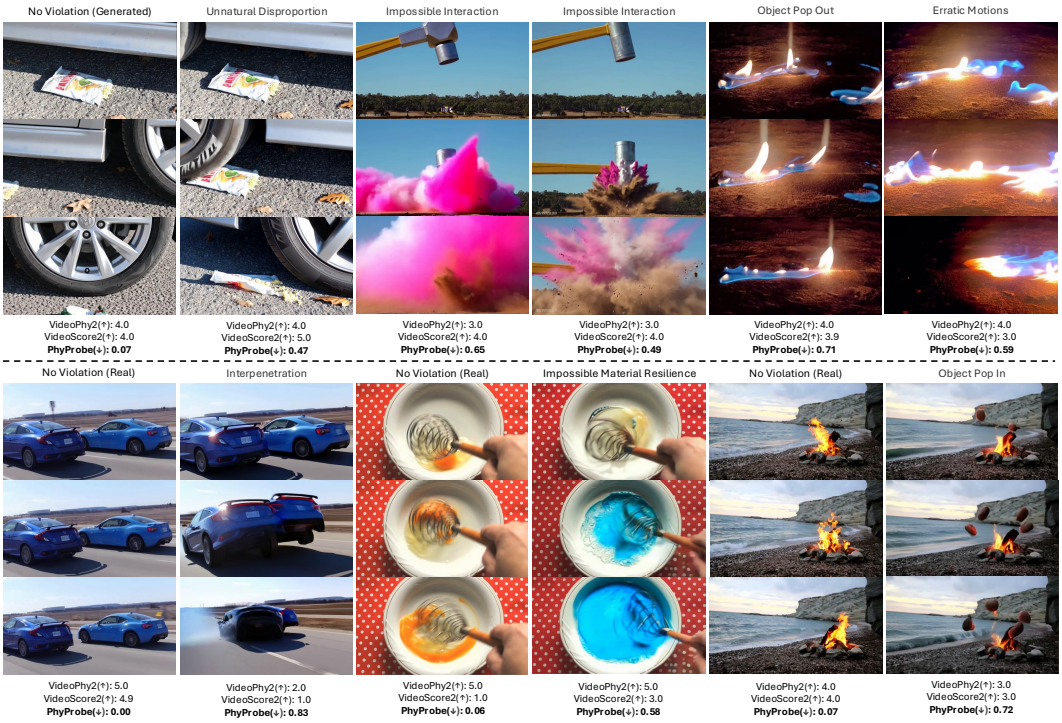


Figure 6: Qualitative comparison of image-to-video (I2V) generations with varying levels of physical consistency. PHYPROBE assigns low violation scores to videos with coherent object motion and interaction dynamics, while assigning higher scores to videos exhibiting implausible behaviors such as inconsistent motion, object deformation, or physical discontinuities. These examples illustrate that PHYPROBE captures continuous variation in physical violation severity beyond binary plausible/implausible judgments.

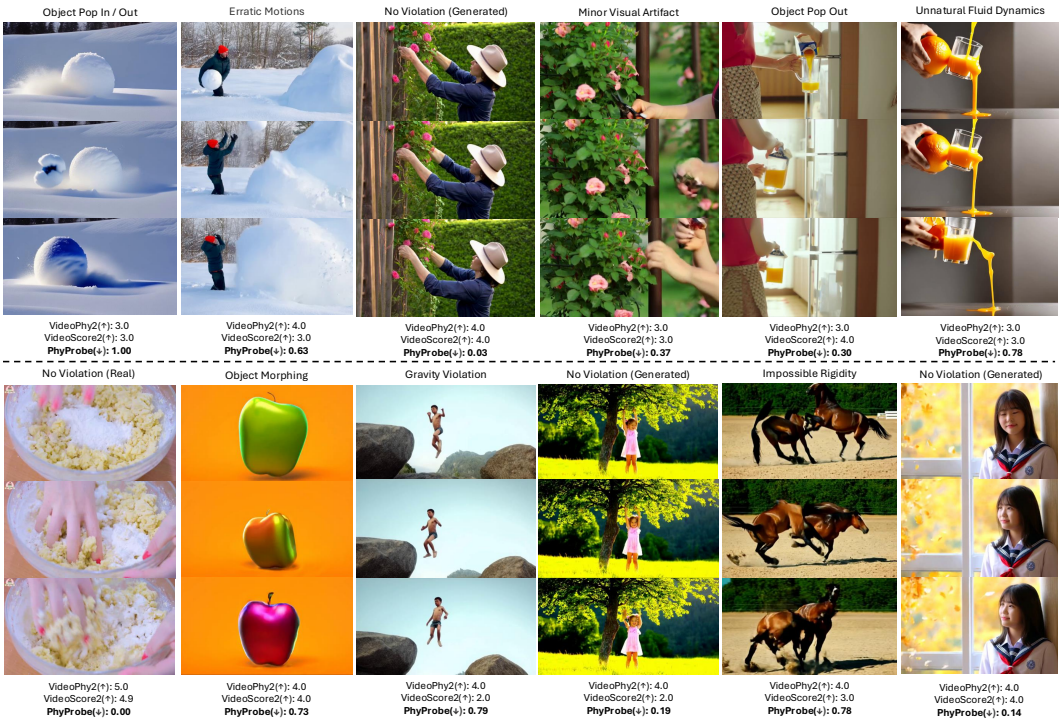


Figure 7: Qualitative comparison of text-to-video (T2V) generations under no semantic correspondence. Despite depicting entirely different scenes and prompts, PHYPROBE is still able to form meaningful pairwise physical consistency comparisons by evaluating the severity of physical violations directly from video dynamics. Videos with coherent motion and interaction patterns receive lower violation scores, while videos exhibiting implausible motion, object deformation, or inconsistent dynamics receive higher scores.

