PhysAlign: Learning Physical Priors for Dynamical Event-Driven Video Generation via Representation Alignment

Mengxian Li[1], Zhan Wang[1], Fan Qi[1,2⋆], and Changsheng Xu[3]

[1] Tianjin University of Technology, Tianjin, China
{jonasrem7,wangzhan14}@stud.tjut.edu.cn, fanqi@email.tjut.edu.cn
[2] Inner Mongolia University, Hohhot, China
[3] Institute of Automation, Chinese Academy of Sciences, Beijing, China
csxu@nlpr.ia.ac.cn

Abstract. Current video generation models produce highly realistic visuals but frequently fail to maintain physical and causal consistency when facing complex dynamic events, such as collisions or collapses. We attribute this limitation to the fact that existing models primarily fit visual data distributions, entangling dynamics with appearance without explicit physical priors to govern motion evolution. To address this, we introduce PhysAlign, a framework for Dynamical Event-Driven Physically Consistent Video Generation. PhysAlign decouples physical priors from visual appearance via a Physical Dynamics Extractor and a discrete Universal Physical Codebook. Recognizing the temporal locality of dynamic events, an Event-Aware Temporal Gating module dynamically controls the injection of physical priors, while Physics-Guided Alignment distills continuous physical topologies from a video foundation model. Furthermore, we construct EventPhy, a 26K-video dataset with structured, VLM-generated causal annotations to benchmark dynamical event-driven video generation. Extensive evaluations show that PhysAlign achieves state-of-the-art physical reasoning, obtaining leading Semantic Adherence and Physical Commonsense scores on challenging benchmarks like PhysicsIQ and VideoPhy, while preserving high visual fidelity. Project Page: https://github.com/FanQi-AI/PhysAlign

Keywords: Video Synthesis · Diffusion Models · Representation learning

1 Introduction

The field of video generation has experienced significant advancements [9,12,20, 26,31,41]. Generative models utilizing architectures such as Diffusion Transformers (DiT) [20, 30] achieve high visual quality and temporal coherence, demonstrating the potential to construct world models [18, 23, 43]. However, existing models lack an understanding of physical laws and causal relationships [5, 6].

⋆ Corresponding author.

When processing dynamic events that trigger sudden changes in the state of objects, such as collisions, pushing, or failures of support, these models struggle to generate correct results. This reveals a critical limitation of current models: while they excel at simulating simple, continuous motion, they lack the causal reasoning required to predict sudden changes in state induced by dynamic events [11,28]. To enhance the physical understanding of current models, we propose the task of Dynamical Event-Driven Physically Consistent Video Generation. Conditioned on the initial state of the scene and future dynamic events, this task requires generative models to predict coherent subsequent frames wherein the evolution of state for relevant objects follows fundamental physical causality.

To enhance the physical realism of generative models, previous studies explore general physical augmentation extensively [21,27]. These methods generally fall into two categories. The first category relies on explicit constraints, injecting physical laws into the generation process through external physics simulators [24] or performing step-by-step physical reasoning using large language models and vision-language models [45, 46]. Although these methods provide precise control for specific scenarios, their reliance on external systems introduces high computational overhead and severely limits the generalization ability of the models. The second category turns to implicit physical representations, attempting to internalize physical priors directly into generative models. This includes distilling structural and motion priors from foundation models through feature alignment [8, 50], optimizing physical representations via reinforcement learning [17], and introducing specialized physical expert modules for feature guidance [38,42]. However, these methods reveal critical limitations when applied to the proposed task of dynamical event-driven video generation. Because existing approaches primarily focus on maintaining general spatiotemporal smoothness, they become ineffective when facing dynamical events that trigger sudden changes in state [48,51]. Furthermore, struggling to decouple the rules of physical dynamics from the distribution of visual data, these methods frequently rely on the fitting of visual continuity as a shortcut for optimization, thereby hiding crucial signals of mechanical interaction. This weakness prevents accurate causal reasoning and ultimately causes the models to produce physical anomalies when predicting the outcomes of events.

To address these challenges, we propose PhysAlign, a unified framework integrating the guidance of specialized physical modules with the distillation of foundation models to achieve Dynamical Event-Driven Physically Consistent Video Generation. First, Physical Dynamics Extractor utilizes techniques of photometric augmentation to disturb surface visual distributions, forcing the model to focus on the extraction of underlying spatiotemporal dynamics. Building upon this, we introduce a discrete Physical Codebook that serves as an information bottleneck to eliminate redundant appearance information and extract the physical priors for retrieval during inference. Furthermore, an Event-Aware Temporal Gating module predicts the time windows of interactions to dynamically adjust the injection strength of physical features, ensuring precise synchronization with the occurrence of events. Finally, a strategy of Physics-Guided Alignment

based on the distillation of spatiotemporal feature relations maps the deep spatial topology and continuous temporal evolution from pre-trained foundation models into the generative space, thereby enhancing physical reasoning without lowering the quality of visual generation.

The proposed method successfully addresses video generation driven by dynamical events, outperforming existing approaches across multiple datasets. Our main contributions are as follows:

- **We propose the task of Dynamical Event-Driven Physically Consistent Video Generation and introduce PhysAlign, a novel generative framework.** By decoupling physical dynamics rules from visual appearances and utilizing event-aware temporal gating, this method achieves targeted injection of physical priors to ensure robust alignment with physical causality.
- **We construct a specialized 26K-video dataset dedicated to dynamical event-driven video generation, encompassing 16 distinct interaction scenarios.** To bridge the gap in existing benchmarks, we leverage VLMs to generate structured, causal annotations that explicitly detail event mechanisms and state changes, thereby providing a rigorous foundation for model training and evaluation.
- **We conduct extensive evaluations across multiple benchmarks to demonstrate the effectiveness of PhysAlign.** Experimental results show that our method significantly outperforms existing baselines in physical reasoning, accurately predicting dynamical state changes while preserving high visual quality and semantic fidelity.

