3DSPA: A 3D SEMANTIC POINT AUTOENCODER FOR EVALUATING VIDEO REALISM

Anonymous authors
Paper under double-blind review

ABSTRACT

AI video generation is evolving rapidly. For video generators to be useful for applications ranging from robotics to film-making, they must consistently produce realistic videos. However, evaluating the realism of generated videos remains a largely manual process – requiring human annotation or bespoke evaluation "datasets" which have restricted scope. Here we develop an automated evaluation framework for video realism which captures both semantics and coherent 3D structure and which does not require access to a reference video. Our method, 3DSPA is a 3D spatiotemporal point autoencoder which integrates 3D point trajectories, depth cues, and DINO semantic features into a unified representation for video evaluation. 3DSPA models how objects move and what is happening in the scene, enabling robust assessments of realism, temporal consistency, and physical plausibility. Experiments show that 3DSPA reliably identifies videos which violate physical laws, is more sensitive to motion artifacts, and aligns more closely with human judgments of video quality and realism across multiple datasets. Our results demonstrate that enriching trajectory-based representations with 3D semantics offers a stronger foundation for benchmarking generative video models, and implicitly captures physical rule violations.

1 INTRODUCTION

Recent years have witnessed rapid progress in generative video models, with systems such as Sora (Brooks et al. (2024)), Kling AI (Kuaishou Technology (2024)), and Luma-Ray (LumaAI (2025)) capable of producing high-resolution, long-duration videos conditioned on natural language descriptions. These systems have started to showcase unprecedented visual fidelity, with coherent multi-objects, smooth camera motion, and diverse scenes. However, the end objective of developing these text-to-video models has always been to generate videos which are not only visually compelling but also realistic—capturing semantic meaning, temporal consistency, and physical plausibility in a way that mirrors a real-world video. If achieved, it will generate tremendous excitement across domains ranging from robotics and embodied AI (Wu et al. (2023); Yang et al. (2025); Fu et al. (2025)) to virtual reality (Christian et al. (2025)), education (Xu et al. (2025)), and creative industries like advertising and film-making.

Understanding the realism of generated videos is more than an aesthetic problem, it directly affects their utility for various downstream applications. In robotics and embodied AI, for example, policies trained in simulated environments that fail to accurately capture real-world dynamics may not transfer successfully to deployment settings. Similarly, in entertainment, audiences are sensitive to subtle cues of unrealistic motion, which can undermine immersion. Thus, a systematic way of measuring whether generated videos are physically plausible and perceptually realistic is a foundational requirement for both scientific and practical use.

However, existing approaches to measuring realism remain limited. The most common strategy is to rely on human annotation, where raters provide subjective assessments of qualities such as naturalness, temporal

smoothness (Wu et al. (2021); Skinner et al. (2023)). While such annotations are informative, they are expensive, time-consuming, and do not scale to the vast number of videos modern generative systems can produce. A second line of work has attempted to build discriminative benchmarks by constructing datasets of paired real and fake videos (Borji (2022)), training classifiers to distinguish them. Yet this requires careful curation of datasets, often domain-specific, and assumes that generated samples are comparable to available real-world footage. Neither approach provides a scalable, general-purpose solution.

Moreover, prior automated measures have largely equated realism with temporal consistency—ensuring that videos do not exhibit frame-to-frame flickering or incoherence. While temporal smoothness is indeed important, it is not sufficient. Realism also requires adherence to the semantics of motion and the physical laws that govern objects in three dimensions. For example, a video where a ball bounces upward indefinitely without slowing down might look temporally smooth but is physically implausible. Likewise, a car turning a corner but sliding sideways without frictional constraints violates semantic expectations of how vehicles move. Prior work has struggled to capture such failures because they demand reasoning about both semantics and 3D structure, not just pixels over time. Most existing evaluations (Allen et al. (2025), operate in 2D feature spaces, neglecting the fact that real-world objects persist in three dimensions, maintain continuity across occlusion, and obey physical laws such as gravity, inertia, and collision.

To address these challenges, we propose **3DSPA** (3D Semantic Point Autoencoder), a novel framework for assessing the realism of generated videos. 3DSPA combines semantic features with 3D point track autoencoding. The key idea is to represent a video as a sequence of tracked 3D points, enriched with semantic embeddings, and train an autoencoder that reconstructs these tracks. By compressing and reconstructing motion trajectories, the model is forced to capture underlying physical and semantic regularities, making deviations from realism detectable.

The main contributions of our paper include -

- First, we demonstrate that 3DSPA functions as a capable 3D point tracker, despite the information bottleneck inherent in autoencoding.
- Second, we show that it reliably detects violations of physical laws in controlled synthetic settings by using the IntPhys2 benchmark (Bordes et al. (2025)).
- Finally, we evaluate 3DSPA on existing works of using human annotations to judge realism using two datasets of generated videos, EvalCrafter (Skinner et al. (2023)) and VideoPhy-2 (Bansal et al. (2025)), and find that it better aligns with human judgments of motion quality and physical realism compared to existing baselines.

Together, these results suggest that incorporating semantics and 3D structure is essential for scalable, automated evaluation of generative video realism.