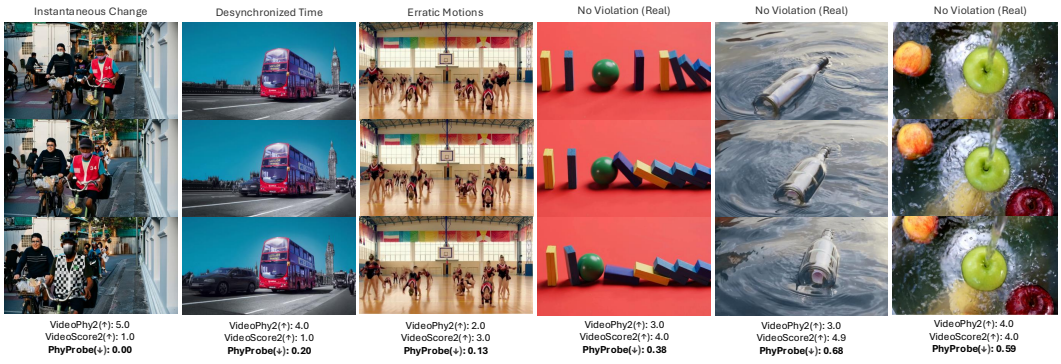


Figure 8: Failure cases of PHYPROBE. In some cases, PHYPROBE assigns elevated violation scores to real videos containing unusual or rarely observed motion patterns, suggesting limitations in the underlying representation space and training distribution. We also observe reduced sensitivity to violations that occur extremely briefly or instantaneously (e.g., sudden appearance, disappearance, or abrupt state changes), where the anomaly may not persist long enough to be strongly encoded in the aggregated video representation.

A.2 More Ablations

▷ **Effect of the PHYPROBE training objective under the same backbone.** In Table 7 we show PHYPROBE beats V-JEPA2 Suprise baseline within the same backbone, and therefore shown PHYPROBE does not solely attribute the gain to a stronger backbone representation.

Table 7: We compare PHYPROBE with V-JEPA2 suprise. Both methods use the identical V-JEPA2 ViT-H/16 backbone. PHYPROBE consistently improves over the original V-JEPA2 Surprise baseline, indicating that the gains are not solely attributable to stronger backbone representations.

		Pairwise Accuracy (%) ↑				**Human Correlation ↑**			
Method	**Backbone**	ImplB^p	VP2^p	PhyDEx^p	VF2^p	ρ_{VP2}	r_{VP2}	ρ_{VF2}	r_{VF2}
V-JEPA2 Surprise	V-JEPA2 ViT-H/16	33.3	46.3	49.0	45.5	0.065	0.065	0.121	0.133
PHYPROBE	V-JEPA2 ViT-H/16	**81.4**	**60.2**	**70.7**	**76.3**	**0.301**	**0.303**	**0.623**	**0.630**
Δ	–	+48.1	+13.9	+21.7	+30.8	+0.236	+0.238	+0.502	+0.497

▷ **Effect of the minimum score gap in evaluation set design.** We further analyze the effect of the minimum ground-truth score gap used to construct the evaluation pairs. As shown in Table 8, PHYPROBE consistently outperforms existing baselines across different minimum-gap thresholds, demonstrating that its advantage is robust to variations in evaluation set construction rather than being tied to a particular pair-selection criterion.

Table 8: Pairwise accuracy (%) on the **text-correspondence subset** of VP2^p under different minimum ground-truth score-gap thresholds. Only video pairs generated from the same text prompt are considered, following the pairwise evaluation protocol; equal-score pairs are excluded. The same checkpoint for each method is evaluated across all rows, while only the evaluation subset is changed.

Minimum	**# Test**	**Pairwise Accuracy (%)**			
Score Gap	**Pairs**	**VideoScore**	**VideoPhy2-AutoEval**	**V-JEPA2 Surprise**	**PHYPROBE**
$\Delta_{\mathrm{eval}} \geq 1$	1280	49.8	28.5	46.3	**66.1**
$\Delta_{\mathrm{eval}} \geq 2$	628	52.0	32.8	44.6	**71.2**
$\Delta_{\mathrm{eval}} \geq 3$	202	52.0	34.2	46.0	**77.2**
$\Delta_{\mathrm{eval}} \geq 4$	35	37.1	45.7	48.6	**74.3**

▷ **Leave-one-dataset-out (LODO) analysis.** To evaluate whether PHYPROBE relies on a particular training source, we retrain the model while excluding TRAVL, VideoPhy2, or VideoFeedback2 from training. As shown in Table 9, performance degrades on the held-out benchmark and on VP2^p, while remaining strong on the others. The results show that useful signals transfer, but the substantial corresponding-benchmark drops also demonstrate meaningful dataset-specific dependence.

▷ **Effect of score normalization.** We compare the proposed violation-flooring normalization with a simple linear normalization while keeping all other training settings unchanged. As shown in Table 10, the proposed normalization consistently improves pairwise accuracy across all four physical video benchmarks, suggesting that preserving greater score resolution in the low-violation regime provides more effective supervision. As preliminary supporting evidence, we also conduct a small blinded human preference study comparing the predicted score differences produced by the two normalization schemes. As shown in Table 11, annotators leaned toward the proposed normalization.

▷ **Distribution of predicted scores.** Figure 9 compares the predicted score distributions of PHYPROBE and VideoPhy2-AutoEval on ImplB^p and VF2^p. PHYPROBE produces substantially better separation between plausible and implausible videos, whereas the baseline exhibits considerable overlap between the two distributions. This indicates that the learned violation scores provide improved discriminability for physical plausibility.