2 Related Work

2.1 Video Generation

Video generation has transitioned from Generative Adversarial Networks [14] to Diffusion models [15]. Early video diffusion models succeeded by extending image architectures like U-Net [34] and 3D U-Net [13] temporally. Employing spatio-temporal denoising networks for end-to-end generation, they exhibited improved motion consistency and finer details over previous approaches. Latent video diffusion and system-level architectures emerged to extend duration, resolution, and controllability. Frameworks like Align Your Latents [10] shifted generation into the Variational Autoencoder (VAE) [19] latent space, reducing computational costs while supporting higher resolutions and temporal consistency.

Recent open-source video foundation models, including Open-Sora [53], Wan [37], and CogVideoX [47], shift from U-Net backbones to DiT [30] architectures. These frameworks typically encode videos into sequences of spatio-temporal patches or tokens, apply latent diffusion modeling via Transformer blocks, and decode representations back to pixels using 3D or spatio-temporal VAEs. Open-Sora [53] adopts a spatio-temporal diffusion Transformer for efficient sequence modeling. CogVideoX [47] combines a 3D causal VAE with a Transformer framework

to support long-duration, high-resolution generation across varying aspect ratios. Wan [37] introduces a novel video VAE alongside system-level scaling and data curation strategies enhancing efficiency and quality. Notable community-developed models like Latte [25] and Open-Sora-Plan [22] further facilitate architecture unification. Although achieving high-quality results by scaling data and parameters, these DiT-based models can still generate physically implausible content. This suggests that expanding training data alone may be insufficient for robust physical reasoning in open-domain dynamics.

2.2 Physical-aware Video Generation

To improve physical realism, explicit constraint methods enhance consistency by injecting physical constraints into the generation pipeline via external engines, trajectories, or inference models. PhysGen [24] integrates a rigid body dynamics simulator predicting force-driven trajectories to guide generation. PhyT2V [45] utilizes chain-of-thought reasoning and iterative self-correction via Large Language Models (LLMs) to enforce physical common sense. PhysAnimator [44] integrates deformable body simulations to improve realism, while PhysDreamer [49] leverages physics-based simulation generating plausible dynamic responses from static objects, producing temporally coherent motions and deformations. Exploring architecture design, WISA [38] decomposes physical principles into a Mixture of Physical Experts Attention module. ProPhy [42] and VLIPP [46] leverage Vision-Language Models (VLMs) [2–4, 40] to explicitly capture dynamics; ProPhy aligns tokens via VLM supervision, whereas VLIPP predicts cross-modal physical trajectories for a video diffusion model. Although providing precise control, these methods frequently face system complexity, substantial computational overhead, and limited generalization.

Conversely, implicit constraint methods transfer structural and physical priors from foundation models into the generative model feature space. MotionDirector [52] customizes motion by decoupling appearance and motion through separate spatio-temporal adaptation modules on a frozen video diffusion model. VideoREPA [50] introduces Token Relationship Distillation aligning spatio-temporal relationships between generative and understanding models. MoAlign [8] separates appearance and motion representations by learning a motion subspace and aligning generative features. The PhysMaster [17] designs a PhysEncoder extracting implicit physical representations, applying reinforcement learning like Direct Preference Optimization (DPO) [32] to enhance physical perception. Utilizing structural priors from representation models, these approaches enable generative models to learn physics-aware spatio-temporal representations rather than relying strictly on data fitting. However, their difficulty in decoupling physical principles from visual appearance distributions limits their capacity to handle abrupt state transitions triggered by dynamical events. In this work, we propose PhysAlign, explicitly isolating physical priors into a discrete codebook and employing an event-aware temporal gating mechanism to dynamically inject constraints at precise interaction moments.

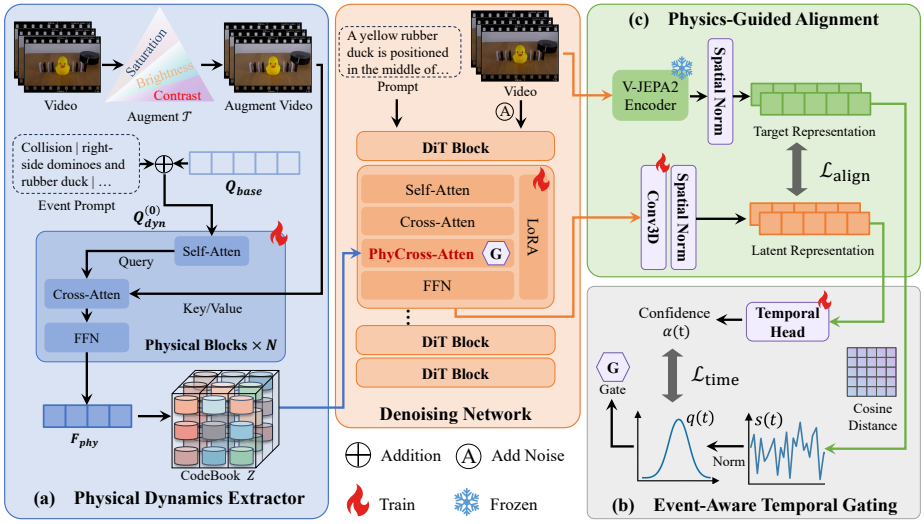

Fig. 1: Overview of the training framework. (a) Physical Dynamics Extractor captures physical representations from videos and writes them into a codebook. (b) Event-Aware Temporal Gating predicts adaptive temporal weights. (c) Physics-Guided Alignment aligns mid-layer features with the target representations of V-JEPA2.

3 Method

3.1 Preliminaries and Problem Formulation

Flow Matching for Video Generation. This work builds on DiT, which represents videos as flattened spatiotemporal token sequences and models their distribution through flow matching. Flow matching learns a velocity field that transports standard Gaussian noise to the data distribution. Given a real video latent (x_1) and Gaussian noise (x_0), their linear interpolation at $(t \in [0, 1])$ is:

$$\mathcal{L}_{FM} = \mathbb{E}_{t,x_1,x_0,c}\left[\|v_\theta(x_t, c) - u_t(x_t|x_1)\|_2^2\right], \tag{1}$$

where $u_t(x_t|x_1) = \frac{\mathrm{d}x_t}{\mathrm{d}t} = x_1 - x_0$ represents the ground-truth target vector field driving the noise x_0 to the data x_1.

