3DSPA: A 3D Semantic Point Autoencoder for Evaluating Video Realism

Bhavik Chandna
University of California San Diego
Vector Institute

Kelsey R. Allen
University of British Columbia
Vector Institute

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 S**emantic **P**oint **A**utoencoder which is trained to reconstruct held out 3D point trajectories which have been augmented with DINO semantic features. By combining *both* semantic and geometric representations, 3DSPA enables robust assessments of realism, temporal consistency, and physical plausibility in generated videos. Experiments show that 3DSPA outperforms leading VLMs by more than 10% in identifying videos which violate physical laws, and more than doubles the correlations between human and auto-eval ratings in judging physical common sense and motion quality across multiple generative video datasets. Our results demonstrate that enriching trajectory-based representations with 3D semantics offers a strong foundation for benchmarking generative video models, and implicitly captures physical rule violations. Source code is available at `https://github.com/TheProParadox/3dspa_code`.

1 Introduction

Recent years have witnessed rapid progress in generative video models, with systems such as Veo 3.1 (Gelman, 2025), Runway Gen-4 (Runway, 2025), Kling AI (Kuaishou Technology, 2024), and Luma-Ray (LumaAI, 2025) capable of producing high-resolution, long-duration videos. These systems have started to showcase unprecedented visual fidelity, with coherent multi-object movement and smooth camera motion across diverse scenes. However, the ultimate 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 real-world video. If achieved, it would create tremendous excitement across domains ranging from robotics and embodied AI (Wu et al., 2023b; Yang et al., 2025; Fu et al., 2025) to virtual reality (Christian et al., 2025), education (Xu et al., 2025), and creative industries such as advertising and film-making. In robotics and embodied AI, policies trained on physically implausible video rollouts may fail to transfer to the real world, while in entertainment, unrealistic motion can break immersion. Scalable measures of plausibility and realism are therefore necessary for practical deployment.

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, and physical plausibility (Liu et al., 2024b; Bansal et al., 2025; Chinchure et al., 2025). While such annotations are informative, they are expensive and time-consuming, and

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

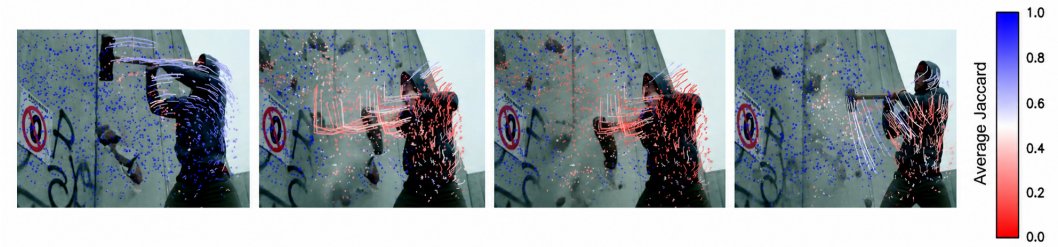


Figure 1: Example 3D point tracks (projected into 2D) reconstructed by 3DSPA for an unrealistic generated video from VideoPhy-2 (video frames progress from left to right in time) (Bansal et al., 2025), depicting a man striking a wall with a hammer and clearly violating physical laws. 3DSPA assigns low realism as measured with the Average Jaccard (AJ) (blue: high, red: low, white: intermediate). The video ranks among the lowest in both 3DSPA's scores and human realism ratings, highlighting strong alignment with human judgment.

do not scale to the vast number of videos modern generative systems can produce. A second line of work has attempted to build benchmarks by constructing datasets of paired real and fake videos (Borji, 2019; Chen et al., 2024; Motamed et al., 2026; Bear et al., 2021; Tung et al., 2023). 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, simply ensuring that videos do not exhibit frame-to-frame flickering or incoherence. While temporal smoothness is 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 reason about pixels rather semantics and 3D structure. Existing automated evaluation metrics make overly limiting simplifications, such as operating in 2D feature spaces (Allen et al., 2025; Liu et al., 2024b) and ignoring issues pertaining to 3D such as occlusion, or operating in 3D but assuming that the world remains static except for the camera (Chang et al., 2025; Asim et al., 2025; Chen et al., 2025b).

To address these challenges, we propose **3DSPA** (3D Semantic Point Autoencoder), a framework for evaluating generated-video realism through semantic-aware 3D point trajectories. 3DSPA represents each video as tracked 3D points augmented with semantic embeddings, and trains an autoencoder to reconstruct their motion over time. The reconstruction quality score is therefore not realism itself, but an operational proxy for compatibility with a learned semantic 3D motion prior. Importantly, 3DSPA shifts the evaluation signal from image-plane temporal consistency toward semantic 3D physical plausibility, helping capture smooth but implausible events such as object disappearance, depth-inconsistent motion, and violations of object permanence or solidity.

The main contributions of our paper include:

- First, we demonstrate that 3DSPA reconstructs held-out 3D point tracks with competitive position and visibility accuracy, despite the information bottleneck inherent in autoencoding.
- Second, we show that 3DSPA reliably detects violations of physical laws in controlled synthetic settings by using the IntPhys2 benchmark (Bordes et al., 2025), up to 20% better than leading alternatives.
- Finally, we evaluate 3DSPA on two existing generated video datasets which provide human annotations of realism: EvalCrafter (Liu et al., 2024b) and VideoPhy-2 (Bansal et al., 2025), and find that it aligns up to 2x better with human judgments of motion quality and physical realism compared to existing baselines.

Together, these results indicate that combining semantic features with 3D structure improves automated evaluation of generated 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. In simulation, earlier works include IntPhys (Riochet et al., 2018) which evaluates models on core physical expectations through possible vs. impossible event videos and CLEVRER (Yi et al., 2019) which aimed to advance causal reasoning in videos via descriptive, explanatory, and counterfactual questions. (Bear et al., 2021) and (Tung et al., 2023) similarly provide simulated videos of different physical phenomena, including rigid and non-rigid objects. Recent works include Physion-Eval (Zhang et al., 2026) which benchmarks physical realism in generated videos using expert annotations of temporally localized physical glitches and fine-grained failure categories, and IntPhys2 (Bordes et al., 2025) which introduces more complex backgrounds and objects to the original IntPhys. In the real world, InfLevel (Weihs et al., 2022) extends these ideas by collecting real-world videos of physics violations. Physics IQ (Motamed et al., 2026) and TRAVL (Motamed et al., 2025) similarly provide a set of real-world videos covering a variety of physical principles. In these benchmarks, models' physical commonsense is measured by either directly comparing their predictions about how a real world video will unfold to the video itself, or by investigating whether models are more *surprised* by an unreal video than by a real one.

Video quality assessment Video generative models are rapidly improving, but evaluating whether their outputs are truly realistic remains challenging. Early metrics such as FVD (Unterthiner et al., 2018) and CLIP-based scores (Radford et al., 2021) measure distributional quality and text-frame alignment, but do not directly capture realism, which also depends on semantic coherence, motion consistency, and physical plausibility. Later FVD-style metrics improve temporal sensitivity, but still rely on distribution-level comparisons with reference videos (Luo et al., 2024; Liu et al., 2024a). Recent automated evaluators and benchmarks target realism more directly, including physics-aware metrics (Chen et al., 2025c) and large-scale suites such as VBench (Huang et al., 2024); however, many models now achieve saturated scores, motivating human-judgment benchmarks such as EvalCrafter (Liu et al., 2024b) and VideoPhy-2 (Bansal et al., 2025). Other works detect synthetic videos using temporal artifacts or spatial-temporal forgery cues (Zheng et al., 2025; Shi et al., 2025), but focus on binary detection rather than realism alignment.

Recent work has also explored 3D structure for evaluating generated visual content. L3DE (Chang et al., 2025) uses monocular appearance, motion, and depth cues to assess AI-generated videos, MEt3R (Asim et al., 2025) measures multi-view consistency through dense 3D correspondences and feature warping, and Grab-3D (Chen et al., 2025b) detects generated videos using vanishing-point temporal consistency. While these methods highlight the value of geometry, they mainly evaluate visual/geometric consistency or binary synthetic-video detection, and mainly focus on videos where only the camera can move. Allen et al. (2025) use a 2D point-trajectory autoencoder for automatic realism evaluation, but primarily capture temporal consistency. In contrast, our method evaluates reference-free video realism through compatibility with a learned semantic 3D motion prior.

3 Model

3DSPA is a reference-free autoencoder that reconstructs held-out semantic 3D point tracks from a compact latent representation of support tracks. This section describes how support tracks are estimated from 2D videos, then encoded, and then finally how held out tracks are reconstructed, to produce the final evaluation score.