A.3 Definition of Physical Consistency

"Physical Consistency" can be defined from different perspectives. Here we follow the operational definition adopted by existing public benchmarks [5, 20, 47], where it is approximated by perceived physical plausibility. While the exact operationalization varies slightly across datasets, they all rely on human judgments of whether physical laws appear to be violated. This variability further motivates aggregating heterogeneous supervision from multiple complementary sources under a unified training

Table 9: **Leave-one-dataset-out (LODO) analysis.** We retrain PHYPROBE while excluding one training dataset (TRAVL, VideoPhy2, or VideoFeedback2) at a time and evaluate pairwise accuracy (%) across four physical video benchmarks. While performance degrades on the corresponding held-out benchmark and, in every case, on VP2^p; it remains strong on the others, indicating partial transfer across sources alongside meaningful dependence on each dataset.

Method	ImplB^p	VP2^p	PhyDEx^p	VF2^p
VP2-AutoEval	28.0	28.5	30.5	36.3
V-JEPA2 Surprise	33.3	46.3	49.0	45.5
VideoScore2	68.7	49.7	60.6	65.0
PHYPROBE	**98.0**	**66.1**	**98.9**	**75.2**
PHYPROBE (w/o TRAVL)	92.1	58.7	98.2	74.7
Δ (w/o TRAVL $-$ full)	-5.9	-7.4	-0.7	-0.5
PHYPROBE (w/o VideoPhy2)	96.1	58.3	99.3	69.5
Δ (w/o VideoPhy2 $-$ full)	-1.9	-7.8	$+0.4$	-5.7
PHYPROBE (w/o VideoFeedback2)	92.8	57.1	99.7	58.0
Δ (w/o VideoFeedback2 $-$ full)	-5.2	-9.0	$+0.8$	-17.2

Table 10: **Effect of score normalization on PHYPROBE.** We compare the proposed violation-flooring normalization against a simple linear normalization while keeping all other training settings identical. Results report pairwise accuracy (%) on four physical video benchmarks.

Method	ImplB^p	VP2^p	PhyDEx^p	VF2^p
PHYPROBE(linear)	96.7	60.0	96.5	71.0
PHYPROBE	**98.0**	**66.1**	**98.9**	**75.2**
Δ (linear $-$ proposed)	-1.3	-6.1	-2.4	-4.2

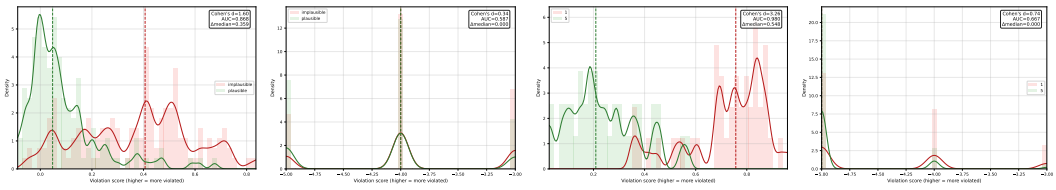
objective. Since physical plausibility and visual quality are often intertwined, human judgments naturally reflect both aspects. In essence, PhyProbe is trained to approximate the degree to which a human would perceive physical laws to be violated.

Common Video Abnormalities. We categorize common abnormalities in generated videos into four groups. Although these categories frequently co-occur with perceived physical inconsistency, not every instance necessarily constitutes a violation of physical laws.

- **Perception abnormalities** affect visual fidelity and appearance without necessarily violating physical laws, including artifacts such as noise, flickering lighting, repeated scenery, missing details, and unnatural color palettes.
- **Objectness abnormalities** relate to the existence, shape, and spatial integrity of entities, such as objects appearing or disappearing, distorted proportions, impossible poses, sudden morphing, or inconsistent spatial layouts.
- **Behavioral abnormalities** capture physically implausible interactions and material properties, such as object interpenetration, unrealistic fluid dynamics, impossible deformation or resilience, and erratic motion that violates expected force and collision dynamics.
- **Time abnormalities** involve disruptions in temporal consistency, including desynchronized motion, instantaneous corrections, abrupt start or stop behaviors, and violations of causal ordering.

Table 11: **Blinded human preference evaluation of scalar target normalization.** We compare the proposed violation-flooring normalization against a simple linear normalization using a blinded preference evaluation. Annotators were shown the predicted score differences from both models in random order and selected which better reflected the perceived difference in physical violation severity. Each of the n pairs was independently rated by four non-expert annotators; percentages are pooled over all annotator judgments. Given the small sample ($n = 30$ pairs) and inter-rater agreement of Fleiss' $\kappa = 0.148$, we regard this as preliminary supporting evidence of a preference trend.

Subset	**Linear**	**Proposed**
ImplB^p ($n = 10$)	17.5%	**82.5%**
VF2^p ($n = 10$)	37.5%	**62.5%**
VP2^p ($n = 10$)	25.0%	**75.0%**
Overall ($n = 30$)	26.7%	**73.3%**
Fleiss' κ	0.148 (4 raters)	
Unanimous agreement	14/30 pairs (46.7%)	



(a) PHYPROBEon ImplB^p (b) VP2 on ImplB^p (c) PHYPROBEon VF2^p (d) VP2 on VF2^p

Figure 9: **Predicted score distributions for plausible and implausible videos.** Score distributions produced by PHYPROBE and VideoPhy2-AutoEval (VP2) on ImplB^p (a,b) and VF2^p (c,d). For VF2^p, only samples with human scores of 1 and 5 are shown, corresponding to clearly implausible and plausible videos, respectively, to reduce ambiguity from intermediate ratings.

Table 12: Sensitivity of PHYPROBE "w/ tie" accuracy (%) to the model-side tie threshold, expressed as a percentile of absolute score differences on each benchmark. The 20th percentile is used for PHYPROBE results in Table 3. "w/o tie" columns compare raw scores without a threshold and are unaffected.

