JOINT MODELING OF VISUAL AND PHYSICAL FIELDS FOR CONSISTENT VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-video generation models have achieved remarkable progress in visual fidelity and temporal coherence, yet generated videos can still exhibit implausible motion dynamics and inconsistent interactions. Existing efforts incorporate physical information (e.g., structured motion or geometric knowledge) mainly via external guidance or constraints, failing to directly consider it within the visual generative representation. In this paper, we proposed to employ physical field signals (e.g., velocity, acceleration, and depth) as a quantitative outcome of motion which naturally serves as unified structured representation across different motion scenarios, and then simultanesouly optimize phyical and visual signals in representation space. However, physical field signals is sparse and noisy in real videos, posing siginificant difficulties in the joint optimization. To tackle these issues, we propose a novel VideoPF method that jointly models video data and physical field signals within a unified generative framework, enabling the synthesis of videos with improved motion consistency. Specifically, we first analyze the degree to which generative models capture physical dynamics. Based on this analysis, we design an adaptive physical-aware data pipeline. To facilitate effective fusion of visual and physical information, we introduce a Gated Physical DiT that selectively integrates physical field signals into the visual generation process. To enhance learning stability and encourage physically plausible dynamics, we propose a Physical Consistency Refinement mechanism and dynamically assess the model's understanding of physical information. Extensive experiments demonstrate the effectiveness of our proposed VideoPF framework.

1 INTRODUCTION

Recent advances in video generation have substantially improved visual fidelity and temporal coherence, enabling generative models to synthesize increasingly realistic and diverse dynamic scenes (Xue et al., 2025a; Sun et al., 2024; Xing et al., 2024). However, visually coherent videos can still exhibit inconsistent motion and implausibly physical interactions, particularly in scenarios involving object contact, collisions, or other temporal changes (Kang et al., 2025; Lin et al., 2025).

These failures suggest that improving video generation requires more than modeling appearance and semantic correlations: generative models also need representations that capture how scene content evolves over space and time. A natural way to provide such capability is through physical information (Bansal et al.; 2025; Meng et al., 2025). In dynamic scenes, velocity and acceleration describe how motion evolves over time, while depth provides complementary information about the spatial configuration of objects. Unlike visual appearance alone, these field representations explicitly expose structured patterns of motion and geometry that are otherwise only implicitly encoded in visual features. They are observable from video and provide complementary descriptions of motion.

Existing works that incorporate physical information into video generation can be broadly categorized into two paradigms. One line of research integrates external reasoning modules or physics simulators to provide guidance during generation (Wang et al.; Liu et al., 2024; Xue et al., 2025b). Although effective in constrained environments, these approaches introduce substantial inference overhead, increase system complexity, and often lack scalability to open-domain video synthesis. Another line of work (Chefer et al.; Zhang et al.) promotes temporal smoothness or leverages larger datasets in the hope that physical realism emerges implicitly in the latent representations. Although

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

these methods demonstrate the value of structured physical information, the physical signals are still primarily treated as conditions or are introduced only at specific stages of the generation pipeline.

Despite their differences, existing approaches primarily treat physical information as an external signal that guides or constrains visual generation, rather than modeling physical and visual representations jointly within the generative representation. Consequently, the physical information is not directly integrated into the representation learned for visual synthesis. This raises a fundamental question: can visual and physical field representations be jointly modeled within a unified generator, allowing physical information to directly participate in learning the generative representation? However, constructing such a joint representation from large-scale video data presents two practical challenges. First, videos containing clearly observable physical dynamics are relatively sparse in natural video collections. Second, physical fields estimated from real videos, such as optical-flow-derived velocity and monocular depth, are noisy and do not provide exact physical supervision.

To address the issues, we propose the VideoPF framework, a novel approach that jointly models physics-inspired fields and visual contents for motion consistent video generation. We first systematically analyze the degree to which generative models capture physical dynamics and construct an adaptive physical-aware data pipeline that integrates real-world video data with structured simulation-derived motion-related auxiliary fields. Rather than supervising individual physical laws, we introduce a physical field abstraction that encodes structured dynamic information. We further design a gated physics attention mechanism to selectively incorporate physical signals into the generative backbone, enabling structured yet flexible modeling of dynamic interactions. To align physical representations with visual generation, we formulate auxiliary learning objectives that encourage joint representation learning across modalities. Finally, we introduce an adaptive regularization strategy that dynamically modulates auxiliary supervision based on physical parameters and temporal progression, enabling context-aware and stable training.

Our contributions can be summarized as follows:

- We propose VideoPF, a unified physical field-guided video generation framework with a Gated Physical DiT for adaptive joint modeling of visual and physical fields and a Physical Consistency Refinement mechanism for field-level supervision.
- To the best of our knowledge, we are the first to formulate physical fields, including velocity, acceleration, and depth, as structured auxiliary representations for video generation, and investigate their joint modeling with visual representations.
- We empirically investigate the degree to which current generative models understand physical phenomena, and design an adaptive physical-aware data pipeline that dynamically retrieves and generates data better suited to the current state of the model.

2 Related Works

Video Generation Models Video generation aims to synthesize high-quality videos and is one of the most important problems. Early approaches to video generation (Vondrick et al., 2016; Skorokhodov et al., 2022) were predominantly based on Generative Adversarial Networks (Goodfellow et al., 2020), where temporal dynamics were modeled through spatiotemporal convolutions or factorized architectures. With the rapid advancement of diffusion, the research focus in video generation has gradually shifted toward diffusion-based (Ho et al., 2020; Song et al.; Rombach et al., 2022) or flow-matching-based approaches (Lipman et al.; Liu et al., a). Initial efforts extended text-to-image generation frameworks to the video domain (Chen et al., 2024a; Girdhar et al., 2024; Singer et al.; Guo et al.; Chen et al., 2025; 2024b; Pan et al., 2025), employing architectures such as 3D U-Nets or incorporating spatiotemporal attention to capture temporal dependencies. Subsequently, owing to the strong scalability of Diffusion Transformers, the performance of models have been continuously improved (Brooks et al., 2024; Gao et al., 2025; Kong et al., 2024; Wan et al., 2025; Gupta et al., 2024; Yang et al.). As the success of foundation models, an increasing number of studies have explored controllable video generation by introducing additional modules (Chen et al., 2024b; Zhang et al., 2025b; Jiang et al., 2025) or designing specialized inference strategies (Zhang et al., 2025a; Chefer et al.; Wang et al., 2025). However, these approaches generally overlook the physical consistency of the generated videos (Kang et al., 2025). As a result, although they can produce temporally coherent sequences, the generated content may violate natural dynamics.

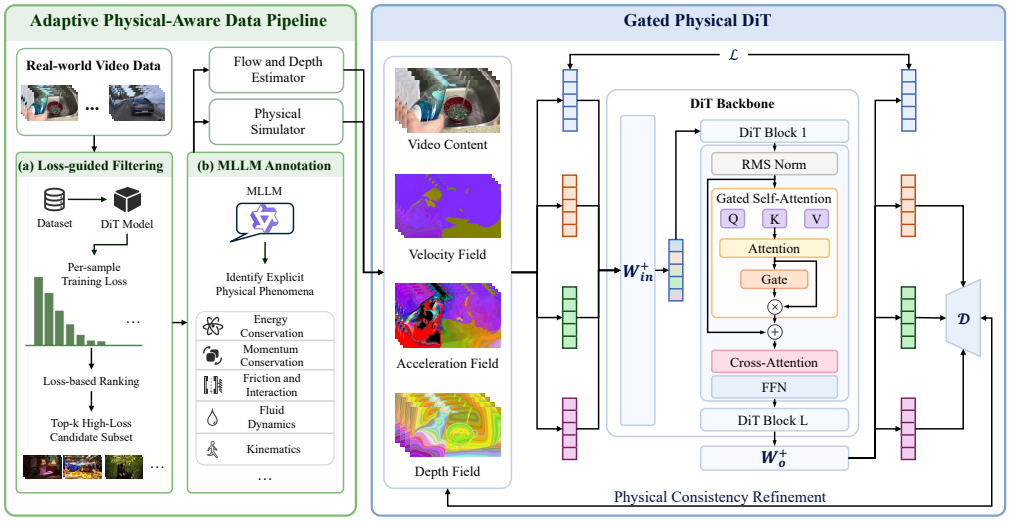


Figure 1: The framework of the proposed VideoPF method. We first employ an Adaptive Physical-Aware Data Pipeline to analyze the data, selectively filtering and generating physics-inspired samples based on the current weaknesses of the model. Subsequently, we jointly model visual and physical field representations by training a Gated Physical DiT. To further enhance the model's ability to model physical fields, we also propose a Physical Consistency Refinement method.

Physics in Video Generation Existing efforts toward improving physical consistency in video generation can be broadly categorized into two paradigms. The first line of work (Chefer et al.; Liu et al., b; Zhang et al.) focuses on enhancing dynamic realism by modifying model architectures or training objectives, encouraging the model to learn more plausible motion patterns through the inherent temporal correlations between consecutive frames. These approaches improve perceptual realism by leveraging natural frame-to-frame continuity. The second paradigm disentangles visual generation from physical understanding (Gosselin et al., 2025; Zhao et al., 2025; Li et al., 2025). It explicitly represents physical evolution using auxiliary signals, such as point clouds (Wang et al.), layout sequences (Yang et al., 2025; Liu et al., 2024), or fine-grained textual descriptions (Xue et al., 2025b; Lv et al., 2024), and reformulates the problem as controllable video generation. However, these approaches often rely on external physical representations, reasoning modules, or simulators, which introduce additional inference-time computation or system complexity. In contrast, we investigate joint modeling of visual and physical field representations within the generative backbone.

Overall, current video generation models have not fully explored joint modeling of visual representations with structured physical field representations within a unified video generator.

3 Methodology

In this section, we provide a comprehensive description of the VideoPF framework. We begin with a set of preliminaries, introducing the definitions and underlying motivations. To better exploit the rich information embedded in video data, we develop an adaptive physical-aware data pipeline. To enable joint modeling and effective integration of visual and physical signals, we further design a Gated Physical DiT architecture. Building upon this, we introduce a physical consistency refinement mechanism to more effectively provide field-level consistency supervision. The overall framework is illustrated in Figure 1.

