DIPHY: DISTILLING TRANSFERABLE PHYSICAL KNOWLEDGE FOR PHYSICS-AWARE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent generative video models have achieved impressive visual quality and increasingly realistic video synthesis, yet they still struggle to generate videos that obey plausible physical dynamics. Existing approaches typically incorporate physical information through textual descriptions, trajectory constraints, or additional conditioning signals, whereas how physical knowledge can be effectively mined and transferred to video generation has not yet been sufficiently investigated. Facing this challenge, we propose **DiPhy**, which **Di**stills transferable **Phy**sical knowledge into a diffusion model for physically plausible video generation. Specifically, a JEPA-based teacher first mines physics-related cues from videos under semantic guidance. These cues are organized into a structured physical codebook that serves as a compact shared space for transferable physical knowledge, which is then distilled into a student module for text-only physical reasoning without video access. The inferred physical codes are subsequently injected into the diffusion backbone through a lightweight physical branch, providing explicit physical guidance during generation. Extensive experiments on physics-oriented video benchmarks demonstrate that our DiPhy improves physical consistency while preserving semantic alignment and overall video quality, with physical commonsense scores improving by approximately 9% and Joint performance by over 10% on VideoPhy2. Moreover, our DiPhy enables efficient video generation with relatively short inference time.

1 INTRODUCTION

Video generation models(Wang et al., 2025a; Yang et al., 2025b; Kong et al., 2024; Zheng et al., 2024; Chen et al., 2024) have recently achieved remarkable progress in visual quality and realism, demonstrating strong capabilities to model complex dynamic scenes. However, these models still struggle to faithfully adhere to physical laws, often resulting in implausible motion, object interactions, and temporal dynamics(Kang et al., 2025; Motamed et al., 2026; Meng et al., 2025b). Such physical inconsistencies undermine their reliability as world simulators(Qin et al., 2025) and limit their ability in broader applications.

Recently, many studies have explored to improve the physical consistency of generated videos. Several methods(Wang et al., 2025b; Yuan et al., 2026; Liu et al., 2024) explicitly formulate physical states or dynamics, providing strong physical constraints and interpretability. However, they often rely on predefined variables and scenario-specific assumptions, which may limit their scalability and generalization to complex real-world settings. Another line of work(Lv et al., 2024; Xue et al., 2025; Wang et al., 2026; Pathak et al., 2026) leverage physical priors encoded in pretrained language models or vision-language models to derive physical guidance from textual prompts, structured plans, or intermediate generations. More recently, methods such as (Zhang et al., 2025; Cheng et al., 2026) exploit video representations extracted from foundation models, like VideoMAE(Wang et al., 2023) and V-JEPA(Mur-Labadia et al., 2026), and leverage these messages as references or alignment targets to enhance the physical plausibility of generated videos. As illustrated in Fig. 1, existing approaches mainly rely on explicit physical simulation or semantic priors from language models, while recent representation-related methods suggest that pretrained video models contain rich physical cues beyond appearance and semantics. This observation naturally raises a further question:

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.

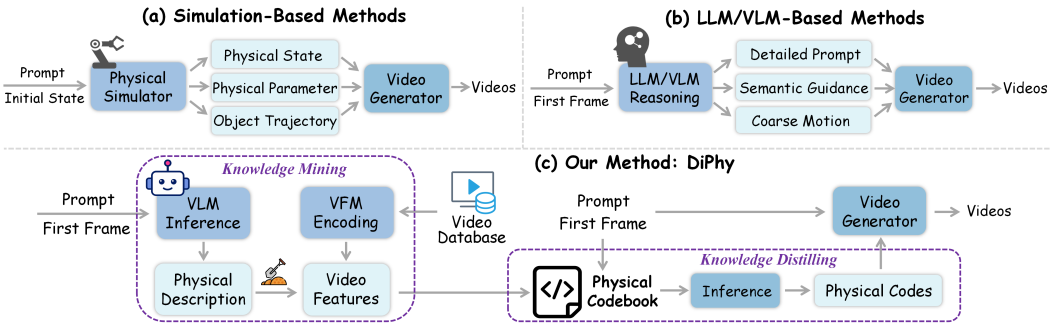

Figure 1: Different paradigms for physics-aware video generation. Simulation-based methods rely on specific physical simulators, while LLM/VLM-based methods utilize semantic priors from LLM/VLMs. DiPhy instead mines transferable knowledge from videos and distills it through a shared physical codebook to video generation.