Problem Formulation. Unlike general video generation tasks, this work focuses on Dynamical Event-Driven Physically Consistent Video Generation. This task aims to simulate the subsequent evolution of a scene under specific physical interactions. Formally, we learn a generative mapping $\mathcal{G}$. The input comprises the initial scene state S_{init}, typically the first frame I_0 and a basic description C_{scene}, and the dynamics event description C_{Event} specifying upcoming physical interactions, such as collisions or collapses. The objective is to generate the future video sequence $\hat{V} = \{\hat{I}_1, \ldots, \hat{I}_T\}$, where $\hat{I}_t$ denotes the generated frame at time step t, and T is the number of future frames to be generated.

$$\hat{V} = \mathcal{G}(S_{init}, C_{Event}). \tag{2}$$

3.2 Physical Dynamics Extractor

As shown in Fig. 1(a), we propose a Physical Dynamics Extractor (PDE) module. This module is designed to extract physical features from real videos and subsequently inject them into the generative model as prior knowledge.

Physical Feature Extraction. During training, models frequently exploit appearance cues as optimization shortcuts, thereby overlooking physical representations. To mitigate this phenomenon, we introduce a dedicated feature extraction pathway. We apply photometric augmentations $\mathcal{T}(\cdot)$, including adjustments to brightness, contrast, and saturation [33], to the input video V_{raw}. By maintaining consistent perturbation parameters across all frames, we generate an augmented video $V = \mathcal{T}(V_{raw})$ that disrupts the distribution of visual appearance while preserving the temporal dynamics. Subsequently, V is encoded by WanVAE to derive the spatio-temporal representation $F_{video} \in \mathbb{R}^{B\times F\times H\times W\times D}$, where B, F, H, W, and D denote the batch size, temporal latent sequence length, spatial height, width, and channel dimension, respectively.

To extract physical evidence from the appearance-augmented video representation F_{video}, we employ an event-driven cross-attention retrieval mechanism. We initialize learnable dynamic queries by combining the event description C_{Event} with a base physical prior $Q_{base} \in \mathbb{R}^{1\times M\times D}$:

$$Q_{dyn}^{(0)} = \text{LayerNorm}\big(Q_{base} + \text{MLP}(C_{Event})\big), \tag{3}$$

where $\text{MLP}(\cdot)$ denotes a multi-layer perceptron that projects the event description C_{Event} into the same channel dimension D as the base physical prior.

These initial queries $Q_{dyn}^{(0)} \in \mathbb{R}^{B\times M\times D}$ are iteratively updated through N stacked Physical Blocks (PB). Within the n-th block ($n \in \{1,\dots,N\}$), self-attention models global dynamic hypotheses, while cross-attention retrieves relevant spatio-temporal physical patterns by treating $Q_{dyn}^{(n-1)}$ as the query and F_{video} as the key and value:

$$\text{Cross-Attn}(Q_{dyn}^{(n-1)}, F_{video}) = \text{Softmax}\left(\frac{(W_Q Q_{dyn}^{(n-1)})(W_K F_{video})^T}{\sqrt{d_k}}\right)(W_V F_{video}), \tag{4}$$

where W_Q, W_K, and W_V are learnable linear projection matrices for the queries, keys, and values, and d_k is the scaling factor representing the channel dimension of the keys. After passing through all N blocks, the final updated query is output as the extracted physical feature $F_{phy} = Q_{dyn}^{(N)} \in \mathbb{R}^{B\times M\times D}$.

The physical feature F_{phy} is injected into specific mid layers of the DiT backbone via an additional PhysCross-Attention layer to constrain the generation process. In the PhysCross-Attention, the latent features of the backbone serve as the Query, while F_{phy} acts as the Key and Value. Mid layers are specifically targeted because they exhibit heightened sensitivity to motion structures and interaction relationships [18, 50].

Physical Codebook and Inference Pipeline. To disentangle appearance information from physical features, we introduce a physical codebook, defined as $\mathcal{Z} = \{z_k\}_{k=1}^{K} \in \mathbb{R}^{K\times D}$, where K denotes the number of codebook entries, and z_k represents the k-th discrete physical primitive with channel dimension D. The quantization of continuous representations into finite primitives establishes an information bottleneck that filters visual noise, thereby compelling the learning of robust physical evolution patterns. Furthermore, the codebook functions as an important intermediate representation for the retrieval of dynamic priors when future frames are inaccessible during inference. To maintain differentiability and capture complex mixed physical states during training, a soft quantization strategy is adopted. Specifically, for each physical feature vector $f_i \in F_{phy}$ produced by the feature extraction pathway, we compute its scaled dot-product similarity against all codebook entries to obtain normalized weights $w_{i,k}$, which are then used to reconstruct the quantized feature $\hat{f}_i$ via weighted summation:

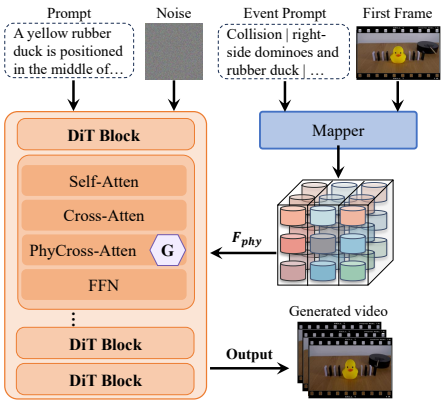

Fig. 2: Overview of the inference pipeline. Given the first-frame image and event prompt, the Mapper queries physics features from the codebook and injects them into PhyCross-Attention.

$$w_{i,k} = \frac{\exp\left(f_i^\top z_k \,/\, \tau_{vq}\sqrt{D}\right)}{\sum_{j=1}^{K} \exp\left(f_i^\top z_j \,/\, \tau_{vq}\sqrt{D}\right)}, \quad \hat{f}_i = \sum_{k=1}^{K} w_{i,k}\, z_k, \tag{5}$$

where τ_{vq} is a temperature hyperparameter that controls the sharpness of the soft assignment distribution. These quantized vectors form the final physical prior $\hat{F}_{phy} \in \mathbb{R}^{B\times M\times D}$, which is subsequently injected into the generation process.