3.1 Preliminary

Flow Matching Flow Matching (Lipman et al.; Liu et al., a) is a continuous-time generative modeling framework that learns a probability flow transporting a simple base distribution to a complex data distribution via neural ordinary differential equations.

Let $x_1 \in \mathbb{R}^d$ denote a data sample from the complex distribution $p_1(x)$ and x_0 denote a sampled noise from the base distribution $p_0(x)$, which is a standard Gaussian distribution $\mathcal{N}(0, I)$, of the same dimension. The generative process is defined via the ODE:

$$\frac{dx_t}{d_t} = u_\theta(x_t, t), t \in [0, 1], \tag{1}$$

where $u_\theta(x, t)$ is a neural velocity field. The density evolution under this flow follows the continuity equation:

$$\frac{\partial p_t(x)}{\partial t} + \nabla \cdot p_t(x) u_\theta(x, t) = 0. \tag{2}$$

Directly minimizing likelihood under the induced flow requires computing divergence terms, which is expensive. Flow Matching instead learns u_θ via supervised regression toward a target velocity constructed from a chosen probability path $x_t = \phi_t(x_0, x_1)$. For the simplest linear interpolation:

$$\phi_t(x_0, x_1) = (1 - t)x_0 + tx_1. \tag{3}$$

The corresponding conditional velocity is

$$u_t(x_0, x_1) = \frac{d}{dt}\phi_t(x_0, x_1) = x_1 - x_0. \tag{4}$$

The model is then optimized by:

$$\mathcal{L}_{FM} = \mathbb{E}_{t\sim\mathcal{U}[0,1], x_1\sim p_1, x_0\sim p_0}[||u_\theta(x_t, t) - u_t(x_0, x_1)||_2^2]. \tag{5}$$

Motivation and Task Definition Real-world dynamic phenomena can be represented through structured field-based descriptions without explicitly specifying the underlying physical laws. In this work, we focus on a set of physically relevant field representations that describe complementary aspects of scene dynamics. Specifically, we consider velocity, acceleration, and depth fields, which capture motion, temporal changes in motion, and spatial configuration, respectively. We collectively refer to these representations as physical fields.

Importantly, the physical fields used in our framework can be obtained from different sources. For real-world videos, velocity and acceleration are estimated from optical flow and its temporal derivatives, while depth is obtained from monocular depth estimation. These signals serve as observable field proxies and may contain estimation noise. For simulated videos, the corresponding fields are directly obtained from the physical simulator and therefore provide cleaner, simulation-derived field supervision. Our framework does not assume that the estimated fields from real videos recover the exact physical state of the underlying scene; instead, they provide structured spatiotemporal representations that complement visual appearance.

In practice, due to the lack of a mature 3D video variational autoencoder (VAE), we follow prior work (Chefer et al.) and project the 3D dynamical velocity and acceleration fields onto the image as $v(x, t), a(x, t) \in \mathbb{R}^{f\times h\times w\times 3}$, where $x \in \mathbb{R}^{f\times h\times w\times 3}$ is the video, f, h, w denote the frames, the height and the width of the video. However, this projection inevitably discards geometric information along the depth dimension. To mitigate this loss, we additionally introduce a depth field $d(x, t)$, which provides geometric information to partially recover spatial structure lost during projection. Overall, by jointly modeling visual appearance x, velocity field $v(x, t)$, acceleration $a(x, t)$, and depth $d(x, t)$, we encourage physically plausible dynamical evolution in the generative process while maintaining scalability to real-world video data.

3.2 Adaptive Physical-Aware Data Pipeline

To investigate the extent to which generative models capture physical dynamics in real-world videos, we begin with a joint analysis of the data distribution and the model–data interaction. The detailed analysis can be found in Appendix B.

We first employ a multimodal large language model (MLLM) to annotate a large-scale public video corpus, identifying whether each clip contains explicit manifestations of physical rules and categorizing the corresponding physical phenomena when present. Among approximately 38,000 annotated videos, only around 2,000 are identified as exhibiting explicit physical expressions. This observation suggests that videos with explicitly observable physical phenomena are inherently sparse in

natural video distributions, where semantically rich yet physically uninformative content dominates. Consequently, directly training on the full dataset risks overwhelming the model with semantic signals while under-exposing it to physically structured dynamics.

However, exhaustively annotating the entire corpus is inefficient under such sparsity. Instead, we aim to construct a more informative candidate subset prior to detailed labeling. Empirically, we observe that samples involving explicit physical dynamics tend to induce higher modeling difficulty for current generative models. Motivated by this observation, we analyze the relationship between per-sample training loss and the presence of explicit physical expressions. The results yield an AUC of approximately 0.73. Although loss cannot directly indicate whether the data carry specific physical information, this moderate yet consistent correlation suggests that it can serve as a practical prioritization signal for candidate subset selection under extreme class imbalance.

Using this loss-guided filtering strategy, we construct a high-likelihood subset and subsequently perform detailed MLLM-based annotation to determine the specific physical rules involved. For these samples, we construct estimated physical field representations by deriving velocity and acceleration from optical flow and obtaining depth from a monocular depth estimator. These estimated fields are observational proxies rather than exact physical states and inevitably contain estimation noise.

To provide cleaner and more controllable physical supervision, we further categorize the annotated samples by physical rule type and identify the categories where the current model exhibits the weakest performance. We then conduct targeted physical simulations using the Material Point Method (MPM) to generate physically consistent sequences under explicit physical constraints. From these simulations, we record ground-truth velocity, acceleration, and depth fields, which are used to reinforce the model in its most deficient physical regimes.

3.3 Gated Physical DiT

After the data preparation stage, we modify the model to jointly predict the distributions of visual information x_t, the physical velocity field v_t, the acceleration field a_t, and the depth field d_t. We achieve this by adapting the DiT model so that it can simultaneously take as input the noisy data quadruplet x_t, v_t, a_t, and d_t, and predict the corresponding velocity field of the distribution flow.

To maximally preserve the capability of the pretrained DiT backbone while incorporating physical information, we restrict our modifications to three components: the input projection layer, the attention mechanism, and the output projection layer.

We first extend the input projection matrix $W_{in} \in \mathbb{R}^{d\times d}$ into a higher-dimensional matrix $W_{in}^{+} \in \mathbb{R}^{4d\times d}$ via structured zero-initialized expansion, enabling it to process the concatenated input $[x_t, v_t, a_t, d_t]$:

$$h^1 = [x_t, v_t, a_t, d_t]W_{in}^{+}. \tag{6}$$

This design maps all modalities into a shared latent space of dimension d, ensuring compatibility with the original DiT architecture without increasing the hidden dimensionality.

For the l-th transformer layer, the model operates on a unified representation h^l. To better capture heterogeneous information across modalities, we introduce a gated self-attention mechanism:

$$Q^l = h^l W_q^l, \quad K^l = h^l W_k^l, \quad V^l = h^l W_v^l, \tag{7}$$

$$z^l = \text{softmax}\left(\frac{Q^l {K^l}^\top}{\sqrt{d}}\right) V^l, \tag{8}$$

$$g^l = \sigma(h^l W_g^l), \tag{9}$$

$$h^{l+1} = h^l + (z^l \odot g^l) W_o^l, \tag{10}$$

where σ denotes the sigmoid function and $\odot$ represents element-wise multiplication. The gating mechanism adaptively modulates the interaction among visual and physical representations, allowing the model to control the contribution of heterogeneous physical signals during joint modeling.

At the output stage, the shared representation is mapped to four representation-specific predictions corresponding to the visual representation, velocity field, acceleration field, and depth field. We

extend the output projection matrix to $W_o^+ \in \mathbb{R}^{d\times 4d}$, yielding

$$[u_\theta(x_t,t), u_\theta(v_t,t), u_\theta(a_t,t), u_\theta(d_t,t)] = h^L W_o^+. \tag{11}$$

Thus, we extend the model to a four-branch framework that processes and predicts all the fields. Finally, we decompose the training objective from Eq.5 into four components as

$$\mathcal{L}_m = \mathbb{E}_{t\sim\mathcal{U}[0,1],x_t,v_t,a_t,d_t}[||u_\theta(m_t,t) - u_t(m_0,m_1)||_2^2], m \in \{x,v,a,d\} \tag{12}$$

$$\mathcal{L} = \mathcal{L}_x + \sum_{m\in\{v,a,d\}} \lambda_m \mathcal{L}_m, \tag{13}$$

where $\mathcal{L}$ is the training objective and $\lambda_m, m \in \{v,a,d\}$ are the hyperparameters.

The key distinction from conventional auxiliary supervision is that the physical fields are not introduced as detached prediction heads or external conditions. Instead, the visual and physical representations share the same generative trajectory and latent representation, and their corresponding flow fields are predicted from the same unified hidden state. Thus, the physical objectives directly shape the shared generative representation used for visual synthesis.

3.4 Physical Consistency Refinement

Although joint modeling provides supervision for the physical field representations in latent space, decoding them through a pretrained video VAE can introduce additional reconstruction errors, compressing physics-inspired field representation through a pretrained video VAE inevitably introduces information loss, particularly for fine-grained dynamics such as flow, acceleration, and depth structure. To mitigate this issue, we introduce a post-hoc physical consistency refinement mechanism that leverages the VAE decoder to project latent predictions back into the physical observation space and uses reconstruction discrepancy to adaptively modulate the training objective.