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.

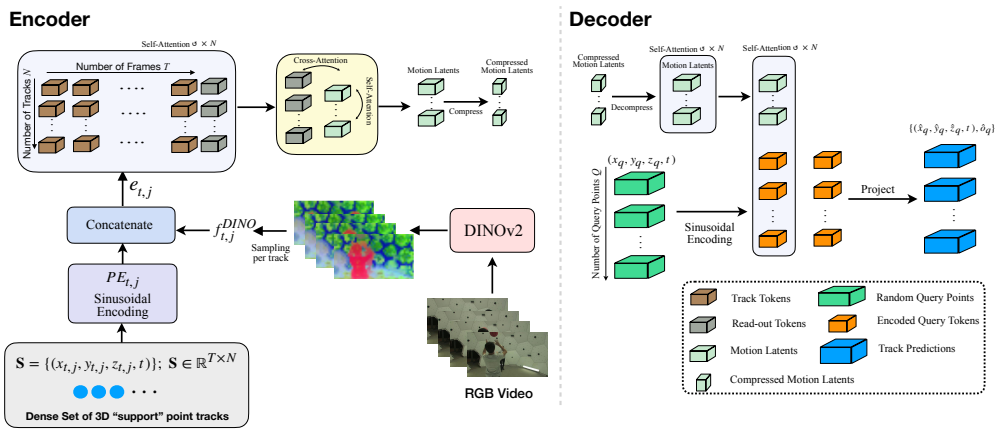


Figure 2: **3DSPA architecture overview.** The encoder integrates 3D trajectories, temporal embeddings, and DINOv2 (Oquab et al., 2023) semantic features into a compact latent representation using occlusion-aware attention and a Perceiver-style transformer architecture (Jaegle et al., 2021). The decoder conditions on query points to reconstruct full 3D trajectories with occlusion flags.

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 operates on a motion latent ϕ_S that 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., 2025), 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. Kubric3D provides ground-truth 3D point positions and a renderer-generated depth buffer. To obtain occlusion flags, we project each 3D point to image coordinates $(u_{t,j}, v_{t,j})$ using the camera parameters and compare its camera-frame depth $z_{t,j}^{\text{cam}}$ with the *ground-truth rendered* depth $D_t^{\text{Kubric}}(u_{t,j}, v_{t,j})$. The occlusion flag is defined as

$$o_{t,j} = \mathbf{1}\left[z_{t,j}^{\text{cam}} > D_t^{\text{Kubric}}(u_{t,j}, v_{t,j}) + \epsilon\right], \tag{1}$$

where $\epsilon = 10^{-4}$ is a numerical tolerance and $o_{t,j} = 1$ denotes an occluded point. This training-time supervision uses the Kubric renderer's ground-truth depth buffer, *not* predicted depth from VideoDepthAnything (VDA). VDA is used to estimate depth for ordinary RGB videos at inference time.

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 training and evaluating models under realistic settings. In our setup, we train on the main set and use the minival split for evaluation. We analyze the effect of training

data composition in Appendix D.1, where we show that adding TAPVid-3D to Kubric3D improves reconstruction performance over Kubric3D-only training which means real-video trajectories help reduce synthetic-domain bias while retaining the precise 3D supervision provided by Kubric3D.

During training, we randomly divide the point trajectories in each video into two equal parts. 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 use this 50:50 support–query split based on the partition sensitivity analysis in Appendix D.3, where it achieves the best performance among the tested splits.

We initialize 3DSPA with a pretrained 2D point-track autoencoder checkpoint (Allen et al., 2025) and train for 300 epochs on 8 H100 GPUs. To account for depth-scale differences between synthetic and real-world training data, we additionally apply a scale-invariant depth regularization term to the reconstructed 3D trajectories. 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:

$$\mathcal{L}_{\text{model}} = \sum_{q,t} (w_{l1} \|\mathbf{p}_{q,t} - \hat{\mathbf{p}}_{q,t}\|_1 + w_{\text{BCE}} \text{BCE}(o_{q,t}, \hat{o}_{q,t})) \,. \tag{2}$$

where $\mathbf{p}_{q,t} = (x_{q,t}, y_{q,t}, z_{q,t})$.

3.1.2 Inference

During inference, 3DSPA operates directly on 2D input videos while constructing 3D point tracks internally. Dense 2D point tracks $(x_{t,j}, y_{t,j})$ with occlusion are first estimated using CoTracker3 (Karaev et al., 2025), providing dense spatial coverage across the video, and then lifted into 3D using metric depth predictions from VideoDepthAnything (VDA) (Chen et al., 2025a). Similarly to training, half of the resulting 3D tracks $(x_{t,j}, y_{t,j}, z_{t,j})$ are used as support tracks, while the remaining tracks serve as queries for reconstruction. The reconstructed query tracks are evaluated against the corresponding held-out tracks produced by the same upstream perception pipeline, without reference to a ground-truth video. Based on the point-density analysis in Appendix D.3, we use 1024 sampled tracks in the main experiments, as performance saturates around this point.

Importantly, the upstream tracking, depth, and semantic models only provide input measurements: CoTracker3 estimates correspondences, VDA provides depth, and DINOv2 supplies semantic context. The realism score itself comes from the learned reconstruction prior: realistic videos produce trajectories that can be compactly encoded and reconstructed, while physically implausible videos yield inconsistent or difficult-to-reconstruct trajectories. Thus, 3DSPA is best understood as a learned semantic 3D reconstruction metric built on standard perception measurements, rather than a direct evaluation of any single upstream model. We further assess sensitivity to tracking and depth quality in Appendix D.2, showing only modest degradation when replacing ground-truth tracks and depth with predicted inputs. Meanwhile, our ablations in Section 4 show that semantics are important and geometry alone is insufficient, suggesting that 3DSPA's performance comes from integrating semantic and 3D cues through the learned reconstruction prior.

For reconstruction-based evaluation, we compute Average Jaccard (AJ) between reconstructed query tracks and the corresponding target query tracks. We do not interpret AJ as a direct measure of realism itself. Rather, AJ serves as a proxy for compatibility with the learned prior; videos whose trajectories are physically and semantically consistent should be easier to compactly encode and reconstruct, while implausible videos should induce geometric, semantic, or occlusion inconsistencies that lower reconstruction quality. Following TAPVid-3D (Koppula et al., 2024), higher AJ indicates better reconstruction. The metric calculates the number of true positives, defined as visible points whose spatial positions are within a δ_{3D} threshold and whose visibility is correctly predicted, divided by the sum of true positives, false positives, and false negatives. False positives correspond to points predicted as visible but either occluded or farther than the threshold, while false negatives correspond to visible points that are predicted as occluded or exceed the threshold.

Method	Aria			DriveTrack			PStudio			Average		
	$AJ\uparrow$	$APD\uparrow$	$OA\uparrow$	$AJ\uparrow$	$APD\uparrow$	$OA\uparrow$	$AJ\uparrow$	$APD\uparrow$	$OA\uparrow$	$AJ\uparrow$	$APD\uparrow$	$OA\uparrow$
TAPIR + ZD	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 + ZD	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 + ZD	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 + ZD	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-FT + ZD	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 3D tracking performance across datasets and performs on par with a finetuned CoTracker3 (CoTracker3-FT+ZD), highlighting its ability to reconstruct consistent and accurate 3D tracks despite the information bottleneck.

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 using three standard 3D tracking metrics from Koppula et al. (2024): Occlusion Accuracy (OA), which measures visibility prediction accuracy; APD, which measures the percentage of 3D position errors within multiple threshold scales δ; and Average Jaccard (AJ), which jointly evaluates 3D position and occlusion accuracy.

Since 3DSPA is a point-track *autoencoder*, its compressed latent bottleneck makes it inherently different from dedicated tracking models. We therefore do not expect it to outperform state-of-the-art trackers, but it should still reconstruct 3D tracks reliably enough to serve as a realism metric. We compare against 3D-lifted 2D trackers using Zerodepth (ZD), as well as 3D tracking methods such as SpatialTracker (Xiao et al., 2024) and SpatialTrackerV2 (Xiao et al., 2025). Since many baselines are trained primarily on synthetic Kubric3D (Greff et al., 2022), we also include CoTracker3 fine-tuned on TAPVid-3D pseudo labels. As shown in Table 1, 3DSPA outperforms most baselines and performs on par with fine-tuned CoTracker3, despite its reconstruction bottleneck.