Can such latent physical cues be explicitly mined, distilled into transferable physical knowledge, and injected into diffusion models to improve physically plausible video generation?

To address this question, we propose **DiPhy**, which **Di**stills transferable physical knowledge for **Phy**sics-informed video generation. In the first stage, a JEPA-based teacher module is constructed to mine physics-relevant information from videos with semantic guidance provided by a vision language model. The valuable information is then further organized into a compact physical codebook, which provides a shared space storing transferable physical knowledge. In the second stage, the knowledge encoded in the codebook is distilled into a student module, enabling it to predict scenario-specific physical codes directly from textual conditions without requiring video observations at inference time. In the final stage, the inferred physical codes are injected into a diffusion backbone through a lightweight physical branch, providing physics-aware guidance throughout video synthesis. Meanwhile, we employ a semantic adapter to preserve semantic alignment and maintain overall quality while improving the physical consistency of generated videos.

Our contribution can be summarized as follows:

1. We propose DiPhy, a physics-informed video generation framework that mines latent physical cues from JEPA-based video representations and distills them into transferable physical knowledge to guide physically plausible video synthesis.
2. To effectively organize the mined physical knowledge, we develop a physical codebook providing a compact shared space for teacher-to-student distillation. It bridges video-based physical knowledge acquisition and text-conditioned generation, allowing the model to exploit learned physical priors without video observations at inference.
3. Extensive experiments on multiple physics-oriented benchmarks demonstrate the effectiveness of our DiPhy, showing improved physical consistency while preserving semantic alignment and overall video quality, with physical commonsense scores improving by approximately **9%** and Joint performance by over **10%** on VideoPhy2.

2 Related Work

2.1 Video Generation Models

Diffusion models(Ho et al., 2020; Song et al., 2021; Rombach et al., 2022) have become a dominant paradigm for generative modeling, with latent diffusion substantially improving the efficiency of high-resolution synthesis. Moreover, transformers have been increasingly adopted as diffusion backbones(Esser et al., 2021; Peebles & Xie, 2023; Esser et al., 2024), improving model scalability and representation capacity. Building upon these advances, recent video generation models(Yang et al., 2025b; Wang et al., 2025a; Zheng et al., 2024) have achieved remarkable progress in visual quality, appearance realism, and overall generation fidelity. However, visual realism does not neces-

sarily imply physical plausibility(Meng et al., 2025a; Bansal et al., 2025; Kang et al., 2025). Current video generation models may still produce unrealistic object motion, interactions, deformation, or temporal evolution(Motamed et al., 2026; Meng et al., 2025b), revealing a remaining gap between visually convincing generation and physically consistent world modeling.

2.2 Physics-Aware Video Generation

To address this limitation, recent research has explored incorporating physical knowledge into video generation. Some methods(Liu et al., 2024; Wang et al., 2025b; Yuan et al., 2026; Foo et al., 2026) **explicitly formulate physical states and governing dynamics**, imposing explicit physical constraints on the generation process and enabling direct control with clear physical interpretation. For example, NewtonGen(Yuan et al., 2026) defines a 9-dimensional latent physical state to capture translation, rotation, deformation, and other physical behaviors, and models diverse dynamics using neural ordinary differential equations (ODEs). PSIVG(Foo et al., 2026) integrates a physical simulator into the generation loop by reconstructing scene geometry and object meshes, and simulating physically consistent trajectories to guide video generation. However, such methods often rely on predefined variables and specific modeling assumptions, which may limit their extension to diverse scenarios and complex physical phenomena.

Another line of work(Yang et al., 2025a; Xue et al., 2025; Pathak et al., 2026; Wang et al., 2026) **leverages physical priors from pretrained language or vision language models** to guide video generation. VLIPP(Yang et al., 2025a) employs a vision-language model as a motion planner to predict coarse trajectories or state changes that approximate real-world dynamics and provide explicit motion guidance for video generation. PhyT2V(Xue et al., 2025) combines language-model guidance with iterative prompt refinement to improve physical dynamics and temporal consistency. PhysVid(Pathak et al., 2026) constructs temporally local and counterfactual prompts with vision language and language models, injecting them as additional conditioning beyond the standard text-to-video pathway. CoECT(Wang et al., 2026) models physics-informed video generation as a sequence of causally connected events, using chain-of-thought reasoning to decompose physical phenomena into elementary event units and summarize them into a coherent prompt. However, these approaches mainly derive physical knowledge through semantic reasoning or intermediate guidance signals, which may only indirectly capture physical dynamics and may not always provide reliable physical guidance.