Percentile	5	10	20	30	40
MonetBench	69.8	71.4	**72.7**	71.8	72.4
RewardBench Overall	63.2	63.6	64.1	64.6	**64.8**
RewardBench MQ	57.9	59.4	62.3	65.0	**67.9**
RewardBench VQ	58.0	59.3	61.4	63.4	**65.6**

B Implementation Details

B.1 Dataset Construction

Dataset Sourcing. We constructed a diverse dataset from prior video physics evaluation works, including VideoScore2 [20], VideoPhy2 [6], BrokenVideos [28], PAI-Bench [61], TRAVL [36], and ImpossibleVideos [2]. These datasets contains generated videos originate from a diverse set of modern video generation systems, including proprietary and open-source Text-to-Video (T2V) and Image-to-Video (I2V) models, such as CogVideoX [58], Wan2.1 [52], Sora [42], Ray2 [31], VideoCrafter2 [12], Veo [14], etc. [49, 41, 35, 25]. Such diversity reduces the reliance on model-specific artifacts.

Data Preprocessing. Each video is resized to a fixed square resolution while preserving the original aspect ratio via letterboxing, filling the empty bars with mid-gray (pixel value 0.5). Frames are sampled uniformly in time at 16 fps, yielding $T = 48$ frames per clip.

Score Assignment and Normalization. For VideoScore2 [20] and VideoPhy2 [6], we assume the 1–5 Likert scales for physical plausibility share a comparable semantic interpretation across datasets. Since PHYPROBE models violation severity rather than plausibility, we invert and normalize these annotations into the range $[0, 1]$, where 0 denotes physically consistent motion and 1 denotes severe physical violation. Motivated by the thresholded nature of human perception reported in related settings [39, 40, 38] , we apply a violation-flooring scheme; the specific mapping is a design choice, supported empirically by the ablation in Table 10 and, preliminarily, by the small preference study in Table 11. Specifically, PC=5 is mapped to 0.0, PC=4 to 0.5, PC=3 to 0.667, PC=2 to 0.833, and PC=1 to 1.0. Under this formulation, videos judged as largely plausible by humans may still contain subtle violations below the perceptual threshold, while increasingly implausible ratings are compressed toward the high-violation regime. For BrokenVideos [28], which provides spatial-temporal mask annotations for visual and physical errors, we derive approximate violation scores from mask statistics including spatial area, temporal persistence, and proximity to the image center. Since these annotations primarily correspond to visibly salient failures, we apply a similar flooring strategy and map the resulting scores to the range $[0.5, 0.9]$. For Kinetics-700 [11], PAI-Bench [61], TRAVL [36], and ImpossibleVideos [2], we use anchor-based supervision. Real videos are assigned a target score of 0, while synthetic failure cases from ImpossibleVideos and TRAVL are anchored toward 1, providing approximate endpoints for calibration of the violation scale. The overall supervision regime is illustrated in Figure 10.

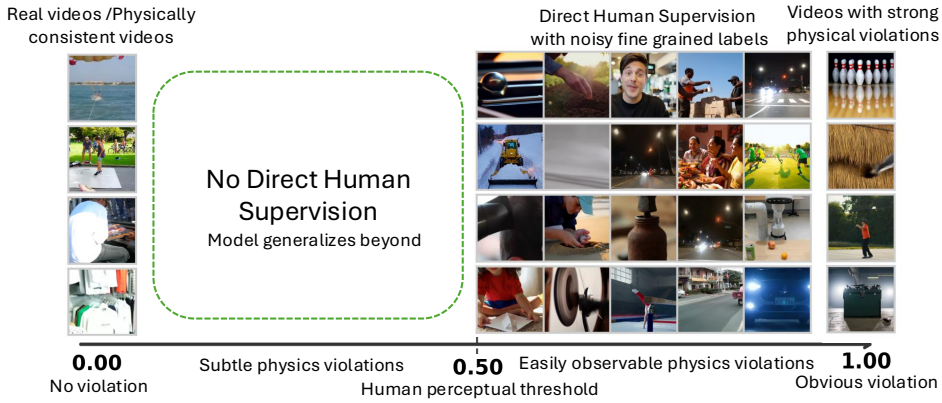


Figure 10: Illustration of the supervision regime for physical plausibility evaluation. Human annotations primarily capture clearly observable violations (scores ≥ 0.5), while the region below the perceptual threshold remains largely unsupervised. Importantly, scores in the range $[0, 0.5)$ do not indicate the absence of violations, but rather subtle or weakly perceptible physical inconsistencies. PHYPROBE produces differentiated scores in this region, but their interpretation remains unvalidated due to the absence of direct severity annotations. Figure 5 and Table 10 provide indirect evidence that finer resolution in the low-violation regime improves supervision.

B.1.1 Construction Detail of Training Datasets and Supervision Labels

The main paper summarizes which supervision signals are provided by each dataset (Table 1). Here, we describe how the corresponding supervision labels are constructed and how each dataset is converted into training examples.