Figure 1 further shows that, even from a 2D input video with noisy VDA depth, 3DSPA reconstructs coherent 3D tracks and assigns low reconstruction quality to physically implausible motion such as the hammer morphing. Together, these results show that 3DSPA can reconstruct 3D point tracks reliably enough to support reference-free video realism evaluation.

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 dataset (Bordes et al., 2025). 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.

Evaluation protocol. For each video v, we compute a scalar plausibility score $s(v)$ from the Average Jaccard (AJ) between the reconstructed query trajectories and their corresponding held-out target trajectories. For each benchmark-defined matched possible–impossible comparison (v^+, v^-), we count a correct decision when $s(v^+) > s(v^-)$. The win rate is the percentage of matched comparisons satisfying this condition, aggregated by physical principle and camera setting. Higher AJ therefore

Model / Category	Permanence	Immutability	Continuity	Solidity
GPT-4o (Hurst et al., 2024)	58.82	59.56	57.35	55.32
Qwen-VL 2.5 (Bai et al., 2025)	54.41	53.68	54.41	51.06
Qwen-VL 3.5 (Qwen Team, 2026)	56.00	55.40	55.99	53.79
Gemini-1.5 Pro (Team et al., 2024)	55.88	56.73	54.80	56.73
Gemini-2.5 Flash (Comanici et al., 2025)	58.82	63.97	55.15	56.38
VideoMAEv2-g (Wang et al., 2023)	50.00	53.69	54.41	59.57
Cosmos-4B (Agarwal et al., 2025)	41.18	48.32	50.00	55.32
V-JEPA 2-h (Assran et al., 2025)	67.65	56.38	57.35	52.13
V-JEPA-h+RoPE (Bardes et al., 2024)	57.35	58.72	**75.00**	58.51
TRAJAN (Allen et al., 2025)	50.00	58.82	48.53	52.13
3DSPA (no 3D)	**76.47**	73.08	73.53	59.57
3DSPA (no DINO)	60.29	66.18	58.82	39.36
3DSPA (ours)	75.00	**76.47**	69.12	**64.47**
Human	99.26	90.44	94.44	95.21

Table 2: **IntPhys2 win rates (%) with a moving camera.** We compare prior models, 3DSPA, and its ablations against human judgments; bold highlights the best non-human result per column.

Model / Category	Permanence	Immutability	Continuity	Solidity
GPT-4o (Hurst et al., 2024)	59.62	58.65	54.81	56.73
Qwen-VL 2.5 (Bai et al., 2025)	53.85	56.73	52.88	50.96
Qwen-VL 3.5 (Qwen Team, 2026)	56.58	59.43	54.92	52.63
Gemini-1.5 Pro (Team et al., 2024)	55.77	56.73	54.80	56.73
Gemini-2.5 Flash (Comanici et al., 2025)	64.42	59.62	54.81	55.77
VideoMAEv2-g (Wang et al., 2023)	63.46	54.81	65.38	48.08
Cosmos-4B (Agarwal et al., 2025)	51.92	50.96	53.85	48.08
V-JEPA 2-h (Assran et al., 2025)	63.46	51.92	50.00	50.00
V-JEPA-h+RoPE (Bardes et al., 2024)	59.62	55.77	57.69	46.15
3DSPA (no 3D)	61.54	**75.00**	**78.85**	69.23
3DSPA (no DINO)	65.38	46.15	50.00	38.46
3DSPA (ours)	**76.92**	73.08	67.31	**70.77**
Human	100.0	97.11	99.04	96.15

Table 3: **IntPhys2 win rates (%) with a static camera.** We compare prior models, 3DSPA, and its ablations against human judgments; bold highlights the best non-human result per column.

indicates greater compatibility with the learned motion prior, rather than serving as a direct test of a particular physical law. No IntPhys2 realism labels are used to train or fine-tune 3DSPA. If the scores are tied, we dont consider them in our scoring.

Baselines and ablations We compare 3DSPA against several state-of-the-art multimodal models (MLLM), self-supervised vision foundation models, and trajectory-based realism metrics, including TRAJAN (Allen et al., 2025). In addition to these external baselines, we evaluate ablations of 3DSPA that isolate the contribution of 3D geometry and semantic information. **3DSPA (no DINO)** keeps the 3D trajectory representation and depth cues, but removes DINOv2 semantic features. **3DSPA (no 3D)** keeps the semantic features, but removes the 3D coordinate and depth information, operating only on 2D point trajectories. Together, these ablations separate the effects of semantic context, 3D structure, and their combination, in detecting physical rule violations.

Performance analysis. Tables 2 and 3 report results for moving and static cameras, respectively. Across the four physical principles, 3DSPA is competitive with or improves upon strong perception and vision–language baselines. Its full representation obtains the highest non-human win rate on immutability and solidity with a moving camera, and on permanence and solidity with a static camera. The no-3D ablation is strongest in several other conditions, indicating that semantic information accounts for a substantial part of the discrimination signal. These comparisons do not imply that AJ directly evaluates a Newtonian equation; rather, it captures whether the observed motion and visibility are compatible with a learned semantic 3D reconstruction prior.

Most of the benefits of 3DSPA in determining possible vs. impossible physics appear to come from the inclusion of DINOv2 features. Although 3DSPA performs best overall, the ablation which removes 3D structure but retains DINOv2 features performs comparably to 3DSPA in several concept categories, indicating that semantic information is essential for recognizing many physical rule violations. At the same time, geometry alone is insufficient — the variant without DINOv2 performs substantially worse, especially on immutability and solidity. These results suggest that 3D structure should be viewed as a complementary signal rather than the sole source of physical realism. 3D structure is

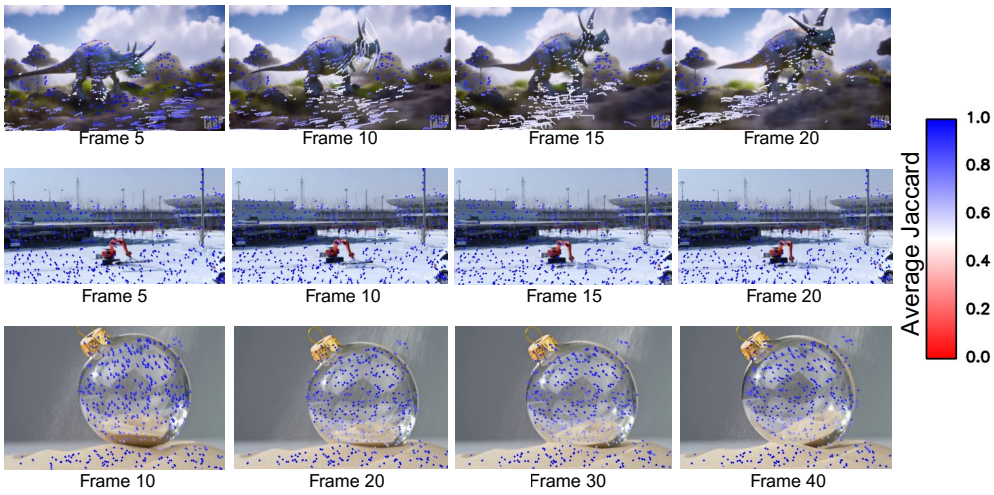


Figure 3: **Qualitative examples of accurate reconstruction on realistic generated videos from EvalCrafter and VideoPhy-2.** High-rated videos produce coherent and stable point tracks under 3DSPA (blue). The top and middle examples are from EvalCrafter (Liu et al., 2024b), and the bottom example is from VideoPhy-2 (Bansal et al., 2025).

most useful for depth-aware motion, occlusion, articulated movement, and object interaction, while semantic features provide the object-level context needed to interpret whether a motion is physically plausible. We provide additional visualizations 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.

We draw on two datasets which include a large set of videos generated by multiple generative video models: VideoPhy-2 (Bansal et al., 2025) and EvalCrafter (Liu et al., 2024b). VideoPhy-2 emphasizes action-centric videos and includes human annotations of physical commonsense and semantic adherence to the text prompt. EvalCrafter (Liu et al., 2024b) evaluates video quality with a larger set of five metrics including motion quality, temporal consistency, and several prompt adherence measures from 11 video generative models. We measure automated metric quality by computing the Spearman rank correlation between model scores and human ratings. For 3DSPA, each video is scored using the Average Jaccard (AJ) of reconstructed query tracks, and we evaluate this reconstruction-based score by its Spearman correlation with human judgments. Figure 3 shows representative high-rated videos from both datasets, where 3DSPA reconstructs coherent and temporally stable point tracks even under transparency and substantial motion. We show more examples of in Appendix G.