For each sampled timestep $t \sim \{t_{min}, 1\}$, where $t_{min} > 0$ is a hyperparameter, we first predict the noise and recover the denoised latent via a scheduler step and decode it into the image domain using the pretrained VAE decoder

$$\widetilde{m} = \mathcal{D}(Step(\hat{u}_\theta(m_t,t),t,m_t)), \tag{14}$$

where $\mathcal{D}$ is the decoder of VAE. We then transform the decoded output into the corresponding physical field using a deterministic conversion function $\mathcal{T}$. The element-wise reconstruction error against simulated physical signals is computed as

$$e_m = ||m_{gt} - \mathcal{T}(\widetilde{m})||_2^2. \tag{15}$$

To avoid unstable physical supervision at early flow-matching stages, where the latent representations are dominated by noise, we introduce a timestep-dependent gating mechanism that gradually increases the contribution of physical supervision as the generation trajectory approaches the data distribution. Specifically, a scalar weight is defined as

$$\alpha(t) = \frac{t}{T}, \tag{16}$$

which increases linearly with the timestep. Then, the physical error map is downsampled via 3D average pooling to obtain a coarse but stable importance estimate, denoted as $\bar{e}_m$.

The training objective for each physical modality is formulated as a re-weighted loss

$$\mathcal{L}_{Refine} = \mathcal{L}_x + \sum_{m\in\{v,a,d\}} \mathbb{E}[\alpha(t)\bar{e}_m \odot \mathcal{L}_m], \tag{17}$$

where the error-derived weight adaptively emphasizes samples and regions where the decoded physical prediction deviates from ground-truth structure. This design effectively couples latent-space denoising with explicit physical-space consistency, encouraging the model to prioritize physically meaningful dynamics without requiring direct supervision in latent space.

Table 1: The quantitative results between our VideoPF method and the baselines. Bold results indicate the best performance, while underlined results denote the second-best performance.

Benchmark	Method	Motion	Appearance	Semantic	Physical
VideoJam-Bench	CogVideoX	0.870	0.712	<u>0.612</u>	0.237
	VideoREPA	<u>0.894</u>	0.740	<u>0.612</u>	<u>0.247</u>
	Wan2.1	0.839	<u>0.752</u>	0.608	0.244
	Ours	**0.937**	**0.764**	**0.629**	**0.262**
Movie-Gen	CogVideoX	<u>0.853</u>	0.717	0.420	<u>0.325</u>
	VideoREPA	0.836	0.737	0.414	0.321
	Wan2.1	0.818	<u>0.763</u>	<u>0.428</u>	0.323
	Ours	**0.864**	**0.770**	**0.454**	**0.356**
VideoPhy-Bench	CogVideoX	<u>0.859</u>	0.677	0.113	<u>0.215</u>
	VideoREPA	0.834	<u>0.723</u>	<u>0.115</u>	**0.220**
	Wan2.1	0.851	0.718	0.114	0.195
	Ours	**0.920**	**0.727**	**0.121**	<u>0.215</u>

4 EXPERIMENTS

We have conducted extensive experiments to show the performance of VideoPF method. We first give a brief introduction to the experimental setup, followed by the quantitative and qualitative results. Ablation studies further demonstrate the effectiveness of each proposed modules.

4.1 EXPERIMENT SETTINGS

Implementation Details We conduct experiments on the Wan2.1-T2V-1.3B model (Wan et al., 2025). All experiments are performed using NVIDIA-A100-40GB GPUs. We sample data from the publicly available OpenVid-HD (Nan et al.) dataset, and extract auxiliary information using FlowAnything (Liang et al., 2025) and DepthAnything (Yang et al., 2024). The model is trained on 10,000 real-world samples and 2,500 simulated samples. The resulting model is capable of generating videos at 15 fps with a spatial resolution of 832×480 and a temporal length of 81 frames.

Evaluation Metrics We compare our method with (i) CogVideoX-2B (Yang et al.), (ii) VideoREPA (Zhang et al.), and (iii) Wan2.1-T2V-1.3B. We evaluate the performance on three benchmarks: VideoJAM-Bench (Chefer et al.), Movie-Gen (Polyak et al., 2024), and VideoPhy-Bench (Bansal et al.). We adopt two types of metrics for evaluation. (i) We use VBench (Huang et al., 2024) to measure motion generation ability (Motion) and visual quality (Appearance). (ii) We use VideoPhy (Bansal et al.) to evaluate the semantic adherence (Semantic) and physical commonsense (Physical) of generation models. More detailed evaluation protocols are provided in Appendix H.2.

4.2 MAIN RESULTS

4.2.1 QUANTITATIVE EXPERIMENTS

Table 1 summarizes the quantitative results across three benchmarks. Overall, our method consistently achieves the best or highly competitive performance across all evaluation metrics, demonstrating consistent improvements across the evaluated benchmarks.

In terms of motion generation, our method significantly outperforms baselines on all of the three benchmarks. This substantial margin indicates its superior capability in modeling complex temporal dynamics. Meanwhile, for visual quality (Appearance metric), our method attains the best performance across all benchmarks, suggesting the improved frame-level realism and visual coherence.

Furthermore, our method consistently improves semantic adherence, surpassing prior approaches on all benchmarks, with particularly notable gains on Movie-Gen. This demonstrates enhanced consistency between generated content and input semantics. For physical commonsense, our method achieves the best results on VideoJAM-Bench and Movie-Gen, and remains competitive

on VideoPhy-Bench, where it achieves comparable performance to the strongest baseline. While Wan2.1 achieves the lowest physical score among the compared methods on VideoPhy-Bench, our method successfully improves this to 0.215, bringing it to a competitive level with other approaches. This demonstrates that the proposed design effectively compensates for the baseline's limitation in modeling physically plausible dynamics.

Overall, our method not only preserves the strong generative capability of the base model, but also significantly enhances its performance in motion modeling and dynamic consistency. More detailed data and baselines with larger parameters are provided in Appendix H.3, H.4, H.5, and H.7.



(a) A light ball is pressed into the water and then gradually rises to the surface.



(b) A lighter red toy car traveling from the left collides with a heavier blue toy car on the right.



(c) A close-up shot of a person cutting a carrot with a knife.



(d) A close-up and slow-motion shot of a runner moving through shallow puddles on a rain-wet surface. Each step produces detailed water splash dynamics.

Figure 2: Qualitative comparison between the proposed VideoPF method and the baselines.

4.2.2 Qualitative Experiments

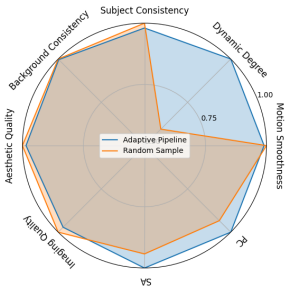
We present qualitative comparisons between our method and the baselines across scenarios involving the physical properties, momentum, and some physical commonsense understanding. The results can be found in Figure 2.

In Figure 2a, while CogVideoX partially captures the notion of buoyancy, it produces physically implausible results where the small ball even flies out of the water. Wan2.1 and VideoREPA keep the object dynamics largely unchanged. In contrast, our method correctly models the full process, where the ball is first pressed downward and then gradually rises back to the surface. In Figure 2b, we consider a collision between a lighter and a heavier object. Wan2.1 yields a degenerate outcome where both cars come to a complete stop after collision, while VideoREPA shows the blue car reversing direction even before collision. Our method instead produces more physically plausible

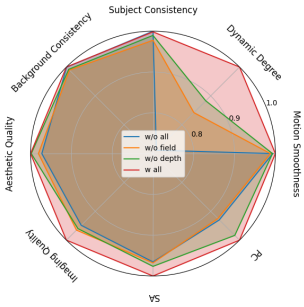
behavior, where the lighter car is pushed backward after the collision, aligning more closely with realistic momentum exchange. For scenarios involving causal commonsense reasoning, as shown in Figure 2c and 2d, only our method preserves motion consistency: the number of carrot slices remains consistent after cutting, and water splashes occur only after the runner's foot makes contact with the surface.

Overall, compared to baseline methods, our method better captures video content by explicitly and jointly modeling physical fields information, resulting in outputs with improved motion consistency and smooth dynamics. Additional generated videos and prompts are provided in Appendix H.9 to better demonstrate the generation performance.

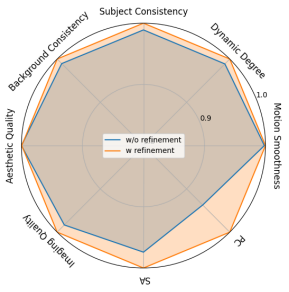
4.3 Ablations



(a) Comparison of whether to adopt the Adaptive Physical-Aware Data Pipeline. "Adaptive Pipeline" means utilizing the proposed module, while "Random Sample" means randomly sampling the same number of videos from the origin dataset.



(b) Comparison of with and without various types of auxiliary information. "w/o all" means training without any auxiliary fields. "w/o field" means training without the velocity and acceleration fields. "w/o depth" means training without the depth field.



(c) Comparison of whether to utilize the Physical Consistency Refinement. "w/o refinement" indicates that the physical consistency refinement method is not applied, and the model is trained for the same number of iterations using fixed auxiliary weights.

Figure 3: Ablation study. We ablate the proposed modules. The best-performing result is normalized to 1, and all other results are scaled accordingly.

We present a comprehensive ablation study evaluating the effectiveness of the proposed components from three perspectives. The detailed values are provided in appendix H.8. Comparing the Adaptive Physical-Aware Data Pipeline with random sampling shows in Figure 3a consistent improvements across all evaluation dimensions, particularly in motion smoothness and physical commonsense, indicating that structured, physically-informed data construction is crucial for learning robust dynamics. Secondly, the analysis of auxiliary information (Figure 3b) demonstrates that incorporating richer physical cues (e.g., velocity, acceleration, and depth) leads to progressively better performance. Removing any subset of these signals degrades performance, while training without all auxiliary fields results in the most significant drop, highlighting the importance of multi-field physical supervision. Figure 3c evaluates the impact of Physical Consistency Refinement. Applying refinement yields noticeable gains, especially in physical commonsense and semantic adherence, whereas training with fixed auxiliary weights (w/o refinement) leads to suboptimal results.

5 Conclusion

In this paper, we present VideoPF, a physical-field-based framework for video generation that improves dynamic plausibility through joint modeling of physics-inspired auxiliary representations. Rather than explicitly modeling physical laws, VideoPF incorporates structured physical field representations into the generative process and jointly models them with visual content. Experiments demonstrate that this formulation provides complementary supervision for improving temporal dynamics and motion quality, while also benefiting physics-oriented evaluation.