- **PAI-Bench.** PAI-Bench [61] contains Image-to-Video model-generated videos paired with human-recorded reference videos **sharing the same image prompt input**. For each reference video, we sample up to 50 model-generated videos as negative partners, re-sampling across training epochs to diversify pair coverage. The reference video is always assigned as the winner, providing a preference learning (**rank signal**) that pushes model-generated videos below their human counterparts. We also anchor the reference videos with **calibration signal** as reference videos are physically consistent.
- **VideoPhy2.** VideoPhy2 [6] annotates videos with a physical correctness (PC) score on a 1–5 integer scale. We form pairs **within the same action category** and assign the video with the higher PC score as the winner, treating the raw scores as continuous regression targets (**magnitude signal**) and also preference learning targets (**rank signal**).
- **VideoFeedback2.** VideoFeedback2 [20] similarly provides physical correctness (PC) score (1–5) at the prompt level. We pair videos **without image or text prompt correspondence** and retain only pairs whose PC scores differ by at least 1, filtering out near-tie comparisons whose ordering is less reliable. The video with the higher PC score is the winner, treating the raw scores as continuous regression targets (**magnitude signal**) and also preference learning targets (**rank signal**).
- **BrokenVideos.** BrokenVideos [28] feature videos with collective failures, ranging from visuals to physical violations. It also contains videos exhibiting visual artifacts detected via segmentation masks. We derive a composite anomaly score $s = 0.6 \cdot \text{mask_area} + 0.3 \cdot \text{center_proximity} + 0.1 \cdot \text{mask_presence_ratio}$, where lower scores indicate fewer artifacts. We create pairs **without image or text prompt correspondence** and retain a score gap ≥ 0.4, which exposes the model to a broader distribution of artifact severity, treating the scores as continuous regression targets (**magnitude signal**) and also preference learning targets (**rank signal**).
- **ImpossibleVideos.** ImpossibleVideos [2] features counterfactual scenes that violate physical, biological, geographical laws, and it contains physically implausible AI-generated videos against real-world videos. Pairs are then formed globally between plausible and implausible videos **without image or text prompt correspondence**, with the plausible video always assigned as the winner. It act as a preference learning pair to provide **rank signal** and we also anchor the plausible videos and implausible videos with **calibration signal** as plausible videos are physically consistent and implausible videos are systematic failures.
- **TRAVL.** TRAVL [36] contains videos annotated via multiple question-answer pairs, each carrying a per-question plausibility label (0 = plausible, 1 = implausible). We derive a per-video plausibility label by majority vote across all QA pairs, discarding videos with tied votes. Pairs are then formed globally between plausible and implausible videos **without image or text prompt correspondence**, with the plausible video always assigned as the winner. It act as a preference learning pair to provide **rank signal** and we also anchor the plausible videos and implausible videos with **calibration signal** as plausible videos are physically consistent and implausible videos are systematic failures. This dataset is complementary to ImpossibleVideos: both provide a binary plausible-vs-implausible signal, but TRAVL's labels are grounded in explicit physical reasoning questions rather than direct human annotation.
- **Kinetics-700.** We form pairs **within the same action category** in the Kinetics-700 dataset [11]. Since both videos are real and physically consistent, pairs are treated as ties under the **ranking signal**. This provides a real-video anchor for the model to learn the **calibration signal**.

B.1.2 Construction Detail of Evaluation Datasets and Ground Truth Labels

For every benchmark below, the model scores each video independently; pairwise accuracy and correlation are computed from those per-video scores. For benchmarks with human tie labels, PHYPROBE declares a tie when $|s(v_A) - s(v_B)|$ falls below the 20th percentile of absolute score differences on that benchmark. The same percentile is used for all benchmarks; its effect is reported

in Table 12. The "w/o tie" columns of Table 3 compare raw scores without a threshold and are unaffected by it.

Physical Domain

- **ImplB^p.** The official ImplausiBench [36] contains 300 videos (150 real and 150 generated by image-to-video models conditioned on the real video's first frame) with gold-standard human judgments. We pair the real and its corresponding generated one to form correspondence pairs.
- **VP2^p.** We derive the pairwise dataset from the official VideoPhy2-test dataset [6]. Each video in official VideoPhy2-test has a rating between 1 to 5 rated by a human annotator in physical commonsense. We kept text prompt corresponding pairs based on rating gap ≥ 1 as pairs. Equal-score pairs are excluded. Total 1280 pairs.
- **PhyDEx^p.** We derive the pairwise dataset from PhyDetEx's PID-test dataset [54] which contains 250 real and 250 generated videos. We sample and form 2000 pairs based on a real video paired with a generated one as pairs without image or text correspondence.
- **VF2^p.** We derive the pairwise dataset from the official VideoFeedback2-test dataset [20]. Each video has a rating between 1 to 5 rated by a human annotator in physical commonsense. We form pairs based on rating gap ≥ 1 as pairs without image or text correspondence. Equal-score pairs are excluded. 2000 pairs are formed.
- **Human Correlation VP2.** The official VideoPhy2-test set released by VideoPhy2 is used for human correlation [6]. Total instances 3397, each video has a rating between 1 to 5 rated by a human annotator in physical commonsense. We use model to score video individually for computing the human correlation.
- **Human Correlation VF2.** The official VideoFeedback2-test released by VideoScore2 is used for human correlation [20]. Total instances 500, each video has a rating between 1 to 5 rated by a human annotator in physical commonsense. We use model to score video individually for computing the human correlation.

Quality Domain

- **MonetBench.** MonetBench is the official pairwise test set from VisionReward [56], which each pair is a human preference pair.
- **Rapidata-I2V.** Rapidata-I2V is a pairwise preference dataset released on HuggingFace datasets by Rapidata [45, 46], which each pair is a human preference pair. The videos come from recent generators such as Seedance1Pro and Hailuo 2.0.
- **RewardBench.** VideoGen-RewardBench is the official pairwise test set from VideoReward [29], which each pair is a human preference pair.

B.2 Model Training

Training. We use pretrained video encoders as feature extractor, candidates include Wan 2.2 VAE [50], VJEPA2 [1], and Perception Encoder [8]. We found Perception Encoder Core G14-448 gives the best results. The encoder is kept frozen during training. On top of the representation, we train a lightweight scoring head $g(\cdot)$ implemented as linear layers with statistic aggregation head. The scoring head outputs an unconstrained value $\tilde{s}(v) \in \mathbb{R}$; the reported violation score is $s(v) = \text{clip}(\tilde{s}(v), 0, 1)$, applied at inference only. Training is performed using the combined objective Equation 5. We set $\lambda_{\text{mag}} = 10$ and $\lambda_{\text{cal}} = 100$. PHYPROBE can be trained within 24 hours using 4 NVIDIA H100 GPUs using batch size = 8 for 10K steps. To balance heterogeneous data sources, we apply dataset-level sampling weights to prevent large datasets from dominating the learning signal. In detail, per-step sampling probabilities are: PAI-Bench 0.059, VideoPhy2 0.176, VideoFeedback2 0.294, BrokenVideos 0.118, ImpossibleVideos 0.059, TRAVL 0.176, Kinetics-700 0.118.