VideoPhy-2. 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, so we use PC as the human judgments. VideoPhy-2 also provides VIDEOPHY-2 AutoEval, a task-specific automatic evaluator fine-tuned on VideoPhy-2 human judgments to predict physical commonsense scores.

As shown in Table 4, we compare against VLM-based and video-evaluation baselines which have been used previously to evaluate physics in videos. In our setting, 3DSPA outperforms video-language evaluators such as VideoCon (Bansal et al., 2024a), VideoLLaVA (Lin et al., 2024) and VideoScore (He et al., 2024), blind video quality assessment models such as DOVER (Wu et al., 2023a), and the trajectory-based TRAJAN baseline (Allen et al., 2025). Notably, 3DSPA also outperforms VideoCon–Physics (Bansal et al., 2024b), a task-specific evaluator trained for physical commonsense and semantic alignment. 3DSPA also substantially outperforms its ablations in tracking

Model	Spearman (PC) ↑
Dataset-specific evaluator	
VIDEOPHY-2 AutoEval (Bansal et al., 2025)	**0.76**
Reference-free / non-VideoPhy-2-trained methods	
DOVER (Wu et al., 2023a)	0.16
VideoCon (Bansal et al., 2024a)	0.13
VideoCon–Physics (Bansal et al., 2024b)	0.48
VideoLLaVA (Lin et al., 2024)	0.08
VideoScore (He et al., 2024)	0.17
TRAJAN (Allen et al., 2025)	0.19
3DSPA (ours)	**0.74**

(a) Comparison with existing baselines.

Model Variant	Spearman (PC) ↑
3DSPA (no 3D)	0.40
3DSPA (no DINO)	0.50
3DSPA (ours)	**0.74**

(b) Ablation of 3D and semantic cues.

Table 4: **Spearman rank correlation on VideoPhy-2 physical commonsense (PC).** Left: comparison with video-language evaluators, learned video quality metrics, a task-specific physical commonsense evaluator, and the trajectory-based TRAJAN baseline. Right: ablation of 3DSPA showing the contribution of 3D information and DINO semantic features. 3DSPA achieves the strongest reference-free performance.

Model	Visual Quality	T2V	Motion Quality	Consistency	Subjective Likeness
DOVER (Wu et al., 2023a)	0.41	0.22	0.20	0.33	0.3
TRAJAN (Allen et al., 2025)	0.28	0.25	0.24	0.33	0.23
3DSPA (no 3D)	0.31	0.44	0.41	0.39	0.46
3DSPA (no DINO)	0.45	0.55	0.49	0.48	**0.63**
3DSPA (ours)	**0.48**	**0.58**	**0.55**	**0.60**	0.60

Table 5: Spearman rank coefficients between human annotations and automated AJ scores for the EvalCrafter (Liu et al., 2024b) dataset in the categories of visual quality, text-to-video similarity, motion quality, temporal consistency, and subjective likeness across different ablations.

human ratings of physical commonsense. We provide a qualitative comparison with TRAJAN in Appendix F, showing that 2D trajectory reconstruction can fail on smooth but semantically implausible videos, while 3DSPA captures these failures through semantic 3D trajectory reconstruction.

VIDEOPHY-2 AutoEval is an even stronger dataset-specific evaluator, as it is trained directly on VideoPhy-2 human judgments and reaches a Spearman correlation of 0.76. In comparison, 3DSPA reaches 0.74 *without dataset-specific fine-tuning or labeled realism supervision*. This suggests that modeling semantic 3D motion provides a stronger signal for physical commonsense evaluation, allowing us to avoid task-specific fine-tuning. As an additional comparison requested during review, we evaluated three dimensions of VBench++ (Huang et al., 2025) against VideoPhy-2 physical-commonsense annotations. The Spearman correlations for subject consistency, background consistency, and motion smoothness are 0.199, 0.211, and 0.268, respectively, compared with 0.742 for 3DSPA. These results suggest that common appearance- and smoothness-focused consistency scores do not, by themselves, capture the same physical-commonsense judgments as semantic 3D trajectory reconstruction. We additionally analyze the computational cost of 3DSPA and representative baselines. 3DSPA achieves comparable or faster inference time relative to VLM-based evaluators and TRAJAN. Detailed runtime breakdowns are provided in Appendix C.

EvalCrafter For each video, a set of human annotators rated the **visual quality**, **text to video consistency**, **motion quality**, **temporal consistency**, and **subjective likeness**. Similar to Videophy-2, we compute Spearman rank coefficients between human ratings and the Average Jaccard (AJ) for each of the ablations 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 5 shows that 3DSPA achieves the best performance; further highlighting that reconstructing 3D semantic point trajectories provides the strongest predictor of a variety of human annotations for generated videos.

5 Discussion

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. The ablations clarify the role of 3D in 3DSPA - the performance is not driven by geometry alone, but by the combination of semantic and 3D cues. Semantic features provide the object-level context needed to interpret physical plausibility, while 3D structure adds complementary information for depth-aware motion, occlusion, articulated movement, and object interactions. Across the evaluated settings, 3DSPA is competitive with or outperforms strong vision–language models in detecting synthetic physical rule violations such as solidity and immutability, and in tracking human physical-commonsense judgments of generated videos. Because the model is trained without realism annotations, its generalization to held-out IntPhys2, VideoPhy-2, and EvalCrafter provides empirical evidence for its use as a learned plausibility proxy, not a guarantee of physical correctness in every video. More broadly, 3DSPA also raises important societal considerations. As generated videos become more realistic, they may become more persuasive when used for synthetic misinformation or "fake news". This makes robust realism evaluation especially important, but it also underscores the need to pair such evaluation tools with AI safety and responsible deployment practices.

5.1 Limitations

Despite its strong empirical performance, 3DSPA inherits limitations from its upstream perception pipeline. Both support and query tracks are derived from estimated 2D correspondences and monocular depth, so tracking failures, depth ambiguities, motion blur, reflections, transparent objects, sparse texture, clutter, and heavy occlusion can lower AJ even when a video is physically plausible. The sensitivity analysis in Appendix D.2 finds relatively stable realism-ranking performance across four tested tracker–depth combinations, but it does not establish invariance to perception errors. Conversely, high reconstruction quality does not guarantee physical correctness: AJ measures compatibility with a learned semantic 3D motion prior, not direct satisfaction of physical equations. A low score can reflect implausible motion, perception failure, or an unfamiliar but valid motion pattern. Qualitative failure cases are shown in Appendix E.

Highly non-rigid and low-texture phenomena, such as fluids, smoke, water bodies, and waterfalls, are also important cases for further study. Although such motion can make point tracks less stable, our experiments show that 3DSPA can still capture severe violations, such as fluids appearing or disappearing unrealistically, through low localized AJ scores (Appendix H). Future work could evaluate fluid based motion more systematically.

Finally, 3DSPA operates on fixed-length clips rather than explicitly conditioning on longer motion history, which may limit detection of delayed physical effects or long-horizon realism failures. While there are already serious intermediate-horizon realism failures which 3DSPA addresses, extending 3DSPA with longer temporal context could provide stronger tests of longer-term dynamics.

5.2 Future work

More broadly, we view 3DSPA not only as an evaluation metric, but also as a step toward improving generative video models. Because 3DSPA provides a reference-free signal for semantic 3D motion consistency, future work can investigate whether such realism scores can be used to filter generated videos for real-to-sim transfer, guide model selection, or regularize the training of video generators toward more physically plausible motion.

Acknowledgments and Disclosures

KRA gratefully acknowledges support from NSERC, the Vector Institute, and CIFAR. Funding for this work was provided by KRA's CIFAR AI Chair and Azrieli Global Scholar award in the Learning in Machines and Brains program.

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.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6034–6044, 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.

Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. Videocon: Robust video-language alignment via contrast captions. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13927–13937, 2024a.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024b.

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 M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

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*, 179:41–65, 2019.

Chirui Chang, Jiahui Liu, Zhengzhe Liu, Xiaoyang Lyu, Yi-Hua Huang, Xin Tao, Pengfei Wan, Di Zhang, and Xiaojuan Qi. How far are ai-generated videos from simulating the 3d visual world: A learned 3d evaluation approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10307–10317, 2025.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024.

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*, pages 22831–22840, 2025a.

Wenhan Chen, Sezer Karaoglu, and Theo Gevers. Grab-3d: Detecting ai-generated videos from 3d geometric temporal consistency. *arXiv preprint arXiv:2512.13665*, 2025b.

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*, 2025c.