The codebook is trained with $\mathcal{L}_{VQ}$ to align the codebook entries and the continuous physical features:

$$\mathcal{L}_{VQ} = \|\mathrm{sg}[F_{phy}] - E\|_2^2 + \beta\,\|F_{phy} - \mathrm{sg}[E]\|_2^2\,, \tag{6}$$

where $E = [e_1, \ldots, e_M]$ denotes the nearest codebook entries with $e_i = z_{k^*}$, $k^* = \arg\max_k \; f_i^\top z_k$, and $\mathrm{sg}[\cdot]$ is the stop-gradient operator, and β is a hyperparameter that controls the weight of the commitment loss.

Following the training of the codebook, a critical disparity emerges between the training and inference phases. During training, the PDE processes real video to derive the physical features F_{phy}. However, because future frames are unavailable during inference, these physical priors need to be estimated from initial conditions. To resolve this issue, a lightweight predictor, called the Mapper, is introduced as illustrated in Fig. 2.

Specifically, the Mapper receives the first frame I_0 and the event description C_{event} as inputs to generate continuous predicted physical features $\hat{F}_{phy}^{pred} \in \mathbb{R}^{M \times D}$. To get the physical prior, these features are quantized. We achieve this quantization by identifying the nearest neighbors of the features within the pre-trained codebook, yielding $\hat{F}_{phy}$. Subsequently, this quantized prior is injected into the denoising backbone through PhysCross-Attention layers. During Mapper training, we freeze the PDE module. The continuous features F_{phy}, extracted from the real video, function as the ground-truth target. The Mapper is trained utilizing an MSE loss:

$$\mathcal{L}_{mapper} = \left\| \hat{F}_{phy}^{pred} - \mathrm{sg}[F_{phy}] \right\|_2^2, \tag{7}$$

where sg[·] denotes the stop-gradient operator. It is crucial to emphasize that this loss is computed within the continuous feature space, strictly prior to the codebook quantization step. Bypassing the non-differentiable quantization operation during the training phase of the Mapper ensures a stable gradient flow.

3.3 Event-Aware Temporal Gating

Dynamic events, such as collisions and support failure, exhibit strong temporal locality. To address this, we introduce an event-aware temporal gating module depicted in Fig. 1(b) to selectively activate physical constraints within precise time windows. Because acquiring fine-grained temporal annotations via Vision-Language Models [3] is computationally prohibitive and often imprecise, we propose a self-supervised label generation strategy based on feature change rates. Operating on the premise that dynamic events correlate with drastic variations in deep semantic features, we extract a sequence of deep spatial feature maps $\{F_1, \ldots, F_T\}$ from real videos using pretrained foundation models such as V-JEPA2 [1], DINOv3 [35], or VideoMAEv2 [39]. Each feature map is denoted as $F_t \in \mathbb{R}^{H_f \times W_f \times D_{feat}}$, where H_f, W_f, and D_{feat} represent the spatial height, width, and channel dimension, respectively. Let $f_t^{(h,w)} \in \mathbb{R}^{D_{feat}}$ denote the feature vector at spatial location (h, w) of frame t. We first compute the cosine distance between adjacent frames at each spatial location:

$$d_{h,w}(t) = 1 - \frac{f_t^{(h,w)} \cdot f_{t-1}^{(h,w)}}{\|f_t^{(h,w)}\|_2 \, \|f_{t-1}^{(h,w)}\|_2}, \quad t = 2, \ldots, T. \tag{8}$$

To suppress static backgrounds and capture the most significant dynamic cues, we aggregate these local variations via top-K spatial mean pooling. We define $K = \lfloor \rho \cdot H_f W_f \rfloor$, where $\rho \in (0, 1]$ is the spatial selection ratio. The frame-level dissimilarity is computed as:

$$\bar{d}(t) = \frac{1}{K} \sum_{(h,w) \in \mathcal{T}_K(t)} d_{h,w}(t), \quad t = 2, \ldots, T, \tag{9}$$

where $\mathcal{T}_K(t)$ denotes the set of K spatial locations with the highest distance values at time step t, and $\bar{d}(1) = 0$. We then apply a one-dimensional Gaussian kernel $\mathcal{G}_\sigma$ along the temporal axis and normalize the result into a soft probability distribution via a temperature-scaled softmax:

$$s(t) = (\mathcal{G}_\sigma * \bar{d})(t), \quad q(t) = \frac{\exp(s(t)/\tau_{time})}{\sum_{t'=1}^{T} \exp(s(t')/\tau_{time})}, \tag{10}$$

where $*$ denotes the temporal convolution operator and τ_{time} is a temperature parameter that controls the sharpness of the resulting distribution.

We predict this event distribution from latent features by extracting intermediate DiT features and processing them through a lightweight temporal head. The module outputs a frame-wise probability distribution $p(t)$ representing the event occurrence likelihood at each timestep. KL divergence supervision aligns this prediction head with the self-supervised teacher distribution:

$$\mathcal{L}_{time} = D_{KL}(q \| p). \tag{11}$$

During training, the teacher distribution $q(t)$ serves as a temporal gating signal for the PhysCross-Attention layers. Specifically, let $h_{in}^{(l)}$ and $h_{out}^{(l)}$ denote the input and output hidden states of the l-th DiT block, respectively. The modified injection mechanism becomes: $h_{out}^{(l)} = h_{in}^{(l)} + \alpha(t) \cdot \text{PhyCrossAttn}^{(l)}(h_{in}^{(l)}, \hat{F}_{phy})$, where $\alpha(t) = q(t) \cdot T$ is the per-frame gate value, scaled such that a uniform distribution yields $\alpha(t) = 1$ (i.e., no gating effect). During critical event phases ($\alpha(t) > 1$), the gate amplifies the influence of physical constraints, while during static phases ($\alpha(t) < 1$), the gate attenuates these features, thereby reverting toward standard generation. This selective modulation introduces physical priors without compromising the visual fluency of non-interactive segments.