2 Related Work

Physical Benchmarking for Video Models Benchmarking intuitive and broader physical understanding has been central to recent progress in machine perception. Earlier works include IntPhys (Riochet et al. (2018)) evaluating models on core physical expectations through possible vs. impossible event videos. PHYRE (Bakhtin et al. (2019)) frames intuitive physics as counterfactual puzzle-solving, requiring agents to reason about object interactions. For broader reasoning, CLEVRER (Yi et al. (2020)) advances causal reasoning in videos via descriptive, explanatory, and counterfactual questions. More recently, Physion++ (Bear et al. (2023)) extends visual physics prediction tasks to richer scenarios involving rigid, soft, and fluid dynamics, providing a comprehensive evaluation suite. IntPhys2 (Bordes et al. (2025)) was recently released which is based on the violation of expectation framework challenge models to differentiate between possible and impossible events within controlled and diverse virtual environments.

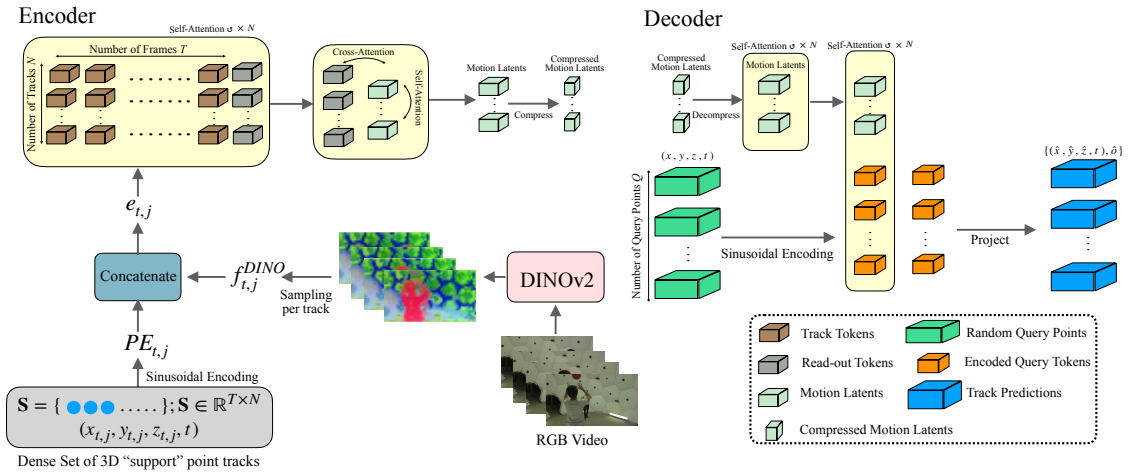

Figure 1: **3DSPA Architecture Overview :** The encoder integrates 3D trajectories, temporal embeddings, and DINOv2 ((Oquab et al., 2023)) semantic features into a compact latent representation using occlusion-aware attention and Perceiver-style transformer architecture (Jaegle et al. (2021)). The decoder conditions on query points to reconstruct full 3D trajectories with occlusion flags.

Video Quality Assessment Text-to-Video generative models are rapidly progressing, but it is still unclear how far they are from being able to generate videos which are indistinguishable from reality. Even evaluating this progress remains tricky. Earlier methods include FVD (Unterthiner et al. (2018)) and CLIP (Radford et al. (2021)) to evaluate the quality of frames and the text-frame alignment respectively. However, these metrics cannot capture realism more broadly, which simultaneously incorporates semantics and geometric structure. Recent works aim to create benchmarks with automated evaluators which tackle realism more directly, e.g. (Chen et al. (2025b)), and VBench (Huang et al. (2024)). However, many models have started saturating these benchmarks, achieving high scores of 90%+, since they are simply not challenging enough. Benchmarks such as EvalCrafter (Skinner et al. (2023)) and VideoPhy2 (Bansal et al., 2025) instead resort to human raters to perform comprehensive evaluation, and therefore avoid issues of benchmark saturation. Automated evaluators in these settings include optical flow and vision-language models, but no approach fully captures human assessments of motion quality and realism.

3 Model

Our goal is to provide an automated metric which can capture human ratings of realism for any video. To that end, we introduce 3DSPA : 3D Semantic Point Autoencoder. 3DSPA can be viewed as an extension of TRAJAN (Allen et al., 2025), a 2D point trajectory autoencoder that is trained to map a support set of point trajectories into a fixed-size motion latent representation which is further used to reconstruct query point tracks. While TRAJAN does a good job in capturing motion information in latents and reconstruction, it has some limitations. First, the model has no knowledge of the surrounding environment, only the point trajectories themselves. This means it cannot reason about scene context, object interactions, or occlusions that can be crucial for judging motion plausibility. Second, restricting motion to 2D trajectories is not sufficient for evaluating realistic dynamics, which naturally occur in 3D. As a result, TRAJAN cannot fully

capture the complexity of real-world motion. 3DSPA is instead designed to reconstruct 3D point tracks from random queries across space and time and provide semantic-aware motion latent representation.