Aditya Chinchure, Sahithya Ravi, Pushkar Shukla, Vered Shwartz, and Leonid Sigal. Spotlight: Identifying and localizing video generation errors using vlms. *arXiv preprint arXiv:2511.18102*, 2025.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 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.

Jason Gelman. Veo 3 and veo 3 fast are now generally available on vertex ai. `https://cloud.google.com/blog/products/ai-machine-learning/veo-3-fast-available-for-everyone-on-vertex-ai`, July 2025. Google Cloud Blog.

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*, pages 3749–3761, 2022.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123, 2024.

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*, pages 21807–21818, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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*, pages 4651–4664. PMLR, 2021.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6013–6022, 2025.

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.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984, 2024.

Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fr\'echet video motion distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124*, 2024a.

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

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.

Ge Ya Luo, Gian Mario Favero, Zhi Hao Luo, Alexia Jolicoeur-Martineau, and Christopher Pal. Beyond fvd: Enhanced evaluation metrics for video generation quality. *arXiv preprint arXiv:2410.05203*, 2024.

Saman Motamed, Minghao Chen, Luc Van Gool, and Iro Laina. Travl: A recipe for making video-language models better judges of physics implausibility. *arXiv preprint arXiv:2510.07550*, 2025.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 948–958, 2026.

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.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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*, pages 8748–8763. PmLR, 2021.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018.

Runway. Runway gen-4: Ai video generation with world consistency. `https://runwayml.com/research/introducing-runway-gen-4`, 2025. Runway Research.

Guoqiang Shi, Aiping Tan, Hailin Tao, and Yuefeng Du. Tcmabnet: A spatio-temporal network for fine-grained fake video. In *2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA)*, pages 334–339. IEEE, 2025.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Advances in Neural Information Processing Systems*, 36:67048–67068, 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*, pages 14549–14560, 2023.

Luca Weihs, Amanda Yuile, Renée Baillargeon, Cynthia Fisher, Gary Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. Benchmarking progress to infant-level physical reasoning in ai. *Transactions on Machine Learning Research*, 2022.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20144–20154, 2023a.

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*, 2023b.

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*, pages 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, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Qin Zhang, Peiyu Jing, Hong-Xing Yu, Fangqiang Ding, Fan Nie, Weimin Wang, Yilun Du, James Zou, Jiajun Wu, and Bing Shuai. Physion-eval: Evaluating physical realism in generated video via human reasoning. *arXiv preprint arXiv:2603.19607*, 2026.

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D3: Training-free ai-generated video detection using second-order features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12852–12862, 2025.

A Training Setup & Hyperparameters

We train our model using AdamW (Loshchilov and 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, z_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 6 and 7 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 6: 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.

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.

We also give a qualitative example of how 3DSPA captures physical rule violations in Figure 5. Recall that the IntPhys2 dataset is constructed by pairing the same initial frames with either a realistic ending (real) or one which violates a physical rule (unreal). In Figure 5, the ball should interact with the ramp and fly through the air (top, real). In this case, the point tracks on the ball are well

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

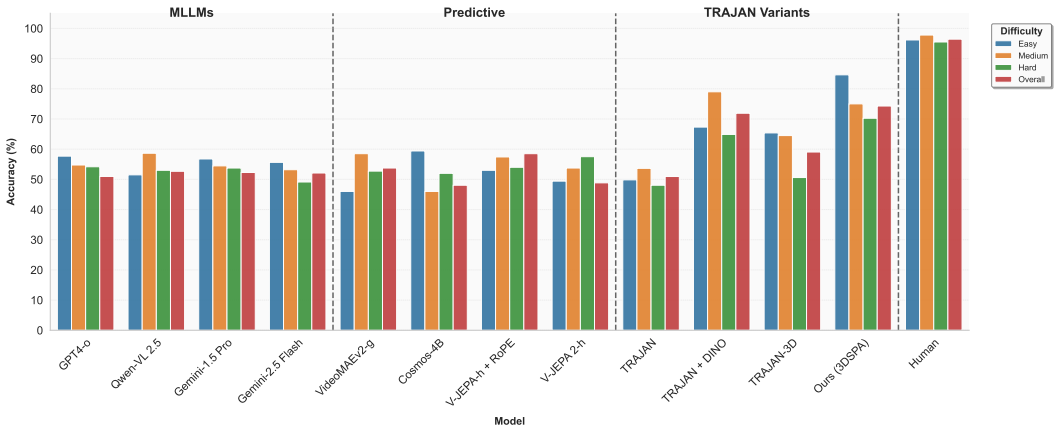


Figure 4: Performance comparison across models on the IntPhys2 benchmark for each of the *easy*, *medium* and *hard* categories.

reconstructed. However, in the bottom video, where the ball just continues down the slope without flying through the air, the point tracks are poorly reconstructed.

C Inference Time Comparison

We benchmark the inference cost of 3DSPA and representative baselines on a single NVIDIA A100-40GB GPU. All methods are evaluated on videos with 50 frames per clip, averaged over 50 different clips. Reported timings include full end-to-end processing required to produce a final evaluation

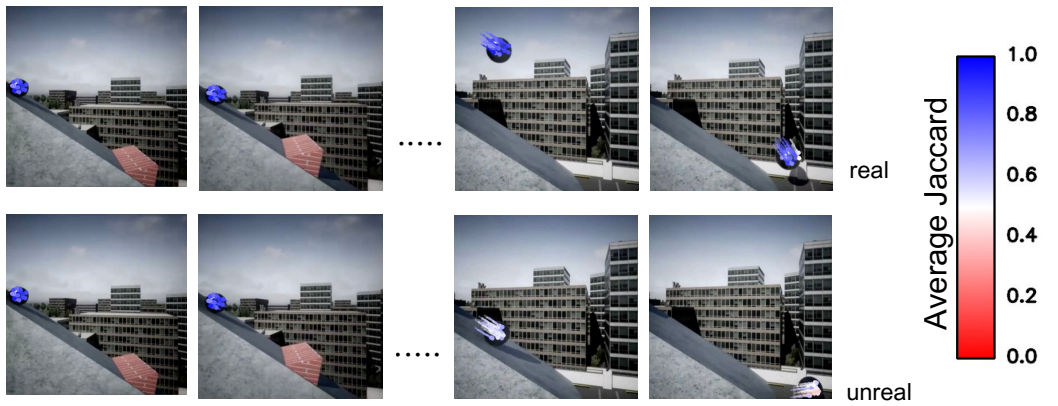


Figure 5: **Visualization of point tracks on IntPhys2 videos.** For each video, we project tracked points and color them using their Average Jaccard (AJ) consistency score (red = low / unreal, blue = high / real). Real, physically consistent videos (top) show smooth, coherent tracks with high AJ scores throughout the video, while unreal videos with physical rule violations (bottom) exhibit poorly reconstructed point trajectories.

Model	Inference Time (s)
VideoCon–Physics	8.13
VideoCon	6.21
VideoLLaVA	13.52
VideoScore	4.89
VIDEOPHY–2–AUTOEVAL	10.67
TRAJAN	2.18
3DSPA (ours)	**5.61**

Table 8: Average inference time per video for 3DSPA and TRAJAN as compared to other finetuned VLM models

Training Data	Avg AJ ↑	Avg APD ↑	Avg OA ↑
Kubric3D only	11.8	19.5	84.4
Kubric3D + TAPVid-3D-main	**14.0**	**19.8**	**85.8**

Table 9: **Effect of training data on TAPVid-3D minival.** Training with both Kubric3D and TAPVid-3D improves reconstruction performance over Kubric3D-only training, suggesting that real-world TAPVid-3D data improves generalization beyond synthetic supervision.

score for a video. Note that for 3DSPA, the pipeline includes four stages: (1) CoTracker3 for dense 2D point tracking, (2) Video Depth Anything (VDA) for per-frame depth estimation, (3) DINOv2 for semantic feature extraction, and (4) 3DSPA autoencoder inference for trajectory encoding and reconstruction. All components are executed sequentially on GPU without batch parallelization across videos.

Table 8 reports the average inference time per video. Despite incorporating multiple perception modules, 3DSPA takes almost similar time as most VLM-based evaluators while offering stronger sensitivity to physical realism and motion artifacts. Compared to TRAJAN, 3DSPA incurs approximately a $2\times$ overhead due to depth and semantic feature extraction, which enables richer 3D semantic reasoning.