3.4 Physics-Guided Alignment

To endow the generative model with robust physical reasoning capabilities, we introduce a physics-guided feature alignment distillation module shown in Fig. 1(c). We align the latent features of the DiT backbone with the deep features of a frozen V-JEPA2 [1]. We extract the features $F_T \in \mathbb{R}^{B \times D_t \times F_t \times H_t \times W_t}$ from a frozen V-JEPA2 [1] and apply spatial normalization $\text{Norm}(\cdot)$ to isolate structural information from feature magnitude [36], yielding $\hat{F}_T = \text{Norm}(F_T)$. For the generative framework, the features $F_S \in \mathbb{R}^{B \times D_s \times F_s \times H_s \times W_s}$ from the mid layers of the DiT backbone are processed by a 3D Convolutional layer [36] to match the teacher's channel dimensions and spatio-temporal resolution, resulting in shared dimensions D, F, H, and W. These features are subsequently normalized to yield $\hat{F}_S = \text{Norm}(\text{Conv3D}(F_S))$.

We adopt the Token Relationship Distillation strategy proposed by VideoREPA [50]. Let $x_{t,i} \in \mathbb{R}^D$ denote the token vector at spatial index i within frame t from the normalized feature maps. We construct spatial and temporal relationship matrices via pairwise cosine similarity:

$$M^{spa}_{t,i,j} = \frac{x_{t,i} \cdot x_{t,j}}{\|x_{t,i}\|_2 \, \|x_{t,j}\|_2}, \quad M^{temp}_{t,t',i,j} = \frac{x_{t,i} \cdot x_{t',j}}{\|x_{t,i}\|_2 \, \|x_{t',j}\|_2}. \tag{12}$$

The final physical consistency loss $\mathcal{L}_{align}$ minimizes the discrepancy between the spatio-temporal relationship matrices of the student model (M_S) and the teacher model (M_T). To prevent overfitting to noise in the teacher's representations, we introduce a margin threshold m to suppress gradients from minor discrepancies. This optimization objective is formulated as:

$$\begin{aligned} \mathcal{L}_{align} = & \frac{1}{F(HW)^2} \sum_{t,i,j} \max\left(0, \left| M^{spa}_{S,t,i,j} - M^{spa}_{T,t,i,j} \right| - m\right) \\ & + \frac{1}{F(F-1)(HW)^2} \sum_{t \neq t',i,j} \max\left(0, \left| M^{temp}_{S,t,t',i,j} - M^{temp}_{T,t,t',i,j} \right| - m\right). \end{aligned} \tag{13}$$

The model is optimized end-to-end using a joint loss function, formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{FM} + \lambda_{VQ}\mathcal{L}_{VQ} + \lambda_{time}\mathcal{L}_{time} + \lambda_{align}\mathcal{L}_{align}, \tag{14}$$

where $\lambda_{VQ}, \lambda_{time}$, and λ_{align} are hyperparameters balancing these objectives.

4 Experiment

4.1 Implementation Details

Training Details We adopt the pre-trained Wan2.2-TI2V-5B [37] model, fine-tuning its backbone via LoRA (rank 16) while fully training the PDE module. Experiments are conducted on four NVIDIA A800 GPUs. Input videos are standardized to 480×832 resolution, 81 frames, and 16 fps. Training runs for 8000 steps using AdamW with a global batch size of 16 and a cosine annealing schedule. Learning rates are 1×10^{-4} for LoRA and 2×10^{-4} for the PDE module.
Data Collection and Annotation We categorize dynamical events into sixteen primary types, such as Collision, Support Failure, and Fluid Impact. Based on these established categories, suitable video samples are filtered from three high-quality datasets to construct the custom dataset. These sources encompass two real-world datasets, namely WISA-80K [38] and OpenVidHD [29], alongside one synthetic dataset, Physion v1.5 [7]. Utilizing Qwen3-VL [3], a curated subset comprising approximately 26K videos is extracted and systematically annotated. The annotation pipeline is structured into two sequential stages. Initially, a comprehensive textual description is generated to capture the overall visual content of each video. Subsequently, a fine-grained event description is synthesized based on the specific physical dynamics of the video. To ensure structural consistency, this event description strictly adheres to a predefined format: ⟨Event Type⟩ | ⟨Involved Entities⟩ | ⟨Mechanism of Interaction⟩ | ⟨Description of State Change⟩. More details are provided in Section A of the Supplementary Material.

Table 1: Experimental results of our method compared to existing SOTA and baseline models on the PhysicsIQ, EventPhy, Physion1.5, VideoPhy, and VideoPhy2 datasets. **Higher** values for SA, PC, and Joint indicate better performance.

Methods	PhysicsIQ		EventPhy		Physion1.5		VideoPhy		VideoPhy2		
	SA	PC	SA	PC	SA	PC	SA	PC	SA	PC	Joint
Wan2.2-TI2V-5B	25.18	18.93	13.04	23.03	10.62	18.55	33.76	28.71	17.67	56.67	16.67
CogVideoX-2B	23.17	13.55	12.61	20.87	7.55	17.32	32.42	28.03	13.36	56.67	13.33
CogVideoX-5B	25.66	19.10	13.11	22.34	9.84	18.79	44.62	30.45	17.36	46.70	17.00
PhyT2V(Round4)	27.42	25.08	15.64	30.93	10.36	23.17	45.13	34.47	26.67	65.00	**26.67**
VideoREPA-2B	26.81	24.69	15.01	30.02	10.01	24.49	45.38	35.03	19.67	63.33	19.00
VideoREPA-5B	29.37	25.29	15.56	31.05	11.05	24.90	47.29	35.61	22.67	66.67	22.00
VLIPP	29.15	26.85	15.64	31.23	10.62	25.62	51.17	36.01	**26.67**	63.33	25.33
Ours	**31.54**	**29.75**	**16.35**	**34.83**	**11.05**	**31.49**	**53.98**	**40.54**	23.67	**66.67**	23.33