3.1 ARCHITECTURE

3DSPA adopts an encoder–decoder setup, where the encoder E operates on a dense set of support point tracks $\mathbf{S} = \{s_{t,j}\}$, with each track defined as $s_{t,j} = (x_{t,j}, y_{t,j}, z_{t,j}, o_{t,j})$. Here, (x, y, z) denote the 3D position and o is a binary occlusion flag at time t for the j-th track. The model is trained to reconstruct a separate set of query trajectories $\mathbf{Q} = q_{t,j}$, which are randomly sampled from the video.

For each trajectory j, we embed its 3D positions $(x_{t,j}, y_{t,j}, z_{t,j})$ together with time t using sinusoidal encoding (denoted by $\mathrm{PE}_{t,j}$). In parallel, we sample DINOv2 (Oquab et al. (2023)) embeddings $f_{t,j}^{\mathrm{DINO}}$ from the corresponding video frame regions. These two representations are concatenated as $e_{t,j} = [\mathrm{PE}_{t,j} \,\|\, f_{t,j}^{\mathrm{DINO}}]$, and then projected into C channels.

A learnable "readout" token is initialized randomly and appended, and self-attention is applied across all the tokens with an occlusion-aware mask $(1 - o_{t,j})$ to ignore hidden points. After attention, only the readout token is retained, producing a compact C-dimensional descriptor for the track. To integrate information across tracks, we adopt a Perceiver-style transformer (Jaegle et al. (2021)) where a set of 128 latent tokens cross-attend to all track descriptors and then interact through self-attention. Finally, the latent tokens are compressed to yield a fixed 128×64 representation ϕ_S, capturing both motion dynamics and semantic appearance cues.

The decoder in 3DSPA follows the same design as TRAJAN, but now operates on a motion latent ϕ_S that already integrates 3D dynamics and semantic context. We train the decoder to reconstruct held-out query tracks. Concretely, given ϕ_S and a query point (x_q, y_q, z_q, t_q), the decoder predicts the full trajectory passing through that point. We first up-project the latent tokens in ϕ_S with an operator U and add a query *readout* token obtained from a sinusoidal encoding of (x_q, y_q, z_q, t_q). Self-attention is applied over all tokens, after which only the readout token is retained. A final linear projection maps this token to the predicted trajectory $(\hat{x}_t^q, \hat{y}_t^q, \hat{z}_t^q, \hat{o}_t^q)$ across all frames.

3.1.1 TRAINING

Following CoTracker3 (Karaev et al. (2024)), we train 3DSPA on a combination of synthetic and real datasets to ensure both controlled supervision and real-world generalization. We use the Kubric3D dataset generator (Greff et al. (2022)) to create 38k synthetic scenes with ground-truth 3D trajectories. While Kubric3D directly provides $(x_{t,j}, y_{t,j}, z_{t,j})$ for every point j at time t, it does not include explicit occlusion labels. To obtain occlusion flags, we project each 3D point into the image plane and compare its depth $z_{t,j}$ against the rendered depth map $D_t(x_{t,j}, y_{t,j})$ at that pixel. For depth maps we use VideoDepthAnything (VDA) metric model (Chen et al. (2025a)). The occlusion flag is then defined as $o_{t,j} = \mathbf{1}[\, z_{t,j} > D_t(x_{t,j}, y_{t,j}) + \epsilon \,]$, where ϵ is a small tolerance of 1e-4 to account for numerical precision. Thus, $o_{t,j} = 1$ indicates that the point is occluded at time t. In addition, we use the TAPVid-3D dataset Koppula et al. (2024), a large benchmark covering diverse real-world scenarios. TAPVid-3D contains 4,569 videos in the main split and 150 videos in the minival split, with lengths ranging from 25 to 300 frames. Unlike Kubric3D, TAPVid-3D provides full ground-truth annotations, including 3D point trajectories as well as occlusion flags $o_{t,j}$, making it directly suitable for evaluating models under realistic settings. In our setup, we train on the main set and use the minival split for evaluation. This benchmark also provides the necessary metrics for evaluation which we discuss later.

During training, we randomly divide the trajectories in each video into two equal halves. The first half is used as support tracks, which are passed through the encoder to produce the motion latents. The second half is used as query tracks, which the decoder must reconstruct given only the motion latents.

We trained 3DSPA with the AdamW Loshchilov & Hutter (2017) optimizer using a learning rate of 1e-4 warm-up followed by cosine decay. We initialize 3DSPA with a pretrained TRAJAN checkpoint and train for 300 epochs. Depth predictions are regularized with a scale-invariant penalty to handle differences in scale between synthetic and real domains. The DINO module (Oquab et al. (2023)) is frozen during training. Additional training details, including hyperparameter settings, batch sizes, and ablation studies, are provided in Appendix A.