Training Techniques. To stabilize training, we adopt several complementary strategies. The projection head is optimized with AdamW [30] with learning 3×10^{-4}, weight decay = 0.01, and with a linear warmup over 500 steps followed by cosine annealing to zero over 10K total steps. To obtain a stable inference model, we maintain an Exponential Moving Average (EMA) of the projection head weights with decay 0.9999, and use the EMA checkpoint for all evaluations. When computing the ranking loss, we apply label smoothing whose strength depends on the annotated score gap. For a pair (v_i, v_j) with gap $g_{ij} = |y_{v_i} - y_{v_j}|$, we set $c_{ij} = \max(0.1,\ g_{ij} / \max_{(k,l)} g_{kl})$,

with the maximum over pairs from the same dataset in the current batch, and use smoothing $\epsilon_{ij} = 0.10 - 0.05\, c_{ij} \in [0.05, 0.095]$: the largest-gap pair in a batch slice receives $\epsilon = 0.05$, and the floor on c_{ij} keeps every pair contributing, with near-tied pairs smoothed at $\epsilon = 0.095$. Pairs without scalar annotations use $c_{ij} = 1$. Regression targets use a Huber loss with $\delta = 0.5$ to reduce sensitivity to annotation noise, while the hard anchor targets use MSE with a higher loss coefficient to enforce the score range boundaries firmly. Finally, the pair pool for each dataset is resampled every 2.5K steps to expose the model to a fresh set of contrasting video pairs throughout training.

B.3 Evaluation

Baselines. All baselines, including VideoPhy2 [6], VideoScore2 [20], VisionReward [56], VideoReward [29], were directly sourced from their publicly released code on github and their provided prompt template. For VJepa2-Suprise score, we follow the implementation as described in WMReward [60] using the pretained VJEPA-2 model [1]. For Gemini-3.1-Flash-Lite, we use the offical API call and adopt the prompt template from Figure 11 for accuracy and Figure 12 for correlation.

As some of the baselines contain scores in multiple aspects, we pick the scores that is related with physical consistency or visual quality. We compared our method against the physical commonsense score (1-5) from VideoPhy2 and VideoScore2, and compared against the Visual Quality (VQ) and Motion Quality (MQ) from VisionReward and VideoReward.

Benchmarks. As we are computing the pairwise accuracy, we reorder the existing benchmark datasets and form pairs by matching their correspondence, as shown in Table 13. For perceptual quality benchmark, we adopt the existing benchmarks directly as they are already in pairwise form. Details

Table 13: Evaluation benchmarks used in this work, covering both **physical plausibility** and **perceptual quality** under diverse settings of correspondence and pair types. Physical benchmarks focus on consistency of dynamics, while quality benchmarks evaluate alignment with human preferences.

Domain	Eval Dataset	Pairs	Type	Corres.	Desc.
Physical	ImplausiBench [36]	150	R–G	Img+Txt	Paired with real and physical violations
	PhyDetEx [54]	2K	R–G	None	Diverse real vs synthetic violations
	VideoPhy2-Test [6]	1.2K	G–G	Txt	Paired with score differences
	VideoFeedback2-Test [20]	2K	G–G	None	Paired with score differences
Quality	MonetBench [56]	1K	G–G	Txt	Human preference on video quality
	Rapidata-I2V [45, 46]	295	G–G	Img+Txt	Human preference on video quality
	RewardBench [29]	25K	G–G	Txt	Human preference on video quality

Human Correlation. We reported the human correlation from the official VideoPhy2 [6] test set and VideoFeedback2 [20] test set.

B.4 Reproducibility

To ensure reproducibility, all stochastic components of our pipeline are controlled via a fixed global seed of 42. All evaluation protocols, including correlation scoring and automated pairwise comparison are initialized with the same seed. For pairwise evaluation, we randomize the A/B presentation order to mitigate position bias; critically, all swap assignments are pre-computed from a seeded RNG prior to dispatching any worker threads, ensuring that results are invariant to worker count and non-deterministic task completion order.

Physical Consistency Pairwise Evaluation Prompt

```
You are evaluating the physical consistency of two generated videos.
Physical consistency refers to whether objects move, deform, interact,
and persist in ways that respect the constraints of the physical world.
Focus ONLY on physical consistency. Ignore visual quality, aesthetics,
text rendering, or semantic coherence.

Physical violations include:
- objects defying gravity
- passing through surfaces
- impossible deformations
- discontinuities in motion
- violations of object permanence
- implausible contact or collision
- fluids/cloth/hair behaving impossibly

Video A:
<video>{video_a_base64}</video>
Video B:
<video>{video_b_base64}</video>

Which video is more physically consistent?
**Vote:** [A or B]
**Justification:** [1-2 sentences]
```

Figure 11: Prompt template for pairwise physical consistency evaluation between two generated videos.

Physical Consistency Scoring Prompt

```
You are evaluating the physical consistency of a generated video.
Physical consistency refers to whether objects move, deform,
interact, and persist in ways that respect the constraints of
the physical world.
Watch this video carefully and evaluate its physical plausibility.
Score the video on a scale from 1 to 5:
    1 - severe violations of physical laws
        (impossible motion, objects, forces)
    2 - major physical implausibilities
    3 - some implausible elements but mostly reasonable
    4 - minor implausibilities, mostly realistic physics
    5 - fully physically plausible, natural and realistic
        throughout

Respond with a JSON object containing a single key "score"
with an integer value 1-5.
```

Figure 12: Prompt template for scalar physical consistency scoring of a single video.

C Data Study

Concerns of train / test overlap. Supervision instances are disjoint between training and test sets for every benchmark (Table 14 and 15). However, VideoFeedback2-test shares all prompts and generators with VideoFeedback2-train, so results on VF2 reflect held-out annotations under the same prompt and generator distribution rather than a distribution shift. VideoPhy2-test, by contrast, shares no prompts and only partially overlaps in generators.