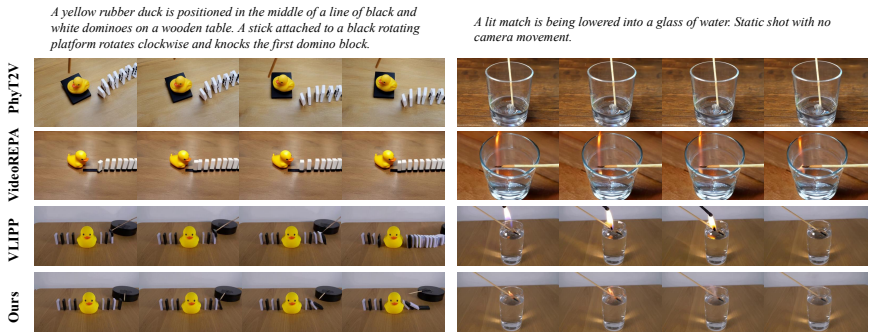

Fig. 3: Qualitative comparison between our method and other methods.

Evaluation The proposed model and baselines are evaluated across VideoPhy [5], VideoPhy2 [6], PhysicsIQ [28], Physion v1.5 [7], VBench [16], and an EventPhy test set. Quantitative analysis primarily employs two VideoPhy [5] metrics: Semantic Adherence (SA) assessing text-to-video alignment, and Physical Commonsense (PC) evaluating physical law adherence. Furthermore, standard VBench metrics compute overall quality scores. The proposed method is quantitatively and qualitatively compared against leading universal video generation models, including Wan2.2-TI2V-5B [37], CogVideoX-2B [47], and CogVideoX-5B [47], alongside SOTA physics-enhanced generation methods like VLIPP [46], VideoREPA [50], and PhyT2V [45].

4.2 Quantitative Comparisons

To evaluate our method on dynamical event-driven video generation, we conduct experiments across three datasets, with the quantitative results detailed in Tab. 1. On the PhysicsIQ benchmark, which emphasizes complex physical logic reasoning, our method gets the highest scores of 31.54 for SA and 29.75 for PC. For the EventPhy dataset, centering on event-level physical generation, the

Table 2: Quantitative comparison of video generation performance on the VBench dataset. The best results are highlighted in **bold**.

Metric	VideoREPA	PhyT2V	VLIPP	Ours
Subject Consistency	95.56	88.78	93.39	**97.31**
Background Consistency	96.42	93.24	95.83	**97.80**
Temporal Flickering	98.23	96.56	97.76	**98.97**
Motion Smoothness	98.87	97.29	98.47	**98.88**
Dynamic Degree	33.33	73.33	53.33	**80.00**
Aesthetic Quality	**50.40**	45.35	48.99	48.92
Imaging Quality	66.59	58.09	65.41	**71.46**
Quality Score	74.68	73.71	75.22	**79.05**

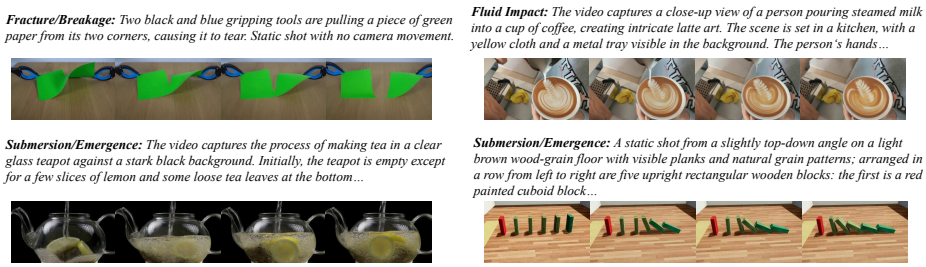

Fig. 4: Additional generated samples, covering various dynamical event types.

model yields 16.35 for SA and 34.83 for PC. Compared with VideoREPA and PhyT2V, our method demonstrates enhanced robustness in managing physical events and multi-object interactions. Furthermore, on the Physion1.5 dataset, our method maintains highly competitive performance, recording an SA of 11.05 and a PC of 31.49. These results demonstrate that our method achieves excellent performance on the dynamical event-driven video generation task.

We further investigate the generalization capability of the proposed method across a broader range of physical scenarios. Specifically, subsequent experiments are conducted on the VideoPhy and VideoPhy2 benchmarks, with the quantitative results presented in Tab. 1. On the VideoPhy dataset, our method achieves the highest scores of 53.98 for SA and 40.54 for PC. For the highly challenging VideoPhy2 dataset, a Joint metric is used to denote the proportion of generated samples where both SA and PC scores are ≥ 4. In this benchmark, our method gets a leading PC score of 66.67, indicating a strong adherence to physical laws. While certain baselines obtain higher SA (26.67) and Joint (25.33) scores, the proposed approach maintains a highly competitive balance, recording an SA of 23.67 and a Joint score of 23.33. Collectively, these findings validate the generalization ability of the proposed method within complex dynamic scenes, suggesting that it effectively captures and reproduces real-world physical processes across diverse contexts. In addition to physical accuracy, the fundamental video generation quality and consistency of the proposed method are evaluated on the VBench [16] dataset, as detailed in Tab. 2. The proposed approach achieves the highest overall Quality Score of 79.05, demonstrating its superiority over existing

Table 3: Ablation study results of different module configurations.

Method	SA	PC
Wan2.2-TI2V-5B(baseline)	13.04	23.03
baseline+LoRA	15.11	26.30
baseline+REPA	15.64	28.87
w/o PDE	15.70	29.05
w/o ETG	16.00	32.23
Ours	**16.35**	**34.83**

Table 4: Ablation study results of feature alignment across layers.