D Robustness and Sensitivity Analysis

We evaluate sensitivity to training data, upstream tracking/depth quality, and support–query design. Across the settings tested here, 3DSPA shows moderate variation, suggesting that its reconstruction signal is not limited to a single dataset choice or idealized perception input, while remaining sensitive to upstream perception quality.

D.1 Effect of Training Data

3DSPA is trained using both Kubric3D and TAPVid-3D. Kubric3D provides accurate synthetic 3D supervision, while TAPVid-3D exposes the model to real-world appearance, motion, and occlusion patterns. To evaluate the role of training data composition, we compare the full training setup against a Kubric3D-only variant on TAPVid-3D minival.

As shown in Table 9, adding TAPVid-3D improves average AJ from 11.8 to 14.0 and OA from 84.4 to 85.8. This indicates that synthetic 3D supervision is useful, but real-video trajectories improve generalization and reduce overfitting to synthetic scene statistics. Since IntPhys2, VideoPhy-2, and EvalCrafter are held out and distributionally distinct from the training datasets, this also suggests that 3DSPA learns transferable semantic 3D motion priors rather than memorizing dataset-specific patterns.

Input Setting	Avg AJ ↑	Avg APD ↑	Avg OA ↑
GT tracks + GT depth	**15.8**	**21.2**	86.6
GT tracks + Pred. depth	15.1	20.7	86.4
Pred. tracks + GT depth	14.9	20.8	**86.9**
Pred. tracks + Pred. depth	14.0	19.8	85.8

Table 10: **Sensitivity to tracking and depth estimation on TAPVid-3D minival.** Performance degrades only modestly when ground-truth tracks and depth are replaced with predicted inputs, indicating robustness to realistic upstream perception errors.

Tracker + depth estimator	IntPhys2 win rate (%) ↑	VideoPhy-2 Spearman ↑
CoTracker3 + VDA (default)	71.6	**0.742**
BootsTAPIR + VDA	**72.0**	0.726
CoTracker3 + Depth Anything V2	69.7	0.716
BootsTAPIR + Depth Anything V2	70.2	0.719

Table 11: **Sensitivity of downstream realism evaluation to upstream tracker and depth estimator.** Results are the additional experiments reported in the author rebuttal. The same trained 3DSPA model is evaluated using different upstream perception combinations.

D.2 Sensitivity to Tracking and Depth Estimation

At inference time, 3DSPA uses external modules to estimate 2D tracks, depth, and semantic features. These modules provide perceptual measurements, but they do not define the realism score. The score comes from whether the resulting 3D trajectories with semantic information can be compactly encoded and reconstructed by the learned motion prior.

To test whether performance is dominated by upstream tracking or depth quality, we replace ground-truth tracks and depth with predicted inputs on TAPVid-3D minival. Table 10 shows that average AJ decreases from 15.8 with ground-truth tracks and depth to 14.0 with predicted tracks and predicted depth, while OA remains similar. This degradation quantifies how the reconstruction metrics change with measured upstream errors under these inference-time conditions.

This experiment alone does not disentangle perception quality from physical plausibility, but it shows the amount of reconstruction degradation on TAPVid-3D when the perception inputs are predicted rather than ground truth. We separately test downstream realism-ranking robustness with different tracker–depth modules below.

Downstream realism-ranking sensitivity. We additionally evaluated four combinations of 2D trackers and monocular depth estimators: CoTracker3 or BootsTAPIR, each paired with VideoDepthAnything (VDA) or Depth Anything V2. Table 11 reports IntPhys2 pairwise win rates and Spearman correlations with human physical-commonsense ratings on VideoPhy-2. IntPhys2 performance varies from 69.7% to 72.0%, and VideoPhy-2 Spearman correlation varies from 0.716 to 0.742 across the tested configurations. Thus, the ranking signal persists when replacing the default perception models, although the score is not fully independent of upstream tracking and depth errors.

D.3 Sensitivity to Support-Query Partition and Point-Track Density

We further evaluate the sensitivity of 3DSPA to two reconstruction design choices: the support–query partition and the number of sampled point tracks. The support–query split is performed over point tracks, not video frames, so support tracks span the full video while held-out query tracks are reconstructed for evaluation.

Figure 6 shows that performance on IntPhys2 remains stable across 25:75, 50:50, and 75:25 support–query splits, with the standard 50:50 split performing best. This suggests that 3DSPA is not overly dependent on the exact partition, provided that enough support tracks encode the scene dynamics and enough query tracks remain for evaluation.

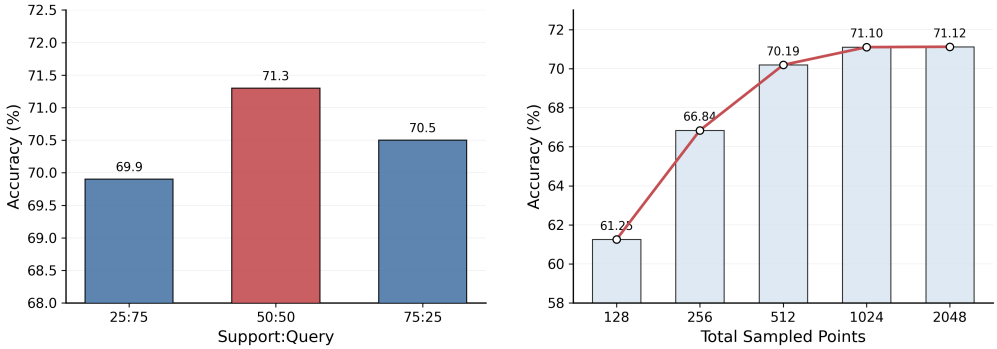


Figure 6: **Sensitivity to support–query partition and sampled point-track density on IntPhys2.** Left: 3DSPA remains stable across different support–query point-track partitions, with the standard 50:50 split performing best. Right: performance improves as more point tracks are sampled and saturates near 1024 points, indicating that dense but not excessive point coverage is sufficient for stable realism evaluation.

We also vary the total number of sampled point tracks. Accuracy improves from 128 to 512 points and saturates near 1024 points, with only marginal gains from 1024 to 2048. We therefore use 1024 points in the main experiments to ensure dense spatial coverage across objects, boundaries, and background regions and it is faster.

These results also clarify why localized failures, such as small-object violations or object disappearance, can affect the score. Since AJ is computed over individual trajectories and visibility states, such failures create repeated geometric, depth, semantic, and occlusion inconsistencies across tracks and time. These structured errors accumulate over the video rather than appearing as isolated noise, allowing 3DSPA to capture localized physical violations.

E 3DSPA successes and failures on 3D point track reconstruction

In this section, we give some examples of how well 3DSPA can reconstruct 3D point tracks on the TAP-Vid3D-Minival dataset. In Figure 7, we show successes of 3DSPA in reconstructing point tracks of both small objects and fast moving objects. However, 3DSPA struggles when depth cues are difficult to interpret (Figure 8). Since we rely on VideoDepthAnything (Chen et al., 2025a), when depth cues are ambiguous, we see poorer 3D tracking. However, the 2D tracks still look reasonable.

F Semantic 3D Reconstruction vs. 2D Trajectory Reconstruction

TRAJAN (Allen et al., 2025) and 3DSPA are both reconstruction-based realism metrics, but they evaluate different notions of video consistency. TRAJAN reconstructs 2D point trajectories and is therefore primarily sensitive to image-plane temporal coherence. This is effective for detecting flicker or unstable motion, but smooth 2D trajectories can still correspond to physically implausible videos. For example, an object may disappear gradually, change identity, or move in a depth-inconsistent way while still producing temporally smooth 2D tracks.

By reconstructing 3D trajectories with occlusion states and semantic context, 3DSPA evaluates not only where points move, but also what object they belong to, how that object moves in 3D, and whether its visibility and persistence are physically consistent over time. This semantic component is particularly important for detecting failures that are not obvious from motion smoothness alone. A disappearing phone, for instance, may have smooth 2D tracks, but it violates object permanence and semantic identity consistency. Similarly, articulated motion such as walking requires understanding coherent 3D limb motion rather than only tracking projected 2D displacement.