AI USE STATEMENT

In this work, we used generative AI tools for [generate synthetic data sets, assist with translation, clean and reformat dataset, design or provide feedback on research methodology or experiments]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, support qualitative and thematic data analysis, interpret results, implement methods] and [assist in the writing of proofs, propose or refine hypotheses,] are not applicable to this work. Additionally, we used generative AI tools for [generating some of the icons, assisting with paper polishing, and aiding in writing parts of the code related to the physics simulator]. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the implementation and training details, including model configuration, optimization settings, datasets, and baselines, in Section 4.1 and Appendix H.1. The physical-aware data construction pipeline, including the MLLM annotation protocol and sample selection procedure, is described in Appendix B and Section 3.2. Details of physical-field transformation, simulation data generation, evaluation protocols, and additional experimental results are provided in Appendices C, D, and H.3to H.10. Computational resources and inference-time analysis are reported in Appendices H.1 and H.6.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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. In *The Thirteenth International Conference on Learning Representations*.

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.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. openai, 2024.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. In *Forty-second International Conference on Machine Learning*.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7310–7320, 2024a.

Hong Chen, Xin Wang, Yipeng Zhang, Yuwei Zhou, Zeyang Zhang, Siao Tang, and Wenwu Zhu. Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3637–3646, 2024b.

Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, Yaofei Wu, and Wenwu Zhu. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning on language-video foundation models. *IEEE Transactions on Multimedia*, 2025.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Anthony Gosselin, Ge Ya Luo, Luis Lara, Florian Golemo, Derek Nowrouzezahrai, Liam Paull, Alexia Jolicoeur-Martineau, and Christopher Pal. Ctrl-crash: Controllable diffusion for realistic car crashes. *arXiv preprint arXiv:2506.00227*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pp. 393–411. Springer, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yuanming Hu, Yu Fang, Ziheng Ge, Ziyin Qu, Yixin Zhu, Andre Pradhana, and Chenfanfu Jiang. A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

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

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pp. 143–150, 1986.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *International Conference on Machine Learning*, pp. 28991–29017. PMLR, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

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

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In *International Conference on Machine Learning*, pp. 35685–35709. PMLR, 2025.

Yingping Liang, Ying Fu, Yutao Hu, Wenqi Shao, Jiaming Liu, and Debing Zhang. Flow-anything: Learning real-world optical flow estimation from large-scale single-view images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Minghui Lin, Xiang Wang, Yishan Wang, Shu Wang, Fengqi Dai, Pengxiang Ding, Cunxiang Wang, Zhengrong Zuo, Nong Sang, Siteng Huang, et al. Exploring the evolution of physics cognition in video generation: A survey. *arXiv preprint arXiv:2503.21765*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, a.

Zhiheng Liu, Xueqing Deng, Shoufa Chen, Angtian Wang, Qiushan Guo, Mingfei Han, Zeyue Xue, Mengzhao Chen, Ping Luo, and Linjie Yang. Worldweaver: Generating long-horizon video worlds via rich perception. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, b.

William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pp. 347–353. 1998.

Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1430–1440, 2024.

Miles Macklin. Warp: A high-performance python framework for gpu simulation and graphics. `https://github.com/NVIDIA/warp`, March 2022. URL `https://github.com/NVIDIA/warp`. NVIDIA GPU Technology Conference (GTC).

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *International Conference on Machine Learning*, pp. 43781–43806. PMLR, 2025.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *The Thirteenth International Conference on Learning Representations*.

Zirui Pan, Xin Wang, Yipeng Zhang, Hong Chen, Kwan Man Cheng, Yaofei Wu, and Wenwu Zhu. Modular-cam: Modular dynamic camera-view video generation with llm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6363–6371, 2025.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3626–3636, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013.

Rui Sun, Yumin Zhang, Tejal Shah, Jiahao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, Bo Wei, and Rajiv Ranjan. From sora what we can see: A survey of text-to-video generation. *arXiv preprint arXiv:2405.10674*, 2024.

Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report, 2025. URL https://arxiv.org/abs/2511.18870.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016.

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

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. Physctrl: Generative physics for controllable and physics-grounded video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Luozhou Wang, Ziyang Mai, Guibao Shen, Yixun Liang, Xin Tao, Pengfei Wan, Di Zhang, Yijun Li, and Ying-Cong Chen. Motion inversion for video customization. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4389–4398, 2024.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024.

Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, Zhensong Zhang, Minglei Li, Jian Yang, et al. Human motion video generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

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

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10371–10381, 2024.

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, et al. Vlipp: Towards physically plausible video generation with vision and language informed physical prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12360–12370, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*.

Shiyi Zhang, Junhao Zhuang, Zhaoyang Zhang, Ying Shan, and Yansong Tang. Flexiact: Towards flexible action control in heterogeneous scenarios. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025a.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yipeng Zhang, Xin Wang, Hong Chen, Chenyang Qin, Yibo Hao, Hong Mei, and Wenwu Zhu. Scenariodiff: text-to-video generation with dynamic transformations of scene conditions. *International Journal of Computer Vision*, 133(7):4909–4922, 2025b.

Qi Zhao, Xingyu Ni, Ziyu Wang, Feng Cheng, Ziyan Yang, Lu Jiang, and Bohan Wang. Synthetic video enhances physical fidelity in video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12135–12146, 2025.

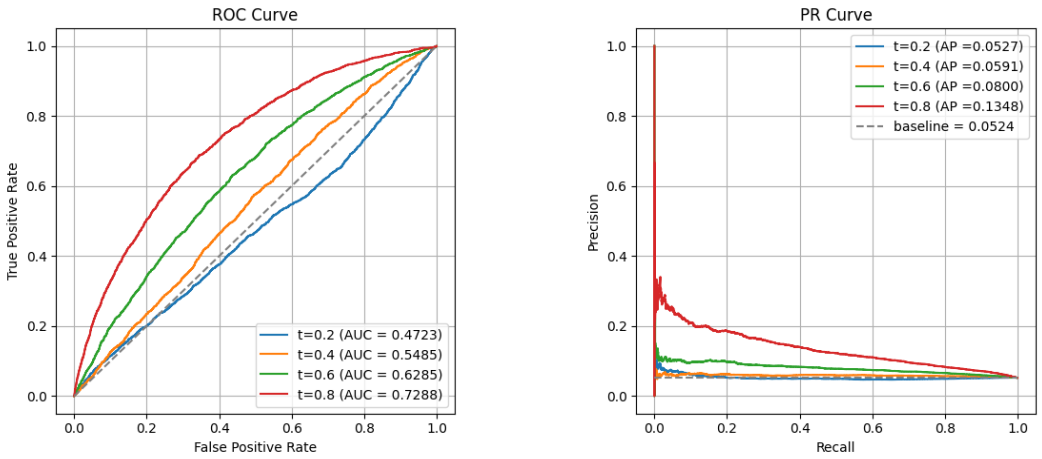
Zeshun Zong, Xuan Li, Minchen Li, Maurizio M Chiaramonte, Wojciech Matusik, Eitan Grinspun, Kevin Carlberg, Chenfanfu Jiang, and Peter Yichen Chen. Neural stress fields for reduced-order elastoplasticity and fracture. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

A Appendix

B Analysis of Dataset and Model

To investigate the base model's capability in understanding and generating data involving physical phenomena, we first conduct an analysis on a publicly available video dataset. Specifically, we sample 40,000 video instances and employ Qwen3-VL-8B (Bai et al., 2025) to annotate each video, examining whether it contains explicit physical phenomena or adheres to identifiable physical laws. The prompt template used for annotation is provided as follows:

> You are a Video Physics Reasoning Assistant. Your task is to analyze a given video and identify the physical laws, motion patterns, interactions, and causal relationships that appear in the scene.
>
> Your output MUST strictly follow the JSON format below. Do NOT include explanations, comments, or any additional text outside the JSON.
>
> You should recognize physical rules including but not limited to:
>
> - Kinematics: constant velocity, acceleration, circular motion, free fall, projectile motion, oscillation.
> - Dynamics: Newton's laws, friction, normal force, gravity, tension, contact forces, stability.
> - Energy: energy conservation, potential–kinetic conversion, elastic potential energy.
> - Collisions: elastic collision, inelastic collision, momentum conservation, recoil.
> - Fluid-related rules: drag, buoyancy, pressure changes.
> - Causality: event A triggers event B, force-induced changes, interaction chains.
>
> Return empty lists for categories where you detect no instances.
>
> Output JSON schema:
>
> ```
> {
>     "physical_rules": [ ... ],
>     "objects": [
>         {
>             "id": "string",
>             "type": "string",
>             "motion_pattern": [ ... ],
>             "interactions": [ ... ]
>         }
>     ],
> ```



(a) ROC Curve (b) PR Curve

Figure 4: Performance comparison across different timesteps.

}
Analyze the following video and extract all physical rules or motion patterns present in the scene. Return ONLY the structured JSON result defined in the system instructions.

Finally, we obtain 37,284 valid annotations, while the remaining 2,716 samples fail due to incomplete or non-compliant responses. Among the valid set, only 1,954 samples are annotated as containing explicit physical information, accounting for approximately 5.24% of the total. This indicates that videos exhibiting clear physical rules or observable physical phenomena are highly sparse in the full dataset.

In principle, all real-world videos should obey specific physical laws. However, our empirical findings reveal a striking sparsity, suggesting a key question: whether physical laws are explicitly expressed is fundamentally different from whether they implicitly govern the underlying dynamics.

Real-world scenes indeed follow physical principles such as Newton's laws and energy conservation. However, most video segments exhibit (i) no evident dynamical evolution, and (ii) no explicit state transitions. For instance, videos of a person speaking at a podium or a dog walking are physically valid, yet they do not involve salient physical interactions or conservation patterns that are directly observable or learnable.

In other words, although physical constraints are always present, their observability in video data is extremely limited. As a result, learning in generative models is predominantly driven by semantic structure rather than explicit physical signals or causal reasonings.