Table 14: Train/test overlap between each evaluation benchmark and the complete training corpus used by PHYPROBE. Entries are reported as X/Y, where X denotes the number of unique evaluation video-label entities that also appear in the complete training corpus and Y denotes the total number of unique evaluation entities. Supervision overlap is computed over annotated supervision instances (e.g., preference labels, quality scores, or pairwise comparisons). Generation Prompt overlap is computed over unique conditioning inputs used to generate the evaluation videos (i.e., text prompts for text-to-video tasks and image–text input pairs for image-conditioned tasks). Generator overlap is computed over unique video generation models. Generator overlap does not imply supervision or train–test sample overlap.

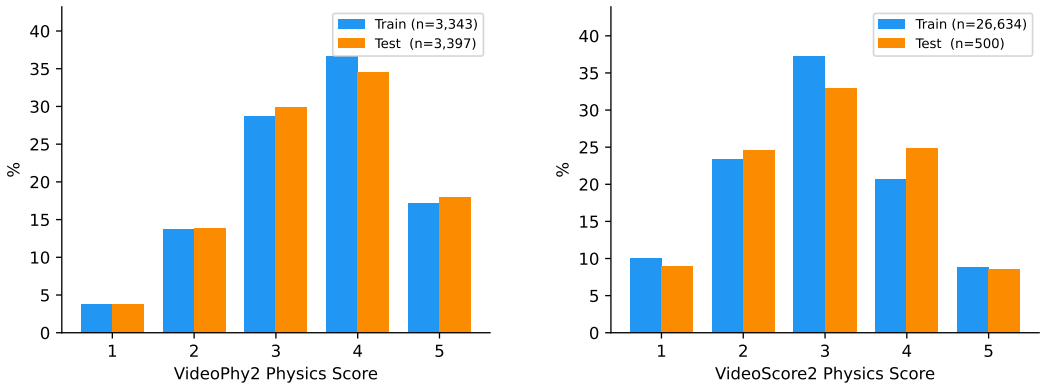
Evaluation Benchmark	Supervision overlap	Generation Prompt overlap	Generator overlap
Physical Domain			
ImplB^p(ImplausiBench)	0/300	Unavailable	Unavailable
VP2^p (VideoPhy2-test)	0/2888	0/599	2/5 (1172/2943 samples)
PhyDEx^p (PhyDetEx)	0/398	0/247	2/3 (55/250 samples)
VF2^p (VideoFeedback2-test)	0/500	500/500	25/25 (500/500 samples)
Quality Domain			
MonetBench	0/2000	2/2000	Unavailable
Rapidata-I2V	0/298	0/298	0/3 (0/298 samples)
RewardBench	0/4691	Unavailable	12/12 (4691/4691 samples)

Table 15: Official train/test overlap for the benchmark datasets used in PHYPROBE. Supervision overlap denotes identical annotated supervision instances shared between the official training and evaluation splits. Generation Prompt overlap denotes identical conditioning inputs (i.e., text prompts or image–text input pairs). Generator overlap denotes shared video generation models and is reported only when generator metadata is available. Note that VideoFeedback2's official splits share prompts and generators; supervision is disjoint, but the test set does not constitute a prompt- or generator-level distribution shift.

Training	Evaluation	Supervision overlap	Generation Prompt overlap	Generator overlap
VideoPhy2-train	VideoPhy2-test	0%	0%	43%
VideoFeedback2-train	VideoFeedback2-test	0%	100%	100%
TRAVL	ImplausiBench	0%	0%	Unavailable

Human Scoring Analysis in VideoPhy2 and VideoScore2. In Figure 13 We study the human annotations in both training and test data from VideoPhy2 [6] and VideoScore2 [20]. We found that Annotators consistently avoid extreme ratings and concentrate around intermediate scores, reflecting the coarse and thresholded nature of human perception toward physical inconsistencies. Anchor-based supervision serves as a mitigation strategy by providing stable reference points at the low- and high-violation ends of the spectrum, helping calibrate the absolute violation scale across heterogeneous datasets.

Heuristic Labels for BrokenVideos. The BrokenVideos dataset does not provide explicit pairwise preference annotations. We therefore construct heuristic preference labels based on automatically extracted anomaly statistics. Specifically, we compute a composite anomaly score from mask area, center proximity, and mask presence ratio (Table 16). To improve label reliability for cross-prompt comparisons, we retain only pairs whose anomaly scores differ by at least 0.4. As shown in Table 18, this threshold is approximately $2.9\times$ the interquartile range (IQR) of the score distribution, representing a conservative criterion that selects well-separated preference pairs. Finally, we validate the resulting heuristic labels through a human annotation study involving four independent annotators. Table 17 shows that the heuristic agrees with the human majority on 80.6% of non-tied pairs, while the observed heuristic–human agreement is comparable to the measured human–human agreement, supporting the use of the proposed heuristic as a reliable source of auxiliary supervision.



(a) VideoPhy2 Annotations

(b) VideoScore2 Annotations

Figure 13: **Human annotation distributions across different physical consistency evaluation datasets.** Annotators consistently avoid extreme ratings and concentrate around intermediate scores, reflecting the coarse and thresholded nature of human perception toward physical inconsistencies. Anchor-based supervision serves as a mitigation strategy by providing stable reference points at the low- and high-violation ends of the spectrum, helping calibrate the absolute violation scale across heterogeneous datasets.

Table 16: Distribution of per-video statistics in the BrokenVideos training set. Lower mask area, center proximity, and mask presence ratio indicate fewer localized anomalies. Annotation coverage ratio is reported only as a data quality statistic and is not used in the composite anomaly score.