The training loss is :

$$\mathcal{L}_{\text{3DSPA}} = \sum_{q,t} \Big(w_{l1} \|(x_{q,t}, y_{q,t}, z_{q,t}) - (\hat{x}_{q,t}, \hat{y}_{q,t}, \hat{z}_{q,t})\|_1 + w_{BCE}\text{BCE}(o_{q,t}, \hat{o}_{q,t}) \Big).$$

3.1.2 INFERENCE

At inference, we operate directly on 2D input videos but we require 3D point tracks. Dense 2D point tracks $(x_{t,j}, y_{t,j})$ with occlusion are first estimated using CoTracker3 Karaev et al. (2024), and subsequently lifted into 3D with metric depth predictions from VideoDepthAnything (VDA) metric model (Chen et al. (2025a)). The resulting 3D tracks $(x_{t,j}, y_{t,j}, z_{t,j})$ are then provided to the trained model in a 1:1 ratio as both support and query tracks. The reconstructed tracks produced by the decoder are finally used for evaluation.

Specifically, we calculate the Average Jaccard (AJ_{3D}) of the reconstructed tracks as a quantitative metric to see reconstruction error. Following TAPVid-3D (Koppula et al. (2024)), as AJ increases, the quality of reconstruction increase and vice-versa. The AJ metric calculates the number of true positives (number of points within the δ_{3D} threshold, predicted correctly to be visible), divided by the sum of true positives and false positives (predicted visible, but are occluded or farther than the threshold) and false negatives (visible points, predicted occluded or predicted to exceed the threshold).

4 RESULTS

To demonstrate that 3DSPA can capture realistic, physical motion, we evaluate three complementary axes: its accuracy in 3D point tracking as described in Section 4.1, its ability to detect physical law violations in possible vs. impossible video pairs as described in Section 4.2, and its alignment with human judgments of realism in generated videos as described in Section 4.3.

4.1 CAN 3DSPA RECONSTRUCT 3D POINT TRACKS?

We evaluate 3DSPA on the TAPVid-3D minival set and report three 3D point tracking metrics: Occlusion Accuracy (OA), which measures the precision of occlusion predictions; APD_{3D}, the average percentage of errors within multiple threshold scales δ; and Average Jaccard (AJ_{3D}), which quantifies the accuracy of both position and occlusion estimation. All these metrics are taken from Koppula et al. (2024)'s work.

Since 3DSPA is an *autoencoder* of point tracks, and therefore inherently less accurate due to its information bottleneck, we do not expect its performance in 3D point tracking to rival state-of-the-art approaches. Nevertheless, it is important that 3DSPA can reasonably accurately reconstruct 3D point tracks. We therefore compare 3DSPA against 3D-lifted versions of state-of-the-art 2D tracking methods and 3D tracking methods like SpatialTracker (Xiao et al. (2024) and SpatialTrackerv2 (Xiao et al. (2025)). Since most of these models were originally trained on the synthetic Kubric3D (Greff et al. (2022)), while our training data combines both Kubric3D (synthetic) and TAPVid-3D (real), we additionally fine-tune CoTracker3 model (Karaev et al. (2024)) on TAPVid-3D pseudo labels and evaluate all models on the minival set. Table 1 summarizes the comparative 3D tracking performance. 3DSPA consistently outperforms most baselines and

Method	Aria			DriveTrack			PStudio			Average		
	AJ_{3D} ↑	APD_{3D} ↑	OA ↑	AJ_{3D} ↑	APD_{3D} ↑	OA ↑	AJ_{3D} ↑	APD_{3D} ↑	OA ↑	AJ_{3D} ↑	APD_{3D} ↑	OA ↑
TAPIR + ZoeDepth	15.7	23.5	79.8	6.3	10.5	81.6	11.2	18.9	78.7	11.0	17.6	80.1
CoTracker + ZoeDepth	17.0	25.7	88.0	6.0	10.9	82.6	11.4	19.9	80.0	11.4	18.8	83.5
BootsTAPIR + ZoeDepth	11.8	16.3	86.7	6.4	10.9	85.3	11.6	19.6	82.6	11.6	18.9	84.9
CoTracker3 + ZoeDepth	15.6	24.1	88.6	13.3	19.6	86.8	9.0	13.6	83.9	12.6	19.1	86.4
SpatialTracker	16.7	25.7	89.3	6.9	12.4	83.7	12.3	21.6	78.5	12.0	19.9	83.8
SpatialTrackerV2	18.6	26.3	90.8	16.4	24.3	90.2	18.1	27.6	86.7	17.7	26.0	89.2
CoTracker3-finetuned + ZoeDepth	16.8	25.5	89.6	13.6	19.9	88.7	10.1	14.3	87.1	13.5	19.9	88.5
Ours	17.7	24.9	89.2	11.9	14.8	85.7	12.3	19.9	82.5	14.0	19.8	85.8

Table 1: 3D point tracking results on the TapVid-3D *minival* set. 3DSPA achieves competitive accuracy across datasets and performs on par with a finetuned CoTracker3, highlighting its ability to reconstruct consistent and accurate 3D tracks.

Figure 2: Example 3D point tracks reconstructed by 3DSPA for a generated video of a man mopping a floor in the VideoPhy-2 dataset (Bansal et al., 2025). This video has a good human rating and was reconstructed pretty well by our model.

achieves performance on par with CoTracker3 (Karaev et al. (2024)) when fine-tuned on the TAPVid-3D main dataset.