Therefore, it is necessary to reorganize the dataset so that it covers as many samples with explicit physical representations as possible. However, due to the inherent sparsity of such samples, it is difficult to annotate them exhaustively in a fine-grained manner. Considering that generative models are trained on the original data distribution, they may not be able to perfectly model data with explicit physical expressions, and their generation quality on such samples is often sub-optimal.

Motivated by this, we aim to leverage certain proxy metrics to perform a coarse-grained separation of the data, first identifying samples that are more likely to contain explicit physical principles, and then refining the selection within this subset.

We adopt the loss score as the primary indicator and conduct a preliminary analysis. Specifically, we perform experiments on the 37,284 annotated samples, analyzing the relationship between the loss value and the presence of explicit physical representations at timesteps $t = 0.2, t = 0.4, t = 0.6$, and $t = 0.8$. The corresponding ROC curves and PR curves are shown in Figure 4:

The results indicate that under an extremely sparse setting, using the loss as an indicator still yields an AUC of approximately 0.7288, as well as an average precision of 0.1348, nearly three times the

positive sample ratio. This suggests a certain degree of correlation between the loss and the presence of explicitly expressed physical structure in the data.

We further observe that this correlation is almost negligible at small timestep t, but gradually strengthens as t increases. On the one hand, when t is small, the data remains in a highly noisy regime regardless of whether explicit physical rules are present, and the model tends to focus more on coarse visual structure rather than dynamic laws or fine-grained temporal variations. On the other hand, the progressively increasing correlation supports the hypothesis that the model's ability to capture and represent physical consistency information – or more precisely, the subset of physical rules that the model struggles to properly model yet – is not uniformly distributed across the generation process, but instead is primarily concentrated in the later stages.

Regarding the PR curve, we observe a sharp peak when recall approaches zero, implying the existence of an "ultra-high-quality physical subset" that can be identified by the loss, albeit in very limited quantity and quickly overwhelmed by noise.

Therefore, in our method, we use the loss as a criterion in combination with annotated physical rules to identify the ten most challenging physical regularities for the current model, ranked as follows:

- Energy conservation
- Momentum conservation
- Friction and interaction
- Fluid dynamics
- Buoyancy
- Air resistance
- Heat transfer
- Kinematics
- Projectile motion
- Circular motion

C PHYSICS SIMULATOR

After completing the data analysis, and given that accurate physical field signals cannot be directly obtained from real-world data, we develop a physics simulator using NVIDIA Warp (Macklin, 2022; Xie et al., 2024; Zong et al., 2023) to generate video data annotated with ground-truth physical field signals.

For the physical simulation module, we adopt the Material Point Method (MPM) (Stomakhin et al., 2013; Hu et al., 2018) for simulation. The detailed descriptions are provided in Appendix D. After the simulation stage, in order to obtain results that are closer to real video data, we first convert the particles into meshes using the Marching Cubes algorithm (Lorensen & Cline, 1998). We then render the simulated results using a path tracing-based rendering pipeline (Kajiya, 1986). This procedure yields results as illustrated in Figure 5.

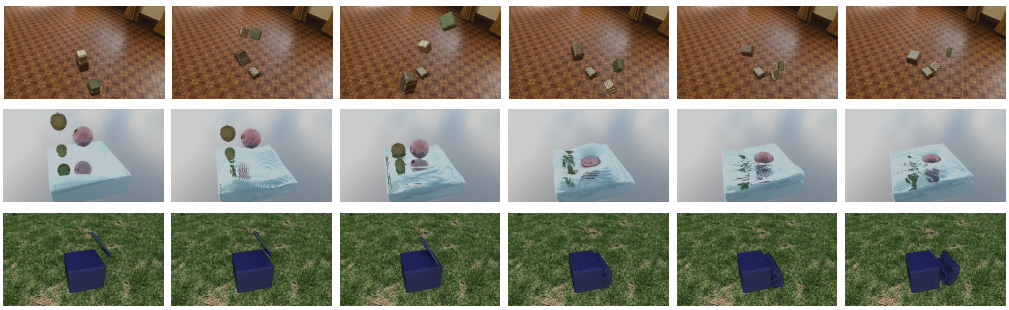


Figure 5: Samples of simulated videos.

In consideration of the scenarios discussed in Appendix B, we design a set of representative simulation environments corresponding to several relatively challenging physical rules:

- Objects with different materials, masses, and shapes are thrown onto sand or water surfaces, which is designed to enhance the model's understanding of gravity and energy dynamics.
- Stacking and collision of heterogeneous objects, which is designed to improve the model's comprehension of friction and momentum conservation.
- Mixing processes between liquids with different properties, as well as interactions between solid and fluid phases, which is designed to strengthen the model's understanding of fluid dynamics and buoyancy.
- Deformation and cutting of objects with different material properties, which is designed to enhance the model's understanding of intrinsic material characteristics.

We employ Qwen3-VL-8B (Bai et al., 2025) to generate captions for the simulated videos and add the necessary physical parameters such as "density" to it. The prompt used is designed as follows:

> This video is a physically simulated result. Here are some of the parameters of the scene:
>
> [
>
> {
>
> "object name": ¡name¿,
>
> "materials": {
>
> "material type": ¡type¿,
>
> "Young's Modulus": ¡value¿,
>
> "Poisson's Ratio": ¡value¿,
>
> "Yield Stress": ¡value¿,
>
> ...
>
> },
>
> "Initialized velocity": ¡value¿,
>
> "Position": ¡value¿,
>
> "Size of Bounding Box": ¡value¿,
>
> },
>
> ...
>
>]
>
> Please caption this video.

D MATERIAL POINT METHOD

The Material Point Method (Stomakhin et al., 2013; Hu et al., 2018) is a hybrid Lagrangian-Eulerian numerical scheme for simulating continuum materials undergoing large deformation. It combines the advantages of particle-based and grid-based representations by evolving Lagrangian material points through an Eulerian computational grid.

D.1 CONTINUUM DYNAMICS

Let the material domain be discretized into a set of material points or particles indexed by $p \in \mathcal{P}$, each with attributes of mass m_p, position $\boldsymbol{x}_p \in \mathbb{R}^3$, velocity $\boldsymbol{v}_p$, and deformation gradient $\boldsymbol{F}_p \in \mathbb{R}^{3\times 3}$. The governing continuum equation is the conservation of linear momentum:

$$\rho(\boldsymbol{x})\frac{d\boldsymbol{v}}{dt} = \nabla\boldsymbol{\sigma}(\boldsymbol{x}) + \rho(\boldsymbol{x})\boldsymbol{b}(\boldsymbol{x}), \tag{18}$$

where ρ is density, $\boldsymbol{\sigma}$ is the stress tensor, and $\boldsymbol{b}$ denotes body forces (e.g., gravity).

D.1.1 PARTICLE-TO-GRID (P2G) TRANSFER

MPM introduces a background computational grid with nodes indexed by $i \in \mathcal{G}$. Let $N_i(x)$ denote the interpolation function associated with grid node i.

Mass and momentum are transferred from particles to grid nodes as:

$$m_i = \sum_p m_p N_i(\boldsymbol{x}_p), \tag{19}$$

$$(m\boldsymbol{v})_i = \sum_p m_p \boldsymbol{v}_p N_i(\boldsymbol{x}_p). \tag{20}$$

Thus, grid velocities are computed as

$$\boldsymbol{v}_i = \frac{(m\boldsymbol{v})_i}{m_i}. \tag{21}$$

The internal force at grid nodes is given by

$$\boldsymbol{f}_i^{int} = -\sum_p \boldsymbol{V}_p \boldsymbol{\tau} \boldsymbol{F}^{-\top} \nabla N_i(\boldsymbol{x}_p), \tag{22}$$

where $\boldsymbol{V}_p$ is the particle volume and $\boldsymbol{\tau}\boldsymbol{F}^{-\top}$ is the first Piola-Kirchhoff stress.

External forces like gravity are:

$$\boldsymbol{f}_i^{ext} = \sum_p m_p \boldsymbol{b}_p N_i(\boldsymbol{x}_p). \tag{23}$$

D.2 Grid Update

On the grid, momentum is advanced using explicit time integration:

$$m_i \frac{d\boldsymbol{v}_i}{dt} = \boldsymbol{f}_i^{int} + \boldsymbol{f}_i^{ext}. \tag{24}$$

A typical explicit Euler update yields

$$\boldsymbol{v}_i^{t+1} = \boldsymbol{v}_i^t + \Delta t \frac{\boldsymbol{f}_i^{int} + \boldsymbol{f}_i^{ext}}{m_i}. \tag{25}$$

Grid positions are updated as

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t + \Delta t \boldsymbol{v}_i^{t+1}. \tag{26}$$

D.3 Grid-to-Particle (G2P) Transfer

After grid updates, velocities and positions are interpolated back to particles:

$$\boldsymbol{v}_p^{t+1} = \sum_i \boldsymbol{v}_i^{t+1} N_i(\boldsymbol{x}_p), \tag{27}$$

$$\boldsymbol{x}_p^{t+1} = \boldsymbol{x}_p^t + \Delta t \boldsymbol{v}_p^{t+1}. \tag{28}$$

D.4 Constitutive Model for Kirchhoff Stress

In the MPM, the stress response of each material point is determined by a constitutive model that maps the deformation gradient $\boldsymbol{F}$ to the Kirchhoff stress $\boldsymbol{\tau}$. Recall that

$$\boldsymbol{\tau} = J\boldsymbol{\sigma}, J = \det(\boldsymbol{F}). \tag{29}$$

In practice, it is convenient to express constitutive laws directly in terms of $\boldsymbol{\tau}$, which is work-conjugate to the velocity gradient.

Let the singular value decomposition of the deformation gradient be

$$\boldsymbol{F} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \sigma_3), \tag{30}$$

with principal stretches σ_i. We now introduce several constitutive models used in this work to simulate materials includes jelly, metal, snow, foam, plasticine, sand, and fluid.

E GRADIENTS ANALYSIS

From an intuitive perspective, our method can be interpreted as explicitly modeling the relationship between physical fields and visual content, thereby achieving improved generation quality. From a system-level standpoint, the overall framework can also be viewed as a form of multi-task learning. Motivated by this, we analyze the gradients of different tasks with respect to the shared parameters in the final layer, aiming to better understand the underlying mechanism behind the observed empirical gains.