Layers	SA	PC
10	13.57	24.38
14	14.67	27.13
18	**16.35**	**34.83**
22	14.15	25.09
26	13.01	23.29

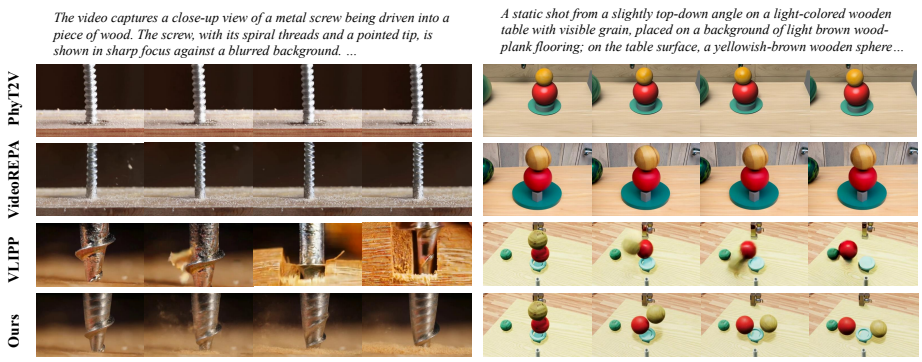

Fig. 5: More qualitative comparison between our method and other methods.

physics-enhanced baselines. Specifically, it exhibits exceptional spatial and temporal coherence, attaining the highest scores across multiple metrics, including Subject Consistency, Background Consistency, Temporal Flickering, and Motion Smoothness. Furthermore, our method secures a leading Dynamic Degree score of 80.00 and an Imaging Quality score of 71.46. Although the Aesthetic Quality score is slightly lower than that of VideoREPA, the comprehensive generation performance remains highly robust. These quantitative results confirm that the proposed method maintains exceptional visual generation quality while ensuring physical realism. More experimental results are available in Section B of the Supplementary Material.

4.3 Qualitative Comparisons

Fig. 3 illustrates the qualitative comparison between our model and SOTA methods [45,46,50]. In the test cases involving dominoes striking a rubber duck and a lit match submerged in water, our method generates physically plausible interactions and abrupt state transitions. In contrast, other approaches exhibit physical violations and logical inconsistencies. Specifically, in the domino scenario, existing methods fail to render the toppling state of the dominoes correctly and occasionally generate new dominoes spontaneously. During match submersion, the matches generated by other methods fail to extinguish after entering the water, or display illogical results like submerging first and igniting later.

To further validate the generalization capability of our method across broader scenarios, Fig. 4 and Fig. 5 present the results of our model on additional predefined categories of dynamical events. These events encompass diverse scenarios, including fracturing/shattering, fluid impact, and submersion/emergence. Experimental results indicate that our method adheres well to text semantics while generating videos with high fidelity, temporal coherence, and alignment with real-world physical laws across various perspectives and backgrounds. This performance highlights the capability of our framework in simulating diverse dynamical events. More results can be found in Section C of the Supplementary Material.

4.4 Ablation Studies

We conduct ablation experiments to demonstrate the effectiveness of our method. All experiments are conducted on the EventPhy dataset.

Ablation on different modules Tab. 3 illustrates the impact of different strategies and module combinations on final performance. Simply introducing basic LoRA or conventional REPA strategies to the baseline model brings limited gains. For example, adding REPA only increases SA and PC to 15.64 and 28.87, respectively, which remains insufficient to handle dynamical events. When we remove the proposed core modules from the complete framework, performance experiences a significant decline. Removing the PDE module drops SA to 15.70 and PC to 29.05, while removing the ETG module similarly degrades the performance. The complete method integrating all proposed modules ultimately achieves the highest scores of 16.35 for SA and 34.83 for PC. These results demonstrate that PDE and ETG have a positive impact on enhancing video physical consistency and capturing spatiotemporal motion features.

Ablation on alignment layers Concerning the selection of alignment layers, previous studies [50] indicate that feature alignment effectiveness is sensitive to the chosen Transformer layer depth in the denoising network. Therefore, we test different alignment strategies from Layer 10 to Layer 26 in Tab. 4. The experimental data exhibits an inverted U-shaped trend. Selecting shallower layers such as Layer 10 with a PC of 24.38 or deeper layers such as Layer 26 with a PC of 23.29 for alignment results in suboptimal generation quality. The model performance reaches its peak at Layer 18 with an SA of 16.35 and a PC of 34.83. This phenomenon is intuitive, as shallow features tend to favor low-level visual textures, whereas excessively deep network layers are overly abstracted into high-dimensional semantics and easily lose fine-grained physical motion information. Layer 18 balances and fuses high-dimensional semantic features with low-level motion dynamics, thereby providing a suitable feature representation space for learning physical laws.

5 Conclusion

This paper introduces PhysAlign, a generative framework designed to address the lack of physical and causal consistency in current video generation models. By

explicitly decoupling physical priors from visual appearances and distilling spatiotemporal topologies from a foundation model, the proposed approach significantly enhances the physical reasoning capabilities of the generative backbone. Extensive evaluations across multiple benchmarks demonstrate that PhysAlign achieves state-of-the-art physical plausibility and semantic alignment, providing a robust solution for simulating real-world physical dynamics.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62576246, 62576179 and 62532004.

References

1. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., et al.: VideoPhy: Evaluating physical commonsense for video generation. In: The Thirteenth International Conference on Learning Representations (2025)
6. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In: The Fourteenth International Conference on Learning Representations (2026)
7. Bear, D.M., Wang, E., Mrowca, D., Binder, F.J., Tung, H.Y., Pramod, R.T., et al.: Physion: Evaluating physical prediction from vision in humans and machines. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1 (2021)
8. Bhowmik, A., Korzhenkov, D., Snoek, C.G.M., Habibian, A., Ghafoorian, M.: MoAlign: Motion-centric representation alignment for video diffusion models. In: The Fourteenth International Conference on Learning Representations (2026)
9. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
10. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22563–22575 (June 2023)
11. Bordes, F., Garrido, Q., Kao, J.T., Williams, A., Rabbat, M., Dupoux, E.: Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. arXiv preprint arXiv:2506.09849 (2025)

12. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7310–7320 (June 2024)
13. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
15. Ho, J., Jain, A.N., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
16. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., et al.: VBench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21807–21818 (June 2024). https://doi.org/10.1109/CVPR52733.2024.02060
17. Ji, S., Chen, X., Tao, X., Wan, P., Zhao, H.: Physmaster: Mastering physical representation for video generation via reinforcement learning. arXiv preprint arXiv:2510.13809 (2025)
18. Joseph, S., Garrido, Q., Balestriero, R., Kowal, M., Fel, T., Bakhtiari, S., Richards, B., Rabbat, M.: Interpreting physics in video world models. arXiv preprint arXiv:2602.07050 (2026)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Proceedings of the 2nd International Conference on Learning Representations (2014)
20. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
21. Li, X., He, X., Zhang, L., Wu, M., Li, X., Liu, Y.: A comprehensive survey on world models for embodied ai. arXiv preprint arXiv:2510.16732 (2025)
22. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., et al.: Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131 (2024)
23. Lin, M., Wang, X., Wang, Y., Wang, S., Dai, F., Ding, P., et al.: Exploring the evolution of physics cognition in video generation: A survey. arXiv preprint arXiv:2503.21765 (2025)
24. Liu, S., Ren, Z., Gupta, S., Wang, S.: PhysGen: Rigid-body physics-grounded image-to-video generation. In: European Conference on Computer Vision. Lecture Notes in Computer Science, vol. 15140, pp. 360–378. Springer (2024). https://doi.org/10.1007/978-3-031-73007-8_21
25. Ma, X., Wang, Y., Chen, X., Jia, G., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. Transactions on Machine Learning Research (2025)
26. Melnik, A., Ljubljanac, M., Lu, C., Yan, Q., Ren, W., Ritter, H.: Video diffusion models: A survey. Transactions on Machine Learning Research (2024)
27. Meng, C., Griesemer, S., Cao, D., Seo, S., Liu, Y.: When physics meets machine learning: A survey of physics-informed machine learning. Machine Learning for Computational Science and Engineering **1**, 20 (2025). https://doi.org/10.1007/s44379-025-00016-0

28. Motamed, S., Culp, L., Swersky, K., Jaini, P., Geirhos, R.: Do generative video models understand physical principles? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 948–958 (March 2026)
29. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., et al.: OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. In: The Thirteenth International Conference on Learning Representations (2025)
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (October 2023)
31. Peng, X., Zheng, Z., Shen, C., Young, T., Guo, X., Wang, B., et al.: Open-sora 2.0: Training a commercial-level video generation model in 200 k. arXiv preprint arXiv:2503.09642 (2025)
32. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Advances in Neural Information Processing Systems. vol. 36, pp. 53728–53741. Curran Associates, Inc. (2023)
33. Ressler-Antal, T., Fundel, F., Alaya, M.B., Baumann, S.A., Krause, F., Gui, M., Ommer, B.: DisMo: Disentangled motion representations for open-world motion transfer. In: Advances in Neural Information Processing Systems. vol. 38. Curran Associates, Inc. (2025)
34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
35. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
36. Singh, J., Leng, X., Wu, Z., Zheng, L., Zhang, R., Shechtman, E., Xie, S.: What matters for representation alignment: Global information or spatial structure? In: The Fourteenth International Conference on Learning Representations (2026)
37. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
38. Wang, J., Ma, A., Cao, K., Zheng, J., Feng, J., Zhang, Z., Pang, W., Liang, X.: WISA: World simulator assistant for physics-aware text-to-video generation. In: Advances in Neural Information Processing Systems. vol. 38. Curran Associates, Inc. (2025)
39. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14549–14560 (2023)
40. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
41. Wang, Y., Liu, X., Pang, W., Ma, L., Yuan, S., Debevec, P., Yu, N.: Survey of video diffusion models: Foundations, implementations, and applications. Transactions on Machine Learning Research (2025), survey Certification
42. Wang, Z., Hu, P., Wang, J., Zhang, T.J., Cheng, Y., Chen, L., et al.: ProPhy: Progressive physical alignment for dynamic world simulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14492–14501 (June 2026)

43. Xiang, K., Zhang, T.J., Huang, Y., He, J., Liu, Z., Tang, Y., et al.: Aligning perception, reasoning, modeling and interaction: A survey on physical ai. arXiv preprint arXiv:2510.04978 (2025)
44. Xie, T., Zhao, Y., Jiang, Y., Jiang, C.: PhysAnimator: Physics-guided generative cartoon animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10793–10804 (June 2025)
45. Xue, Q., Yin, X., Yang, B., Gao, W.: PhyT2V: LLM-guided iterative self-refinement for physics-grounded text-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 18826–18836 (June 2025). https://doi.org/10.1109/CVPR52734.2025.01754
46. Yang, X., Li, B., Zhang, Y., Yin, Z., Bai, L., Ma, L., 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 (ICCV). pp. 12360–12370 (October 2025)
47. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., et al.: CogVideoX: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2025)
48. Yin, Z., Chen, K., Bai, X., Jiang, R., Li, J., Li, H., Liu, J., Xiang, Y., Yu, J., Zhang, M.: A survey: Spatiotemporal consistency in video generation. ACM Computing Surveys **58**(12) (May 2026). https://doi.org/10.1145/3802588
49. Zhang, T., Yu, H.X., Wu, R., Feng, B.Y., Zheng, C., Snavely, N., Wu, J., Freeman, W.T.: PhysDreamer: Physics-based interaction with 3D objects via video generation. In: European Conference on Computer Vision. Lecture Notes in Computer Science, vol. 15060, pp. 388–406. Springer (2024). https://doi.org/10.1007/978-3-031-72627-9_22
50. Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., Cheng, Y.: VideoREPA: Learning physics for video generation through relational alignment with foundation models. In: Advances in Neural Information Processing Systems. vol. 38. Curran Associates, Inc. (2025)
51. Zhang, Z., Li, X., Liu, Y., Wang, Y., Chen, Y., Xue, T., Guo, S.: EGVD: Event-guided video diffusion model for physically realistic large-motion frame interpolation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 3982–3991 (October 2025)
52. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: MotionDirector: Motion customization of text-to-video diffusion models. In: European Conference on Computer Vision. Lecture Notes in Computer Science, vol. 15114, pp. 273–290. Springer (2024). https://doi.org/10.1007/978-3-031-72992-8_16
53. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)