We additionally provide an example of how 3DSPA performs in 3D track reconstruction when only a 2D input video is provided in Figure 2. Despite the noisy depth signal obtained from VideoDepthAnything, 3DSPA reconstructs smooth 3D tracks for the generated video.

Both results demonstrate that 3DSPA is capable of reconstructing 3D point tracks accurately despite its compressed latent space bottleneck, and motivates its candidacy as an automated metric for video realism.

4.2 Can 3DSPA detect physical rule violations?

For an automated metric of video realism to be useful, we need to be sure that it will detect physical rule violations. To assess whether 3DSPA can reliably distinguish physically real and unreal scenarios, we evaluate on the IntPhys2 (Bordes et al. (2025)) dataset. IntPhys2 contains 1,012 videos across 253 scenes, organized as quadruplets of two **possible** (real) and two **impossible** (unreal) outcomes. Each video tests one of four core physical principles: **object permanence**, where objects continue to exist even when occluded; **object immutability**, where objects maintain their shape and structure; **spatio-temporal continuity**, where objects move smoothly through time and space; and **solidity**, where objects occupy space and cannot pass through one another. All videos are rendered in the Unreal Engine with both static and moving cameras, increasing realism and memory demands.

Baselines and ablations We compare 3DSPA against several state-of-the-art vision-language models, self-supervised vision foundation models, and TRAJAN variants that progressively add dimensional and semantic information. The original TRAJAN model uses only 2D point tracks without depth cues or semantic

Model / Category	Permanence		Immutability		Continuity		Solidity	
	Fixed	Moving	Fixed	Moving	Fixed	Moving	Fixed	Moving
GPT-4o (Hurst et al. (2024))	59.62	58.82	58.65	59.56	54.81	57.35	56.73	55.32
Qwen-VL 2.5 (Bai et al. (2025))	53.85	54.41	56.73	53.68	52.88	54.41	50.96	51.06
Gemini-1.5 Pro (Google DeepMind (2024))	55.77	55.88	56.73	56.73	54.80	54.80	56.73	56.73
Gemini-2.5 Flash (Google DeepMind (2025))	64.42	58.82	59.62	63.97	54.81	55.15	55.77	56.38
VideoMAEv2-g (Wang et al. (2023))	63.46	50.00	54.81	53.69	65.38	54.41	48.08	59.57
Cosmos-4B (Agarwal et al. (2025))	51.92	41.18	50.96	48.32	53.85	50.00	48.08	55.32
V-JEPA 2-h (Assran et al. (2025))	63.46	67.65	51.92	56.38	50.00	57.35	50.00	52.13
V-JEPA-h+RoPE (Bardes et al. (2024))	59.62	57.35	55.77	58.72	57.69	**75.00**	46.15	58.51
TRAJAN	44.23	50.00	54.81	58.82	53.85	48.53	46.15	52.13
TRAJAN+DINO	61.54	**76.47**	**75.00**	73.08	**78.85**	73.53	69.23	59.57
TRAJAN+3D	65.38	60.29	46.15	66.18	50.00	58.82	38.46	39.36
3DSPA	**76.92**	75.00	73.08	**76.47**	67.31	69.12	**70.77**	**64.47**
Human	100.0	99.26	97.11	90.44	99.04	94.44	96.15	95.21

Table 2: Win rates (%) on IntPhys2 across physical principles. Top row reports prior models' win rates. Bottom rows benchmark against 3DSPA , ablations, and human performance. 3DSPA and TRAJAN+DINO strongly outperform all alternatives in detecting physically implausible events across most concept categories.

features. **TRAJAN+3D** is a 3D extension where we add an extra spatial dimension in the autoencoder to better capture motion dynamics. **TRAJAN+DINO** instead augments the representation with semantic features from DINOv2, while still excluding 3D information and depth cues. Together, these variants highlight the individual roles of 3D structure and semantic context in detecting physical rule violations.

Performance Analysis. Table 2 shows the performance of 3DSPA against the baselines reported in Bordes et al. (2025) as well as our ablations. 3DSPA and TRAJAN+DINO significantly outperform all alternatives across most concept categories. Perhaps most surprisingly, 3DSPA shows the *most benefit* over alternatives in the permanence ($+10\%$), immutability ($+10\%$), and solidity ($+5\%$) concept categories rather than continuity (approximately -5 to $+2\%$). This suggests that a small amount of 3D point track data is sufficient for models to learn what is physically plausible or not, and that reconstructing semantic 3D tracks is a better signal for learning *realistic, plausible* physical motion than next frame prediction or next token prediction.

Looking at the ablations, most of the benefits of 3DSPA in determining possible vs. impossible physics may be due to the inclusion of DINO features. Although 3DSPA performs best overall, TRAJAN+DINO performs comparably to 3DSPA in most concept categories, indicating that *semantic information* is key for understanding physical principles. By comparison, TRAJAN and TRAJAN+3D perform comparably to previously evaluated predictive and Multimodal LLM (MLLM) approaches. We provide additional results in Appendix B.