Let $h \in \mathbb{R}^{1536}$ denote the hidden representation in the final layer of our codebase. In such a high-dimensional space, it is well known that almost all the randomly sampled normalized vectors are highly likely to be orthogonal due to concentration of measure phenomena in high-dimensional geometry.

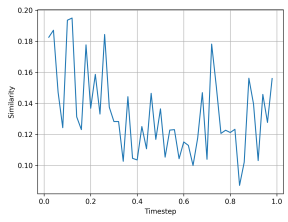
To address the issue, we first compute the gradient of each task-specific loss $\mathcal{L}_m, m \in \{x, v, a, d\}$ with respect to the final-layer hidden representation, i.e.,

$$g_m = \nabla_h \mathcal{L}_m. \tag{31}$$

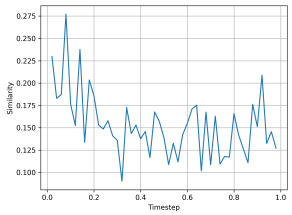
We then perform singular value decomposition for dimensionality reduction:

$$\hat{g}_t \approx U_k \Sigma_k V_k^\top, k = 64, \tag{32}$$

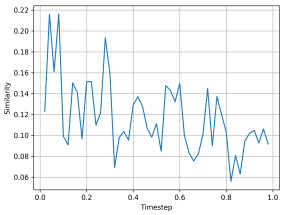
and project the gradients into a 64-dimensional subspace spanned by the leading singular components. This step preserves the dominant directions of task-specific optimization dynamics while filtering out high-dimensional noise. We then compute pairwise cosine similarities between the projected gradients of different tasks, as illustrated in Figure 6.



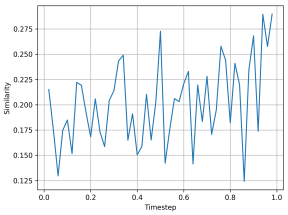
(a) Similarity between $\mathcal{L}_x$ and $\mathcal{L}_v$



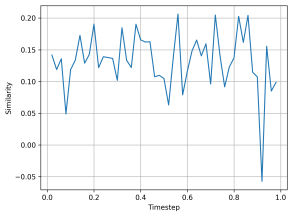
(b) Similarity between $\mathcal{L}_x$ and $\mathcal{L}_a$



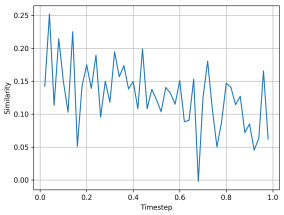
(c) Similarity between $\mathcal{L}_x$ and $\mathcal{L}_d$



(d) Similarity between $\mathcal{L}_v$ and $\mathcal{L}_a$



(e) Similarity between $\mathcal{L}_v$ and $\mathcal{L}_d$



(f) Similarity between $\mathcal{L}_a$ and $\mathcal{L}_d$

Figure 6: Cosine similarities between each $\mathcal{L}_m$ across different timesteps.

Empirically, we observe that in nearly all cases, the cosine similarities between gradients of the four tasks are consistently greater than zero. It indicates that a systematic positive alignment between task gradients. This suggests that these tasks share a common optimization direction, and thus can mutually reinforce each other during joint training. In particular, such positive correlation implies that multi-task learning in our setting does not introduce severe gradient conflict, but instead facilitates cooperative representation learning, enabling the model to better capture physically consistent structures.

However, this does not imply that auxiliary supervision can be arbitrarily scaled. When gradient similarity is positive, auxiliary tasks effectively act as a regularization constraint on the primary

objective, i.e., they bias the solution toward representations that are simultaneously predictive of multiple physical quantities. While this can improve generalization, it also introduces an implicit trade-off: the model must allocate representational capacity to capture auxiliary signals that are not perfectly aligned with the main task objective. Consequently, when both data scale and model capacity are sufficiently large, the benefit of such regularization may diminish.

Nevertheless, as observed in Appendix B, physical-consistency-aware video generation suffers from a severe scarcity of physically grounded supervision signals. In this regime, our joint-modeling formulation provides a particularly effective inductive bias, leveraging shared physical structure across tasks to compensate for limited explicit physical annotations, thereby offering a practical and efficient solution under data-constrained settings.

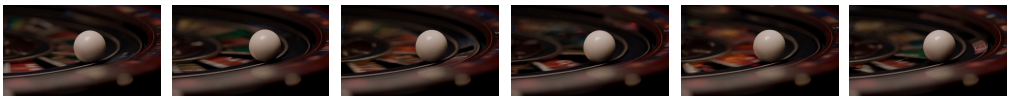
F Transformation Approach

In this work, we discretize physical field variables and map them into the RGB space. During the refinement stage, the decoded RGB representations are reconstructed back into the physical domain. Conventional mappings used in prior approaches are typically non-bijective, which leads to information loss and insufficient information density during refinement.

To address this issue, we adopt a transformation approach based on the Hilbert Curve to transform between the physical space and the RGB space. Owing to its strong locality-preserving property, this mapping better aligns with the structured distribution of meaningful RGB images.

G Discussion and Limitation

Due to resource constraints, the base models employed in this work are relatively small in scale, and thus inevitably inherit their limitations. In particular, the generated results remain suboptimal when handling extremely fast or highly complex motions and interactions. As shown in Figure 7a, the model neglects the interaction between the roulette wheel and the ball, resulting in the ball exhibiting only self-rotation rather than orbital motion. Meanwhile, the loss of fine-grained details due to high-speed motion also prevents our method from accurately capturing the underlying physical relationships. Moreover, the method struggles with scenarios that require extensive reasoning over intricate physical transmission systems, such as Rube Goldberg machines in Figure 7b. Finally, our approach models physical fields that are more closely aligned with macroscopic dynamics. Therefore, in scenarios involving multiple objects, their physical properties can only be evaluated in a relative manner, rather than through absolute comparisons. Besides, our approach does not explicitly enforce physical laws or conservation constraints and the velocity and acceleration fields used in this work are observation-level estimates derived from visual motion, rather than physically grounded state variables. Our goal is therefore to improve perceptual physical plausibility rather than guarantee exact physical correctness. Modeling physically grounded state variables and explicit physical constraints remains an important direction for future work.



(a) A roulette wheel in a dimly lit room or casino floor. In the center of the wheel, there's a small white ball that appears to be spinning rapidly as it moves around the track. The ball spins around the wheel, and the wheel rotates counterclockwise.



(b) A polished steel ball is released from the top of a winding track and rolls smoothly downward under gravity, accelerating as it follows curved rails and gentle ramps. The ball strikes the first carefully aligned domino, triggering a precise chain reaction as a series of dominoes topple sequentially in rhythmic motion across the structure and finally strikes a brass bell.

Figure 7: Failure cases.

From these perspectives, a promising future research direction may lie not so much in improving generative modeling itself, but rather in exploring how models can better understand and faithfully follow the physical information embedded in textual instructions. For instance, one potential avenue is the design and training of a genuinely physical-aware variational autoencoder, which could map physical fields into a latent space that is more amenable to machine understanding.

In addition, the evaluation of generated videos remains a significant challenge. How to reliably extract object-level quantities such as velocity and motion dynamics from pixel-level video frames, and thereby construct deterministic metrics instead of potentially biased scores derived from post-trained large language models, may constitute one of the most critical problems for future research. A better evaluation and understanding of videos can also help mitigate potential privacy issues arising from the misuse of video generation models.

H Detailed Scores and More Results

H.1 Experiment Settings

Implementation Details Our base model is Wan2.1-T2V-1.3B, a widely used text-to-video generation framework. All experiments are conducted on four NVIDIA-A100 GPUs with 40GB memory each. We employ the Muon optimizer (Liu et al., 2025) with a learning rate of 1e-5 and a weight decay of 1e-2. The model is trained on a dataset comprising 10,000 real samples and 2,500 simulated samples, with refinement applied exclusively to the simulated samples. During training and evaluation, we fix the generated video length to 81 frames at 15 fps, with a spatial resolution of 832×480.

Dataset OpenViD-HD (Nan et al.) is a high-quality, open-source dataset designed for text-to-video generation. It aims to provide high-resolution and semantically rich video data to facilitate the training and evaluation of advanced video generation models. The dataset is typically curated through multiple quality control procedures, such as aesthetic scoring, motion filtering, and deduplication, ensuring a relatively high level of data quality. The dataset is publicly available at the official repository.

Baselines We compare our method with the following baselines, which represent the top-tier performance among open-source models under comparable parameter budgets.

- Wan2.1-T2V-1.3B:https://huggingface.co/Wan-AI/Wan2.1-T2V-1.3B
- CogVideoX-2B:https://huggingface.co/zai-org/CogVideoX-2b
- VideoREPA: https://huggingface.co/aHapBean/VideoREPA/tree/main

Benchmarks To better compare our method with the baselines in terms of motion dynamics, appearance quality, and physical consistency, we evaluate all models on three complementary benchmarks.

- VideoJAM-Bench: VideoJAM-Bench is primarily designed to evaluate the dynamic performance of video generation models, as well as their basic ability to preserve physical consistency. It covers a diverse set of dynamic scenarios, including complex human actions and physically driven motions. The benchmark comprises 128 carefully curated prompts for comprehensive evaluation and is available at the Appendix F of https://arxiv.org/pdf/2502.02492.
- MovieGen: MovieGen is designed to evaluate the general generative capability of models in dynamic scenarios, with a particular emphasis on generation quality, diversity, and generalization ability. The benchmark comprises 1,003 prompts covering a wide range of scenes, which are used to systematically assess the overall performance of video generation models across diverse contexts. The benchmark is available at https://github.com/facebookresearch/MovieGenBench.
- VideoPhy-Bench: VideoPhy-Bench is designed to evaluate models' ability to understand and adhere to physical commonsense in video generation. The benchmark curates a set of

scenarios that require explicit physical reasoning. For computational efficiency, we sample 600 prompts. The benchmark is available at https://github.com/Hritikbansal/videophy.