Figure 9 illustrates these differences on two examples from EvalCrafter (Liu et al., 2024b). In the first video, a dog walks on grass and receives a high human motion-quality rating. 3DSPA reconstructs the articulated leg motion coherently in 3D, while TRAJAN produces noisier and less stable reconstructions, incorrectly penalizing a realistic video. In the second video, a phone gradually

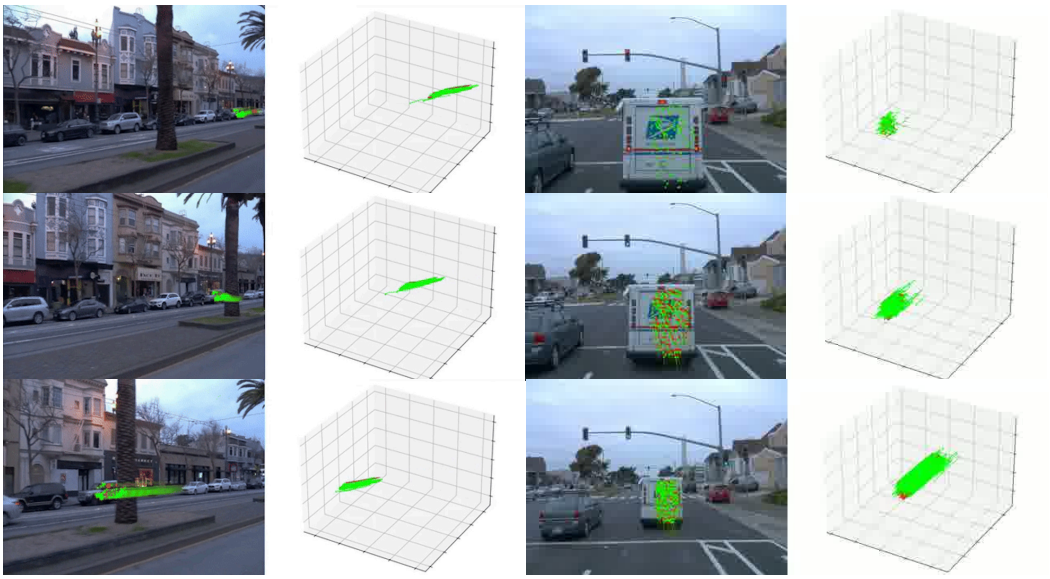


Figure 7: **Positive examples where 3DSPA performs well on TAP-Vid3D-Minival.** Predicted 3D trajectories (green) align closely with ground truth (red) even for videos with small objects and high motion. These videos are from the DriveTrack dataset within TAP-Vid3D-Minival.

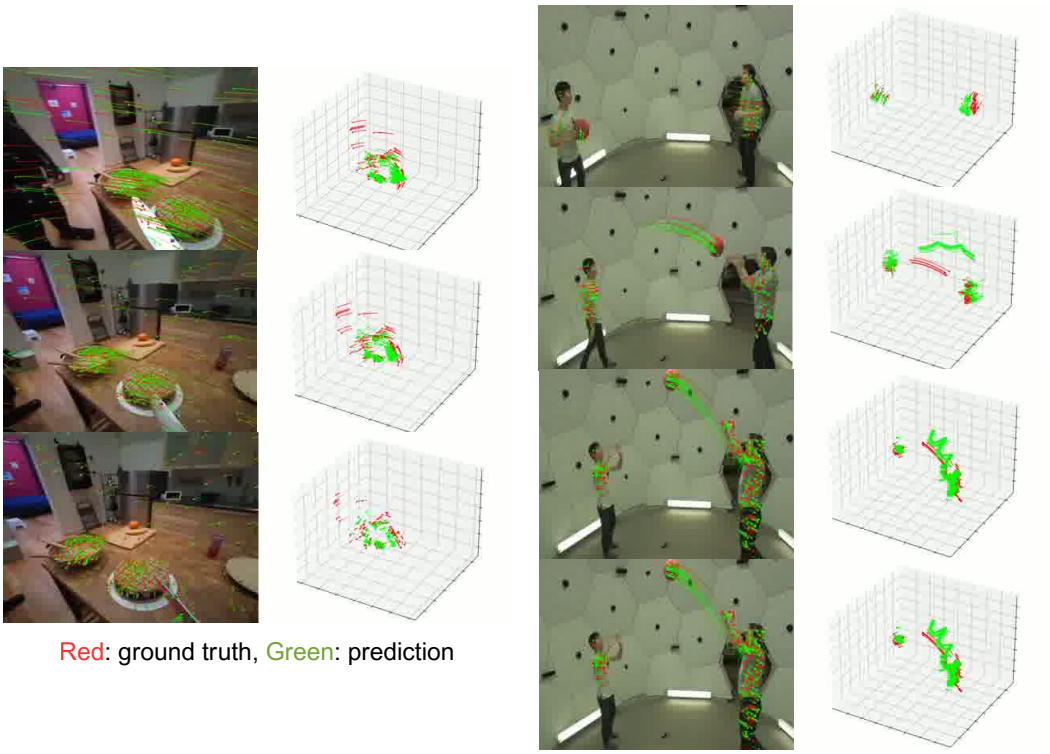


Figure 8: **Failure cases of 3DSPA on TAP-Vid3D-Minival.** We visualize ground-truth (red) and predicted (green) query 3D point trajectories for challenging scenes. Errors typically arise in regions with complex geometry or depth ambiguities, giving high error in depth prediction, although the motion following is good. Corresponding 2D frames (left) and reconstructed 3D trajectories (right) highlight mismatches between predicted and true motion, which mainly arise from incorrectly estimated depth.

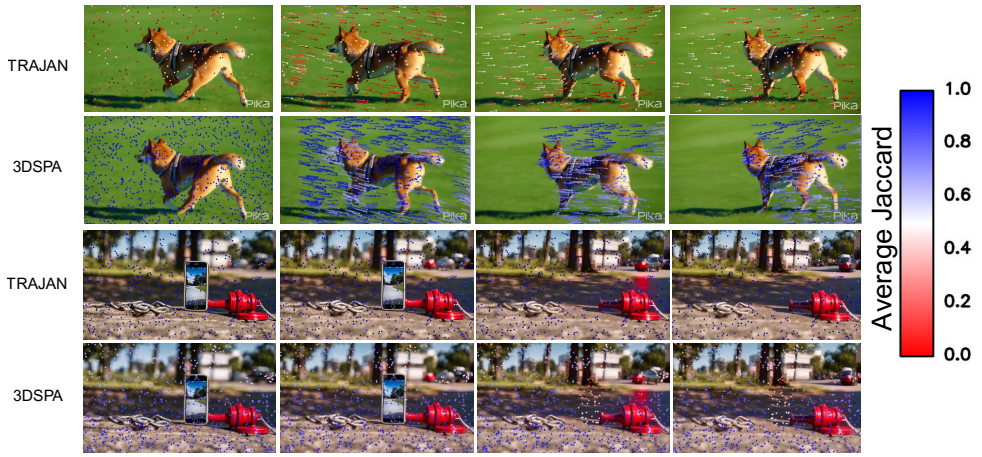


Figure 9: **TRAJAN vs. 3DSPA.** Qualitative comparison on videos from EvalCrafter (Liu et al., 2024b). TRAJAN (Allen et al., 2025) evaluates 2D trajectory reconstruction and primarily captures image-plane temporal consistency, whereas 3DSPA reconstructs semantic 3D trajectories with occlusion information. In the top example, a realistic dog-walking video receives a high human rating (4.5/5); 3DSPA reconstructs coherent articulated motion, while TRAJAN produces unstable tracks. In the bottom example, a phone gradually disappears and receives a low human rating (1.67/5); TRAJAN fails because the 2D motion remains smooth, whereas 3DSPA detects the semantic and physical violation.

disappears. Because the 2D motion remains smooth, TRAJAN assigns a relatively high realism score. In contrast, 3DSPA flags the video as unrealistic because the disappearance creates semantic, geometric, and visibility inconsistencies. These examples show that the gain over TRAJAN does not come from simply adding stronger features, but from changing the realism signal: 3DSPA evaluates whether semantic 3D trajectories can be reconstructed under a learned physical motion prior.

G Visualizations showing 3DSPA's ability to capture bad motion in generated videos

Here we show 3DSPA's ability to capture unrealistic motion across a selection of generated videos. In Figure 11, we instead show examples of videos that were rated *poorly* for motion realism. In these cases, objects are morphing shape or inappropriately disappearing. The tracks are poorly reconstructed exactly where the motion becomes unrealistic.

H Additional Analysis on Fluid Motion

Highly non-rigid and low-texture phenomena such as fluids, smoke, water bodies, and waterfalls remain challenging for trajectory-based realism evaluation. While EvalCrafter and VideoPhy-2 contain such videos these phenomena often appear as background motion. In many cases, human annotators primarily judge the realism of foreground subjects, such as humans, animals, or salient objects, rather than the fine-grained realism of the fluid itself. As a result, these benchmarks do not cleanly isolate fluid-specific realism, making it difficult to determine whether an automated metric is matching human judgments of fluid motion or instead matching judgments of the foreground scene.