4.3 Does 3DSPA capture human evaluations of realism in generated videos?

A key challenge in evaluating generated videos is measuring realism without relying on reference videos. This is particularly relevant when training data are inaccessible or when sampling a large number of outputs is computationally prohibitive. Human evaluation has therefore become the gold standard, since people can naturally judge whether motion appears realistic and physically plausible.

To ground our study, we draw on two datasets which include a large set of videos generated by a large collection of generative video models: VideoPhy-2 (Bansal et al. (2025)) and EvalCrafter (Skinner et al., 2023).

Model	Spearman (PC)
Video Evaluation Models (fine-tuned)	
VideoCon–Physics	0.48
VideoCon	0.13
VideoLlava	0.08
VideoScore	0.17
VIDEOPHY–2–AUTOEVAL	**0.76**
TRAJAN Variants (no fine-tuning)	
TRAJAN	0.19
TRAJAN+DINO	0.40
TRAJAN+3D	0.50
3DSPA (ours)	**0.74**

Table 3: Spearman rank coefficients on the VideoPhy-2 benchmark for physical commonsense (PC). Video Evaluation Models are fine-tuned vision-language models.

Figure 3: An example *unrealistic* video (Physical Commonsense score of 2/5) from VideoPhy-2 which 3DSPA scores poorly (Average 3D Jaccard of 6.95) but TRAJAN scores highly (Average 2D Jaccard of 60.9).

VideoPhy-2 emphasizes action-centric videos and includes human annotations of physical commonsense and semantic adherence to the text prompt. EvalCrafter (Skinner et al. (2023)), evaluates video quality with a larger set of five metrics including motion quality, temporal consistency, and several prompt adherence measures.

VideoPhy-2 We use the VideoPhy-2 benchmark (Bansal et al. (2025)) to assess how well 3DSPA performs as an automated realism metric relative to human judgments. This benchmark emphasizes two key aspects: *semantic adherence* (SA), which measures whether generated videos follow the intended action semantics, and *physical commonsense* (PC), which evaluates whether the motion and interactions in videos are consistent with intuitive physical rules. We are primarily interested in physical commonsense. Bansal et al. (2025) also provide an automated evaluation metric, VIDEOPHY-2 AutoEval, which is a vision-language model fine-tuned to predict a physical commonsense score on a subset of the generated video dataset.

We measure the automated metric quality by correlating model ratings and human ratings with the Spearman rank coefficient. Model ratings are calculated automatically as the Average Jaccard for each video – a proxy for the reconstruction error of the autoencoder.

As shown in Table 3, 3DSPA substantially outperforms 2D variants such as TRAJAN and TRAJAN+DINO in tracking human ratings of physical commonsense, and also provides a significant boost over the 3D baseline. The inclusion of both 3D structural cues and semantic DINO features enables stronger alignment with human assessments, where 3DSPA achieves the highest Spearman rank coefficient among TRAJAN variants. More remarkably, 3DSPA strongly outperforms most vision-language models (VideoCon, VideoScore, and VideoLlava) on this task, and even closely matches VIDEOPHY-2 AutoEval despite not being trained on the provided dataset.

Figure 3 shows an example video which highlights the difference between 3DSPA and TRAJAN in capturing physical commonsense. In this video of a man smashing a concrete wall with a hammer, the motions are smooth but the hammer also partially disappears. TRAJAN assigns a high realism rating because it is only sensitive to the smooth motion. 3DSPA assigns the video a low score because it is additionally sensitive to

the semantics: the hammer cannot just disappear. This underscores the importance of semantic information when evaluating physical realism.

EvalCrafter Similarly to VideoPhy-2, EvalCrafter (Skinner et al. (2023)) is a dataset consisting of generated videos from several frontier generative video models. For each video, a set of human annotators rated the **visual quality**, **text to video consistency**, **motion quality**, **temporal consistency**, and **subjective likeness**. Similarly to VideoPhy-2, we compute Spearman rank coefficients between human ratings and the Average Jaccard (AJ) for each of the TRAJAN variants and 3DSPA . Since many videos in EvalCrafter contain no motion (and therefore could not be assessed as physically realistic or not), we restricted evaluation to videos with medium to high motion, defined as the top 50% of videos ranked by change in 3D point track positions, yielding a test set of 1,849 videos. Table 4 clearly demonstrates that 3DSPA achieves the best performance; further highlighting that integrating both 3D structure and semantic DINO features provides the strongest predictor of a variety of human annotations for generated videos.

Model	Visual Quality	T2V	Motion Quality	Consistency	Subjective Likeness
TRAJAN	0.28	0.25	0.24	0.33	0.23
TRAJAN+DINO	0.31	0.44	0.41	0.39	0.46
TRAJAN+3D	0.45	0.55	0.49	0.48	**0.63**
3DSPA (ours)	**0.48**	**0.58**	**0.55**	**0.60**	0.60

Table 4: Spearman rank coefficients between human annotations and automated AJ scores across different TRAJAN variants for the categories of visual quality, text-to-video similarity, motion quality, temporal consistency, and subjective likeness.