H.2 Evaluation Metrics

For the metrics calculated by VBench, we evaluate six aspects of them: Motion Smoothness, Dynamic Degree, Subject Consistency, Background Consistency, Aesthetic Quality, and Imaging Quality. The first two metrics are motion-related, while the remaining four assess appearance quality. In the main text, we aggregate these metrics following the official protocol provided by VBench, as detailed below:

$$\text{Motion} = \frac{1}{1+0.5}[(\text{Motion Smoothness} - 0.706)/(0.9975 - 0.706) + 0.5 \cdot \text{Dynamic Degree}]$$

$$\begin{aligned}\text{Appearance} = \frac{1}{4}[&(\text{Subject Consistency - } 0.1462)/(1.0 - 0.1462) + \text{Aesthetic Quality} \\ &+ (\text{Background Consistency - } 0.2615)/(1.0 - 0.2615) + \text{Imaging Quality}]\end{aligned}$$

H.3 Detailed Scores of Quantitative Results

In this section, we provide more detailed results compared to Table 1. For the three different benchmarks, the results are reported in Table 2, 3, and 4. Here, MS denotes Motion Smoothness, DD denotes Dynamic Degree, SC denotes Subject Consistency, BC denotes Background Consistency, AQ denotes Aesthetic Quality, IQ denotes Imaging Quality, SA denotes Semantic Adherence, and PC denotes Physical Commonsense. The term agg. refers to the weighted aggregation described in H.2

Table 2: Quantitative results between VideoPF method and other baselines on the VideoJAM-Bench.

Method	Motion			Appearance					Physical	
	MS	DD	agg.	SC	BC	AQ	IQ	agg.	SA	PC
CogVideo	0.961	0.859	0.870	0.908	0.937	0.504	0.539	0.712	0.612	0.237
REPA	0.978	0.813	0.894	0.923	0.945	0.532	0.587	0.740	0.612	0.247
Wan2.1	0.983	0.617	0.839	**0.937**	0.946	0.543	0.613	0.752	0.608	0.244
Ours	**0.989**	**0.867**	**0.937**	0.936	**0.950**	**0.558**	**0.646**	**0.764**	**0.629**	**0.262**

Table 3: Quantitative results between VideoPF method and other baselines on the Movie-Gen.

Method	Motion			Appearance					Physical	
	MS	DD	agg.	SC	BC	AQ	IQ	agg.	SA	PC
CogVideo	0.974	**0.719**	0.853	0.911	0.934	0.528	0.533	0.717	0.420	0.325
REPA	0.983	0.606	0.836	0.911	0.940	0.546	0.588	0.737	0.414	0.321
Wan2.1	**0.990**	0.505	0.818	**0.944**	**0.951**	0.565	0.618	0.763	0.428	0.323
Ours	0.984	0.686	**0.864**	0.932	0.940	**0.582**	**0.659**	**0.770**	**0.454**	**0.356**

Table 4: Quantitative results between VideoPF method and other baselines on the VideoPhy-Bench.

Method	Motion			Appearance					Physical	
	MS	DD	agg.	SC	BC	AQ	IQ	agg.	SA	PC
CogVideo	0.977	0.717	0.859	0.893	0.944	0.456	0.455	0.677	0.113	0.215
REPA	0.987	0.576	0.834	**0.921**	**0.948**	0.489	0.567	0.723	0.115	**0.220**
Wan2.1	**0.989**	0.614	0.851	0.912	0.936	0.489	0.571	0.718	0.114	0.195
Ours	0.982	**0.867**	**0.920**	0.900	0.929	**0.508**	**0.614**	**0.727**	**0.121**	0.215

H.4 COMPARISON WITH HUNYUANVIDEO-1.5

In this section, we compare our method with HunyuanVideo-1.5-8.3B (Team, 2025) on the three aforementioned benchmarks, as shown in Table 5, 6, and 7.

Due to the substantial gap in model scale, our method underperforms HunyuanVideo in terms of motion dynamics and visual quality of the generated videos. Nevertheless, it is noteworthy that across all benchmarks, our approach achieves performance on semantic adherence (SA) and physical commonsense (PC) – two metrics closely related to physical plausibility – that is nearly comparable to HunyuanVideo. In particular, on Movie-Gen, which involves relatively simpler video generation scenarios, our method even surpasses HunyuanVideo in terms of physical consistency. These results further demonstrate the effectiveness of our joint modeling approach in generating physically consistent videos.

Table 5: Quantitative results between our VideoPF method and HunyuanVideo-1.5-8.3B on VideoJAM-Bench.

Method	**Motion**			**Appearance**					**Physical**	
	MS	DD	agg.	SC	BC	AQ	IQ	agg.	SA	PC
Hunyuan	**0.989**	**0.922**	**0.954**	**0.942**	**0.948**	**0.580**	0.628	**0.767**	0.627	**0.267**
Ours	**0.989**	0.867	0.937	0.936	0.947	0.558	**0.646**	0.764	**0.629**	0.262

Table 6: Quantitative results between our VideoPF method and HunyuanVideo-1.5-8.3B on Movie-Gen.

Method	**Motion**			**Appearance**					**Physical**	
	MS	DD	agg.	SC	BC	AQ	IQ	agg.	SA	PC
Hunyuan	**0.991**	**0.797**	**0.917**	**0.937**	**0.948**	**0.589**	0.638	**0.771**	0.412	0.334
Ours	0.984	0.686	0.864	0.932	0.940	0.582	**0.659**	0.770	**0.454**	**0.356**

Table 7: Quantitative results between our VideoPF method and HunyuanVideo-1.5-8.3B on VideoPhy-Bench.

Method	**Motion**			**Appearance**					**Physical**	
	MS	DD	agg.	SC	BC	AQ	IQ	agg.	SA	PC
Hunyuan	**0.991**	0.864	**0.939**	**0.927**	**0.945**	**0.528**	0.609	**0.744**	**0.125**	**0.227**
Ours	0.982	**0.867**	0.920	0.900	0.929	0.508	**0.614**	0.727	0.121	0.215

H.5 COMPARISON WITH VIDEOJAM

For VideoJAM, since the method is not open source, we can only conduct a limited comparison based on the settings described in the original paper. The results are presented in Table 8.

It can be observed that our method is competitive with VideoJAM in terms of motion dynamics, and surpasses VideoJAM-4B in appearance quality.

Combining the results from H.4 and H.5, we find that jointly modeling physical field signals and visual content enables the model to significantly enhance its understanding of physical principles and dynamics, achieving performance improvements disproportionate to its parameter scale, while preserving strong visual fidelity.

These findings suggest that, for models with relatively small parameter sizes, their representational capacity has not yet been fully exploited, and there remains substantial room for further optimization and learning.

Table 8: Quantitative results between our VideoPF method and VideoJAM-4B on VideoJAM-Bench.

Method	Motion			Appearance					Physical	
	MS	DD	agg.	SC	BC	AQ	IQ	agg.	SA	PC
VideoJAM	0.993	0.875	0.948	0.878	0.939	0.469	0.489	0.683	-	-
Ours	0.989	0.867	0.937	0.936	0.947	0.558	0.646	0.764	0.629	0.262

H.6 Comparison with PhyT2V

Although our setting differs from methods that rely on external physical knowledge bases, we nevertheless conduct comparisons with such approaches for a fair evaluation. PhyT2V (Xue et al., 2025b) is a representative example, which leverages iterative interactions with multimodal large language models to exploit their embedded physical knowledge for enhancing generated video results. We adopt GPT-5 (Singh et al., 2025) and Qwen3-VL (Bai et al., 2025) as external physical knowledge sources for comparison. Considering the computational cost of the experiments, we sample 50 prompts from VideoJAM-Bench, and the results are presented in Table 9.

Table 9: Quantitative results between our VideoPF method and PhyT2V on the sampled VideoJAM-Bench. "Phy-Qwen" denotes the use of Qwen3-VL as the external physical knowledge base, while "Phy-GPT" refers to employing GPT-5 as the physical knowledge base, and "our-GPT" represents the results of our method augmented with prompts enhanced by PhyT2V.

Method	Motion			Appearance					Physical	
	MS	DD	agg.	SC	BC	AQ	IQ	agg.	SA	PC
Wan2.1	0.985	0.715	0.876	**0.920**	0.941	0.527	0.601	0.739	0.630	0.223
Phy-Qwen	0.984	0.715	0.875	0.919	0.938	0.532	0.600	0.738	0.625	0.227
Phy-GPT	**0.990**	0.717	0.889	0.913	**0.946**	**0.552**	0.605	0.745	**0.655**	0.251
Ours	0.984	**0.877**	0.928	0.916	0.931	0.545	**0.636**	0.747	0.643	0.245
Our-GPT	0.987	0.873	**0.934**	0.914	0.931	0.548	0.635	0.747	0.649	**0.258**

The results in the table reveal that: (i) this direction of approaches is highly dependent on the capability of the external physical knowledge model or simulator. When the external model is not sufficiently strong, the performance gains are limited. In principle, if an ideal physical simulator were powerful enough to faithfully and unbiasedly reproduce any desired phenomenon, such approaches could indeed generate high-quality videos. However, this also raises a fundamental question: if the physical simulator is already so capable, why introduce the stochasticity of a video generation model, instead of directly relying on a rendering-based pipeline for video synthesis? Therefore, balancing the roles of the external knowledge model and the video generation model remains an unavoidable challenge. (ii) Our method remains highly competitive, and importantly, it is not orthogonal to this part of approaches. While our method focuses on enhancing the intrinsic physical understanding of the model, these approaches improve performance via the external physical simulation, the two are complementary and can be combined. As evidenced by the results, using prompts enhanced by PhyT2V also leads to further improvements when applied to our method.

Finally, it is worth noting the impact of computational overhead: the average generation time per video for different methods is reported in Table 10.

Table 10: The average time required by different methods to generate a single video.