There is an additional evaluation challenge - many fluid regions exhibit weak apparent motion. Since our EvalCrafter evaluation filters out videos with very low 3D point-track motion, as described in Section 4.3, some videos with subtle fluid motion are excluded from the main quantitative analysis. Therefore, we include qualitative fluid examples in Figure 10 to better illustrate both the strengths and limitations of 3DSPA in this setting.

In the top row, the generated video contains a severe physical violation: the wine progressively appears in the glass rather than being consistently poured through physically plausible motion. 3DSPA assigns low localized AJ scores to the inconsistent regions, showing that the model can detect large-scale fluid

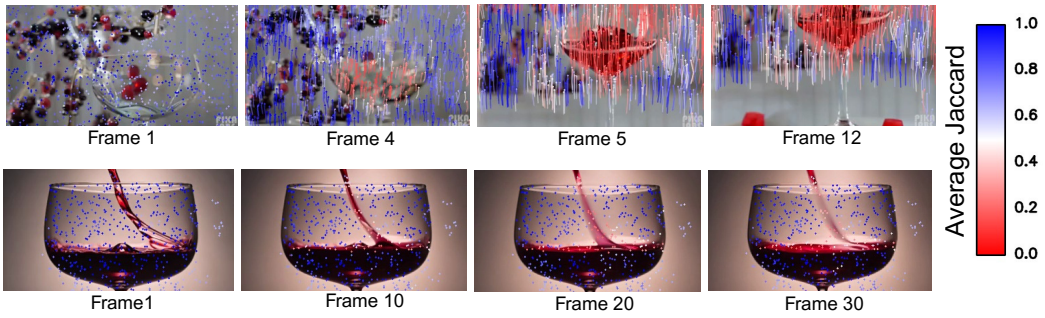


Figure 10: **Fluid-motion examples.** We visualize 3DSPA reconstructions on generated videos involving wine-like fluid motion. Blue points indicate high AJ, red points indicate low AJ, and white points indicate intermediate AJ. **Top:** This video is generated by Pika model. 3DSPA detects a severe physical violation where the wine appears or accumulates unrealistically as the video progresses, producing low localized AJ scores in the inconsistent fluid regions. **Bottom:** This video is generated by Modelscope model. the generated fluid motion is more plausible overall, but still contains subtle artifacts such as partial disappearance or transparency of the liquid across frames. In this case, 3DSPA produces a medium AJ score, highlighting a limitation: fine-grained fluid artifacts can be harder to capture than large object-level or appearance/disappearance violations.

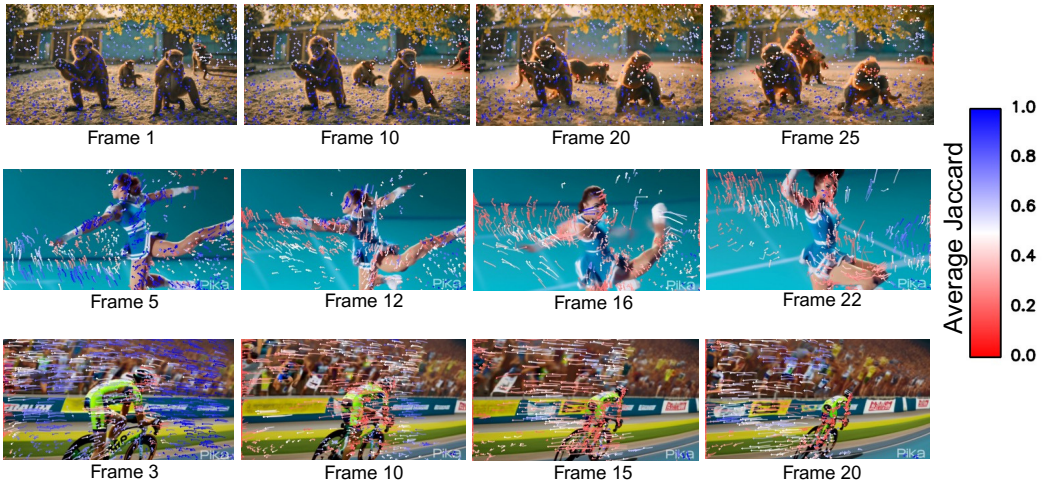


Figure 11: **Additional Visualization on Evalcrafter & Videophy-2.** Low-rated motion videos lead to lower AJ scores and unstable point tracks from 3DSPA, reflecting difficulty in reconstructing unreliable motion. The top videos are from Evalcrafter (Liu et al., 2024b), middle and bottom are from Videophy2 (Bansal et al., 2025).

appearance and disappearance errors. In the bottom row, the generated wine motion is more plausible overall, but still contains subtle failures: parts of the liquid become transparent or disappear across frames. In this case, 3DSPA assigns a medium AJ score. This means while 3DSPA can capture strong fluid-related violations, it is less sensitive to fine-grained fluid artifacts, especially when the motion is low-texture, partially transparent, or not the dominant source of human-rated realism. We view this as an important limitation and future direction. Incorporating stronger physics-aware priors for non-rigid motion, or evaluating on datasets where fluid dynamics are the primary annotated subject, could provide a more systematic test of fluid and weather realism.

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: In the abstract and introduction, we propose a reference-free 3D semantic point autoencoder for evaluating video realism, including the ability to detect physical rule violation and align with human judgments. These claims are supported by the experimental results 4.1–4.3. The paper also clarifies that AJ is an operational proxy for compatibility with a learned semantic 3D motion prior rather than a direct measure of realism.

   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: The paper includes a dedicated Limitations section.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: The paper is primarily empirical and does not present formal theoretical results, theorems, or proofs.

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: Refer to section 3 and 4 and appendix A for all the information. The paper describes the model architecture, training setup, inference pipeline, datasets, evaluation metrics, and baselines in sufficient detail to reproduce the main experiments. Additional hyperparameters and training details are provided in Appendix A, and sensitivity/ablation analyses are provided in the appendix. Public code and model checkpoint instructions are available at `https://github.com/TheProParadox/3dspa_code`.

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: [No]

   Justification: Training and inference code, a TAPVid-3D evaluation script, model checkpoint instructions, and links to public benchmark datasets are available at `https://github.com/TheProParadox/3dspa_code`. However, complete evaluation scripts for all reported benchmark comparisons and baselines are not yet packaged for end-to-end reproduction.

   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: The paper specifies the training datasets, train/evaluation splits, optimizer, learning rate schedule, number of epochs, batch size, support-query split, number of sampled tracks, input models, and evaluation metrics. These details are given in Section 3 and Appendix A.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

Justification: The paper reports quantitative results across multiple datasets and ablations, including Spearman correlations and win rates, but it does not report error bars, confidence intervals, or statistical significance tests for the main experimental comparisons. We did not train multiple models, so results are always computed for one model on the full test datasets available, as also done in the prior work which we compare to.

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: [No]

   Justification: Section 3 reports training on 8 H100 GPUs, and the inference-time appendix reports single-A100 (40 GB) timings. Total training GPU-hours, individual run durations, and overall project compute are not reported.

   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 work uses existing public or standard research datasets and focuses on evaluating generated-video realism, physical plausibility, and alignment with human judgments. We do not collect private user data, identify individuals, or deploy the system for surveillance or deceptive media generation. We also discuss broader societal risks in the paper, including the possibility that increasingly realistic generated videos may increase the persuasiveness of synthetic misinformation, and emphasize the need for detection, provenance, and responsible deployment practices in the discussion section.

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: The paper discusses positive societal impacts by motivating applications in robotics, embodied AI, virtual reality, education, and creative industries, where reliable realism evaluation can help assess and improve generated videos. It also discusses negative societal impacts in the Discussion, noting that increasingly realistic generated videos may become more persuasive when used for synthetic misinformation or "fake news," and emphasizing the need for detection, provenance, and responsible deployment practices.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model, scraped dataset, or dual-use foundation model. The proposed method is an evaluation metric for generated-video realism.

    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: [No]

    Justification: The paper cites the existing datasets, models, and software it uses, but does not yet explicitly document and verify the licenses and terms of use for every relevant asset.

    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: The public `https://github.com/TheProParadox/3dspa_code` repository includes a README describing the architecture implementation, dependencies, training and inference examples, benchmark data links, and model-checkpoint access.

    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 conduct new crowdsourcing experiments or new human-subject studies.

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 new human-subject data collection or new crowdsourcing experiments.

    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: LLMs are not an important, original, or non-standard component of the core method.

    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.