5 Discussion and Conclusion

We introduce the 3D Semantic Point Autoencoder (3DSPA), a framework for evaluating video realism using semantic-aware 3D point trajectories. Across several experiments, we found that 3DSPA 's combination of semantic and 3D geometric information was crucial for (1) 3D point track reconstruction, (2) physical rule violation detection, and (3) matching various human annotations of generated videos, including motion quality and adherence to physical commonsense.

Through extensive ablation studies, we determined that semantic information is particularly crucial to determining whether a video is physically realistic – geometry alone is not enough. Perhaps most surprisingly, 3DSPA outperforms state-of-the-art vision-language models both in detecting synthetic physical rule violations such as solidity and immutability, and in tracking human physical commonsense judgements of generated videos.

Overall, 3DSPA offers a scalable alternative to human evaluation of video realism. We believe 3D point tracks naturally capture depth-aware motion, interactions, and occlusion, making them more effective than frame-based metrics for spotting subtle physics violations. In future work, we plan to make trajectories depend on past motion, enabling stronger tests of long-term dynamics and temporal realism, as well as investigate whether these metrics can be used to improve or regularize the training of generative video models.

References

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Kelsey Allen, Carl Doersch, Guangyao Zhou, Mohammed Suhail, Danny Driess, Ignacio Rocco, Yulia Rubanova, Thomas Kipf, Mehdi SM Sajjadi, Kevin Murphy, et al. Direct motion models for assessing generated videos. *arXiv preprint arXiv:2505.00209*, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Anton Bakhtin, Aäron van den Oord, Arthur Chen, Yilun Du, Yuandong Tian, Tim Rocktäschel, and Edward Grefenstette. Phyre: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Daniel Bear, Hsiao-Yu Fish Tung, Yilun Du, Tanmay Gupta, Peter Battaglia, Joshua B. Tenenbaum, and Jiajun Wu. Physion++: A benchmark and evaluation platform for physical scene understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 2022.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22831–22840, 2025a.

Yongfan Chen, Xiuwen Zhu, and Tianyu Li. A physical coherence benchmark for evaluating video generation models via optical flow-guided frame prediction. *arXiv preprint arXiv:2502.05503*, 2025b.

Nicolay Anderson Christian, Jason Turuwhenua, and Mohammad Norouzifard. Ai and generative models in 360-degree video creation: Building the future of virtual realities. *Applied Sciences*, 15(17):9292, 2025.

Xiao Fu, Xintao Wang, Xian Liu, Jianhong Bai, Runsen Xu, Pengfei Wan, Di Zhang, and Dahua Lin. Learning video generation for robotic manipulation with collaborative trajectory control. *arXiv preprint arXiv:2506.01943*, 2025.

Google DeepMind. Gemini 1.5: Unlocking multimodal understanding across modalities. `https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/`, 2024. Accessed: 2025-09-25.

Google DeepMind. Gemini 2.5 flash: Lightweight multimodal reasoning at scale. https://deepmind.google/technologies/gemini/, 2025. Accessed: 2025-09-25.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024.

Skanda Koppula, Ignacio Rocco, Yi Yang, Joe Heyward, Joao Carreira, Andrew Zisserman, Gabriel Brostow, and Carl Doersch. Tapvid-3d: A benchmark for tracking any point in 3d. *Advances in Neural Information Processing Systems*, 37:82149–82165, 2024.

Kuaishou Technology. Kling: Ai video generation model. https://klingai.com/, 2024. Versions 1.0, 1.5, 1.6, 2.0, 2.1 released 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

LumaAI. Luma ray2. https://lumalabs.ai/ray, July 2025. Accessed: 2025-07-08.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rémi Riochet, Nicolas Garcia, Fabien Baradel, Remi Cadene, Michael Arbel, Claire Donnat, Olivier Blayo, Antoine Bordes, R Devon Hjelm, Aaron Courville, et al. Intphys: A framework and benchmark for visual intuitive physics reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

James Skinner et al. Evalcrafter: Benchmarking video generation models with comprehensive evaluation. In *NeurIPS Datasets and Benchmarks*, 2023.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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

Chen Sun Wu et al. Tcc: Time-contrastive networks for self-supervised video representation learning. In *CVPR*, 2021.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023.

Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20406–20417, 2024.

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: 3d point tracking made easy. *arXiv preprint arXiv:2507.12462*, 2025.

Tao Xu, Yuan Liu, Yaru Jin, Yueyao Qu, Jie Bai, Wenlan Zhang, and Yun Zhou. From recorded to ai-generated instructional videos: A comparison of learning performance and experience. *British Journal of Educational Technology*, 56(4):1463–1487, 2025.

Xiuyu Yang, Bohan Li, Shaocong Xu, Nan Wang, Chongjie Ye, Zhaoxi Chen, Minghan Qin, Yikang Ding, Xin Jin, Hang Zhao, et al. Orv: 4d occupancy-centric robot video generation. *arXiv preprint arXiv:2506.03079*, 2025.