Method	Time (seconds per video)
Wan2.1	175.4718
PhyT2V	1464.6826
Ours	295.4291

Whether external LLMs or explicit physical simulators are used as physical knowledge bases, substantial additional time is required during inference to perform physical reasoning or planning, which often exceeding the time needed for video generation itself. This overhead can significantly reduce overall system efficiency and scalability, making such approaches less practical for real-time or large-scale applications. It may also introduce latency bottlenecks in interactive settings, increase computational cost, and limit their deployment in resource-constrained environments.

H.7 User Study

We conduct a user study following the two-alternative forced-choice protocol. For each evaluation case, we present users with two videos generated respectively by our method and the baseline method, and ask them to evaluate their relative performance across three aspects: semantic adherence, visual quality, and physical consistency. The interface of our user study is illustrated in Figure 8.

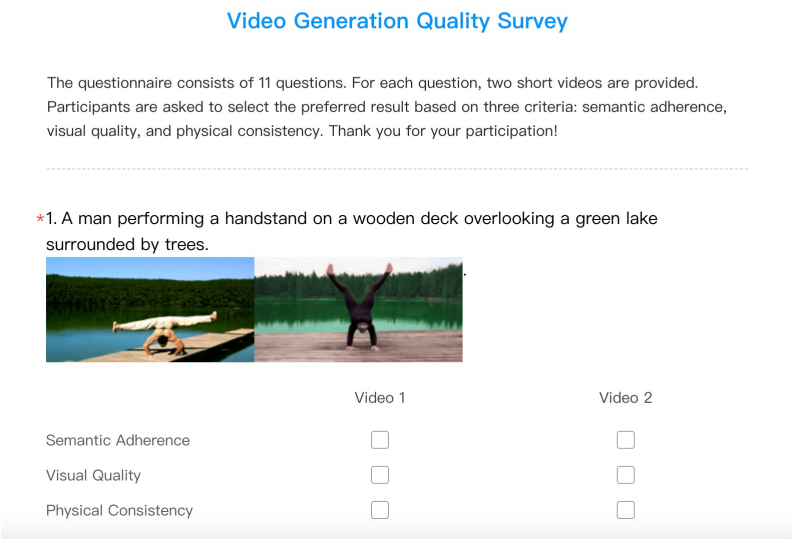


Figure 8: The interface and descriptions of the user study.

For each participant, 11 cases are randomly sampled from the benchmark set to mitigate potential ordering bias. In total, we recruit 30 participants and report the average preference rate of our method relative to the baselines, as summarized in Table 11.

Table 11: Compared with the baseline results, the preference rate of our method.

Method	Semantic Adherence	Visual Quality	Physical Consistency
CogVideoX	0.827	0.788	0.885
VideoREPA	0.65	0.717	0.633
Wan2.1	0.672	0.688	0.688

H.8 Detailed Scores of Ablation Studies

In this section, we provide the detailed quantitative results for the ablation studies presented in Section 4.3, as shown in Table 12, 13, and 14. It can be observed that the proposed module consistently improves both motion dynamics and physical consistency in the generated outputs.

H.9 More Qualitative Results

In this section, we present additional visual results, as shown in Figure 9.

Table 12: Ablation results on whether the Adaptive Physical-Aware Data Pipeline is employed. *Adaptive* denotes the use of the proposed method, whereas *Random* denotes results obtained by randomly sampling an equal amount of data for experimentation.

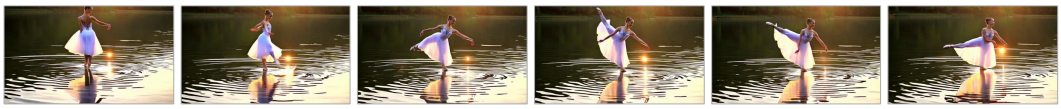
Method	Motion			Appearance					Physical	
	MS	DD	agg.	SC	BC	AQ	IQ	agg.	SA	PC
Random	0.992	0.367	0.778	0.956	0.950	0.551	0.631	0.766	0.573	0.228
Adaptive	0.983	0.617	0.839	0.937	0.946	0.543	0.613	0.752	0.608	0.244

Table 13: Ablation results on whether incorporating auxiliary physical field signals during training. *w/o all* denotes the original architecture without any additional inputs. *w/o field* removes the velocity and acceleration field information. *w/o depth* excludes the depth field. *Our* denotes the full model utilizing all auxiliary information.

Method	Motion			Appearance					Physical	
	MS	DD	agg.	SC	BC	AQ	IQ	agg.	SA	PC
w/o all	0.983	0.617	0.839	0.937	0.946	0.543	0.613	0.752	0.608	0.244
w/o field	0.977	0.730	0.863	0.916	0.937	0.547	0.624	0.747	0.609	0.243
w/o depth	0.982	0.766	0.887	0.928	0.939	0.557	0.621	0.753	0.614	0.258
ours	0.989	0.867	0.937	0.936	0.947	0.558	0.646	0.764	0.629	0.262

Table 14: Ablation results on the use of Physical Consistency Refinement. *w/o refine* denotes the variant without this component, while *our* denotes the model with it enabled.

Method	Motion			Appearance					Physical	
	MS	DD	agg.	SC	BC	AQ	IQ	agg.	SA	PC
w/o refine	0.987	0.857	0.929	0.925	0.937	0.558	0.636	0.755	0.613	0.246
ours	0.989	0.867	0.937	0.936	0.947	0.558	0.646	0.764	0.629	0.262



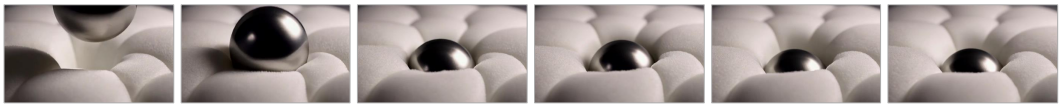
(a) A graceful ballet dancer performing on the surface of a calm, mirror-like lake. She moves elegantly and creates subtle ripples and soft splashes beneath her feet. The water reflects her silhouette perfectly, enhancing the symmetry and visual harmony.



(b) The thick, viscous golden honey slowly poured onto a rotating wooden stirring stick. Under gravity, portions of the honey gradually detach and fall in slow motion.



(c) A solid gold cube is dropped into the sandbox. Upon impact, the cube displaces the sand, and the sand behaves like a granular material: particles scatter and pile up naturally, not like a liquid. The heavy gold cube sinks partially into the sand due to its high density, then gradually comes to rest.



(d) A heavy metal sphere is dropping vertically onto a soft foam cushion. The metal sphere barely rebounds due to energy absorption, lifting only slightly before settling back into the softened surface.



(e) A transparent glass is placed on a table. Clear water is steadily poured into the glass from above.



(f) A massive silk flag is mounted on a tall flagpole, violently billowing in strong wind.



(g) A man performing a handstand on a wooden deck overlooking a green lake surrounded by trees.



(h) A woman doing a headstand on a beach.

Figure 9: Additional generated samples of the proposed VideoPF method.

H.10 FINE-GRAINED SCENARIO COMPARISON

Following the taxonomy of PhyGenBench (Meng et al., 2025), we additionally evaluate four categories of physical scenarios (Force, Light, Heat, and Physical Properties), where our method consistently improves motion quality, semantic alignment, and physical commonsense across nearly all categories, demonstrating robust generalization under diverse physical contexts.

- **Force**. This category encompasses scenarios related to gravity, buoyancy, collision, and other associated physical interactions. 40 prompts.
- **Light**. This category encompasses scenarios related to reflection, refraction, and other associated optical phenomena. 50 prompts.
- **Heat**. This category encompasses scenarios related to melting, freezing, and other associated phase-change phenomena. 30 prompts.
- **Physical Properties**. This category encompasses scenarios related to intrinsic physical attributes of objects, including flammability, hardness, and other inherent material properties. 25 prompts.

Table 15: Comparison of results across different fine-grained scenarios.

Category	Method	Motion	Appearance	SA	PC
Force	Wan2.1	0.798	0.689	0.324	0.274
	Ours	**0.838**	**0.710**	**0.346**	**0.323**
Light	Wan2.1	0.742	0.711	0.391	0.279
	Ours	**0.760**	**0.733**	**0.402**	**0.293**
Heat	Wan2.1	0.689	0.736	0.221	**0.340**
	Ours	**0.708**	**0.748**	**0.227**	0.337
Physical Properties	Wan2.1	0.747	0.680	0.261	0.199
	Ours	**0.821**	**0.723**	**0.398**	**0.236**

From these results, it can be observed that our method consistently outperforms the backbone model across all evaluated scenarios. Moreover, owing to our specific modeling strategy and data processing pipeline, dynamic physical properties emerge as the primary beneficiaries of our approach. In particular, our method substantially enhances the plausibility of motion trajectories and the coherence of spatial structures within the generated videos. While the improvement is less pronounced for heat-related phenomena, which primarily depend on appearance transformation rather than explicit dynamics, consistent gains are observed in motion, semantic alignment, and the remaining physics categories.

Besides, we conducted additional controlled experiments on four simple but representative dynamic scenarios of real-world. We conduct scenarios of free-fall, projectile motion, buoyancy, and sphere collision, and generate 5 videos for each scenario. For each generated video, we detected and tracked the sphere using GroundingDINO (Liu et al., 2023) and SAM (Kirillov et al., 2023), recorded its trajectory, and normalized the coordinates to the (0, 1) interval. Under the assumption that a model is capable of capturing underlying physical properties, the future trajectory predicted from the first part of the observed motion should closely align with the actual generated trajectory. To this end, we used the first four frames following the appearance of the sphere as reference, employed a physical simulator to predict and compute the subsequent trajectory, and compared the predicted trajectory against the actual generated one by computing the RMSE. The results are reported below.

Table 16: Comparison Results Using Controllable Metrics.

Method	free-fall	projectile motion	buoyancy	collision
Wan2.1	0.4504	1.3378	1.5903	0.5175
Ours	0.3868	0.7456	1.2046	0.4395

These controlled experiments complement existing VLM-based evaluations by directly measuring trajectory consistency under known dynamics, providing additional evidence that the proposed method improves physically plausible motion beyond perceptual quality alone.