Field	**Min**	**Q1**	**Median**	**Q3**	**P99**	**Max**
Mask area	0.00011	0.014	0.039	0.094	0.446	0.860
Center proximity	0.00	0.34	0.56	0.79	1.00	1.00
Mask presence ratio	0.021	1.00	1.00	1.00	1.00	1.00

Table 17: Human validation of the proposed heuristic-derived cross-prompt BrokenVideos preference labels design, as the score is designed as $0.6 \cdot$ mask_area $+ 0.3 \cdot$ center_proximity $+ 0.1 \cdot$ mask_presence_ratio. Four independent non-expert annotators evaluated the same set of 85 sampled video pairs.

Metric	**Value**
Heuristic vs. Human	
Annotator 1 agreement	77.6%
Annotator 2 agreement	74.1%
Annotator 3 agreement	77.6%
Annotator 4 agreement	65.9%
Human majority agreement	80.6% (58/72)
Human vs. Human	
Mean pairwise agreement	77.5%
Pairwise agreement range	72–89%
Fleiss' κ	0.42
Unanimous Cases	
4/4 unanimous pairs	51 / 85
Heuristic agreement on unanimous pairs	86.3% (44/51)

Table 18: Distribution of the composite anomaly score used for BrokenVideos pair construction. The score is computed as $0.6 \cdot$ mask_area $+ 0.3 \cdot$ center_proximity $+ 0.1 \cdot$ mask_presence_ratio. The interquartile range (IQR) characterizes the intrinsic spread of the heuristic score. The adopted cross-prompt threshold (0.4) is approximately $2.9\times$ the IQR, making it a conservative threshold that retains only well-separated preference pairs.

Min	**Q1**	**Median**	**Q3**	**IQR**	**Max**
0.012	0.24	0.31	0.38	0.14	0.699

D Baseline Study

Tie-aware analysis. We further performed a tie-aware pairwise analysis of VideoPhy2-AutoEval and VideoScore2 (Table 19). While VideoPhy2-AutoEval achieves higher accuracy after excluding tied pairs on several benchmarks, it predicts ties substantially more frequently than VideoScore2 across nearly all evaluation sets. This behavior is particularly pronounced on Rapidata-I2V, where VideoPhy2-AutoEval exhibits an 81.4% tie rate, leaving relatively few non-tied comparisons and reducing its overall pairwise accuracy (9.5% vs. 50.9% on non-tied pairs). Rapidata-I2V also contains three recently released video generators that are absent from the training data of both VideoPhy2-AutoEval and VideoScore2, which may partially explain the particularly high tie rate observed on this benchmark.

Table 19: **Tie-aware pairwise accuracy of VideoPhy2-AutoEval and VideoScore2 across evaluation benchmarks.** Both baselines output discrete physical-commonsense scores, so a pair receives a model tie whenever the two videos are assigned equal scores. *Model tie rate* is the fraction of pairs on which this occurs. *Acc (all pairs)* is computed over every pair, including human-labelled ties, with a model tie counted as correct only when the human label is also a tie; it corresponds to the "w/ tie" columns of Table 3. *Acc (model non-tied)* restricts evaluation to pairs on which the model produced a non-tied prediction, i.e., accuracy conditional on the model committing to a preference.

Eval Set	Model	Model tie rate	Acc (all pairs)	Acc (model non-tied)
VP2^p	VideoPhy2-AutoEval	58.8%	28.5%	69.1%
	VideoScore2	40.2%	49.7%	60.0%
VF2^p	VideoPhy2-AutoEval	46.0%	36.3%	67.2%
	VideoScore2	38.1%	65.0%	79.6%
ImplB^p	VideoPhy2-AutoEval	63.3%	28.0%	76.4%
	VideoScore2	43.3%	68.7%	75.3%
PhyDEx^p	VideoPhy2-AutoEval	40.0%	30.5%	50.8%
	VideoScore2	35.8%	60.6%	75.1%
Rapidata-I2V	VideoPhy2-AutoEval	**81.4%**	9.5%	50.9%
	VideoScore2	45.8%	45.8%	53.1%
MonetBench	VideoPhy2-AutoEval	69.4%	22.5%	49.0%
	VideoScore2	51.2%	42.3%	46.7%
RewardBench	VideoPhy2-AutoEval	64.1%	24.6%	54.9%
	VideoScore2	37.4%	57.4%	66.2%

E Societal Impact

Reliable evaluation of physical consistency in generated videos may help improve the safety, reliability, and realism of generative video systems by enabling better detection of implausible or unstable outputs. More robust physical evaluation could support applications in simulation, robotics, scientific visualization, education, and content generation, where physically inconsistent behavior may reduce trustworthiness or usability.

At the same time, advances in evaluation may indirectly contribute to improving the realism of generative models, including synthetic media that could be used for misinformation or deceptive content. However, PHYPROBE is designed as an evaluation framework rather than a generation system, and does not itself generate or manipulate media. We also note that physical plausibility should not be conflated with factual correctness or authenticity: a physically plausible video may still depict fabricated events.

F Safeguards

PHYPROBE is intended as a research evaluation framework for measuring physical consistency in generated videos, rather than a system for determining factual authenticity or detecting misinformation. We emphasize that physical plausibility should not be interpreted as evidence that a video depicts real events. In addition, PHYPROBE remains imperfect under distribution shift and may assign incorrect scores to unusual but physically valid motions or complex long-horizon interactions.

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: See Abstract, Section 1, and Section 4.

   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: See Section 5.

   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: [Yes]

   Justification: See Section 3.

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: See Appendix B.4.

   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 built our dataset based on public datasets from prior works. We will release our code and packaged dataset to public in the final camera ready paper.

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: See Appendix B.

   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: [Yes]

   Justification: See Table 2, where we reported our human correlation results are all statistically significant with $p < 0.0001$.

   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: See Appendix B.2.

   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: The research conforms in every respect with 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: See Appendix E.

    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: [Yes]

    Justification: See Appendix F.

    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 cited the original paper that produced the datasets. See Appendix B.

    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 will release our code and packaged dataset to public in the final camera ready paper, with proper documentations.

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: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

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: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

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: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

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.