Kexin Yi, Chuang Gan, Yunzhu Li, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

A Training Setup & Hyperparameters

We train our model using AdamW (Loshchilov & Hutter, 2017) with a cosine learning rate schedule, preceded by a warmup of 10000 steps. The peak learning rate is set to 1×10^{-4}. Training is performed for 300 epochs with a batch size of 256. This extended training schedule, along with the larger batch size, allows the model to better stabilize its motion representation and improve generalization across diverse video scenarios.

As before, we supervise both position and occlusion prediction, applying a L1 loss on (x_t, y_t) coordinates and a cross-entropy loss on the occlusion logit o_t. We maintain the weighting ratio of $5000 : 10^{-8}$, which prioritizes motion fidelity while encouraging invariance to occlusion. We found that balancing these losses equally degraded correlation with human judgments of realism, consistent with our earlier observations.

To improve temporal localization of query points, we replace the naive linear up-projection operator with a strided-window upsampling operator. Specifically, each latent token ϕ_S^l is linearly up-projected and concatenated with a temporal window $[\rho_t : \rho_t + 128)$ along the channel axis. This encourages the decoder to attend to temporally relevant information for a given query point.

A.1 HYPERPARAMETERS

Tables 5 and 6 provide the full set of hyperparameters for positional encoding, projection operators, and transformer modules. Compared to the original TRAJAN configuration, we increase the dimensionalities of the projection layers as we have increased the data and added additional parameters due to DINO and depth features, enabling richer multi-modal fusion of semantic and geometric cues.

Component	**Hyperparameter Value**
Sinusoidal embedding (spatial + temporal + depth)	32 frequencies
Track token projection dimensionality (C)	384
DINO feature projection dimensionality	768
Depth feature projection dimensionality	256
Compression dimensionality	96
Up-projection dimensionality	1280
Query point encoder dimensionality	1280

Table 5: Positional encoding and projection operator hyperparameters for 3DSPA .

B ADDITIONAL RESULTS ON INTPHYS2

Dataset Structure. The videos are further categorized by difficulty:

- **Easy (104 videos):** Simple environments with colorful geometric shapes.
- **Medium (400 videos):** Diverse backgrounds with textured shapes.
- **Hard (336 videos):** Realistic objects within cluttered, complex backgrounds.
- **Unknown (172 videos):** Mixed or ambiguous scenes.

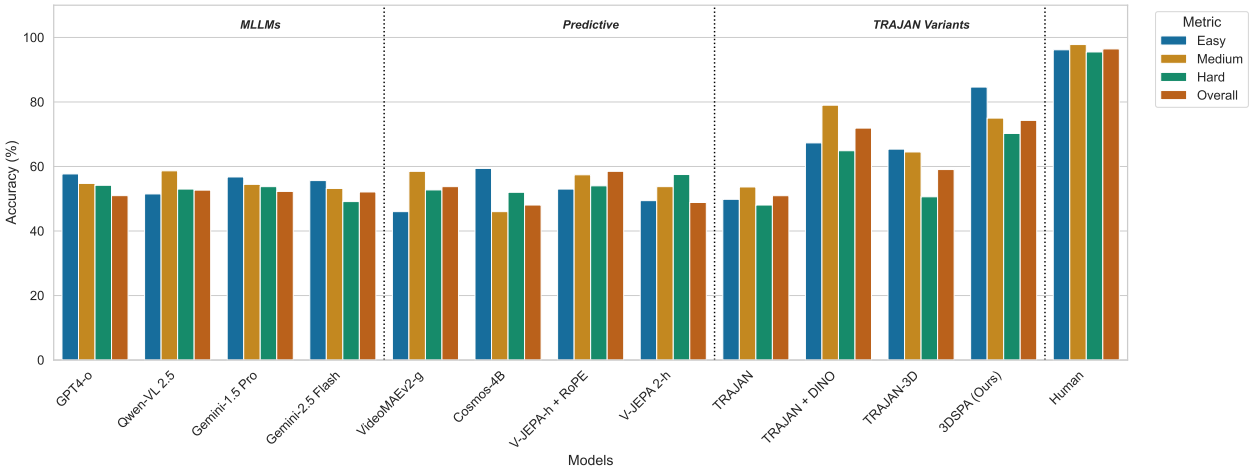

Figure 4: Performance comparison across models on the IntPhys2 benchmark for each of the *easy*, *medium* and *hard* categories.

Transformer Name	Attention Type	QKV Size	Layers	Heads	MLP Size
Input 3D track transformer	SA	96×8	3	8	1536
Perceiver-style transformer	CA	96×8	4	8	2048
Up-projection latent transformer (decoder)	CA	96×8	4	8	2048
Track readout transformer	CA	96×8	4	8	1536

Table 6: Transformer architecture hyperparameters for 3DSPA . SA = self-attention, CA = cross-attention.

We report results across the three difficulty variants to better capture how model performance scales with increasing visual and physical complexity. This breakdown provides motivation for our evaluation, as it disentangles robustness to simple synthetic settings from generalization to realistic and cluttered environments (see Figure 4). Notably, performance on the **hard** category is consistently the lowest in terms of Average Jaccard, highlighting the challenge of reconstructing tracks in realistic scenes with heavy clutter, occlusions, and object interactions.