More recently, several studies(Garcia et al., 2025; Cheng et al., 2026; Zhang et al., 2025) have explored **learning physical information directly from video observations**. Learning Physics From Video(Garcia et al., 2025) shows that physical parameters of continuous dynamical systems can be estimated from videos, suggesting that visual observations contain informative physical cues. PhysRAG(Cheng et al., 2026) constructs a curated physical video database and employs learnable queries to derive physical priors from retrieved reference videos, which are then injected into a video diffusion model through a retrieval augmented generation framework. VideoREPA(Zhang et al., 2025) learns physics for video generation through relational representation alignment, using a Token Relation Distillation loss to transfer physics-relevant information from video foundation models to generative models. These advances motivate us to explicitly mine transferable physical information from video observations and enable the valuable knowledge to be distilled for physics-aware video generation.

3 Method

3.1 Overall Framework

Given a text prompt describing one or more physical phenomena in the real world, our goal is to generate a corresponding video with physically plausible dynamics. As illustrated in Fig. 2, we propose DiPhy, a physics-informed video generation framework that distills transferable physical knowledge into a diffusion model to generate videos that obey physical laws. More specifically, a teacher module is first constructed to mine physics-relevant information from videos under semantic guidance, and organize this information into a compact physical codebook. The transferable knowledge stored in the physical codebook is then distilled into a student module, enabling the student to infer the corresponding physical codes without access to videos. Finally, the inferred codes are

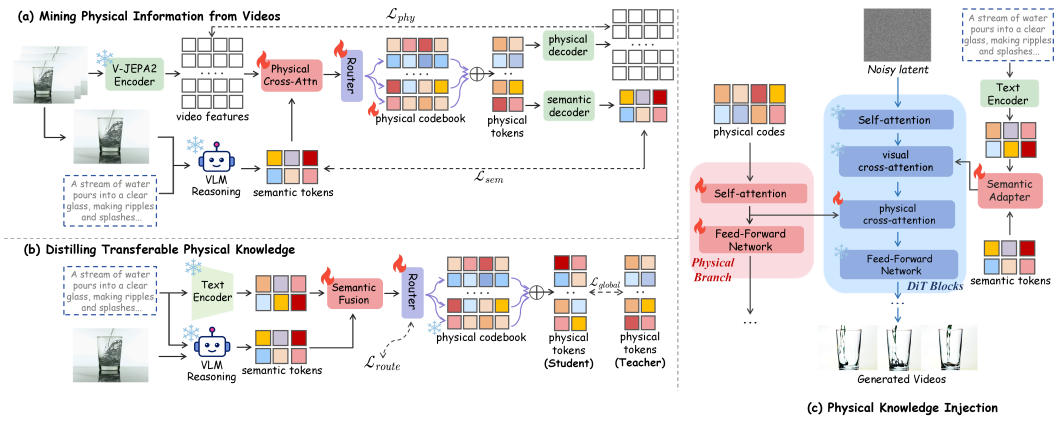

Figure 2: The overall framework of DiPhy. (a) We construct a teacher module to mine physics-related information from videos and develop a physical codebook that stores transferable knowledge. (b) The transferable physical knowledge is distilled into a student module, enabling the student to infer codes corresponding to specific physical phenomena. (c) The inferred knowledge is incorporated into the diffusion backbone through a physical branch, guiding the model toward physically plausible video generation.

injected into the diffusion backbone through a lightweight physical branch, allowing the model to generate physics-informed videos and maintaining a relatively low inference time.

3.2 Mining Physical Information from Videos

As discussed in Subsection 2.2, motion patterns and object interactions can be inferred from video observations, indicating that video features contain valuable cues about underlying physical behavior. Based on this observation, we introduce a teacher module to mine physics-relevant information from video features under semantic guidance. Specifically, given a video V, we first extract its spatiotemporal features $\mathbf{F}_v \in \mathbb{R}^{N_v \times d_v}$ using a frozen video foundation model, where N_v denotes the number of video tokens. Meanwhile, the first frame I_1 and text prompt T are fed into a vision language model to obtain semantic tokens $\mathbf{F}_s \in \mathbb{R}^{N_s \times d_s}$, which can provide semantic guidance about the described phenomenon and assist the teacher to identify spatiotemporal video tokens related to the physical description. We therefore employ cross-attention between F_s and F_v, to aggregate spatiotemporal features $\mathbf{H}$, which are more relevant to the corresponding physical scenario.

$$\mathbf{H}_T = \mathrm{Softmax}\left(\frac{(\mathbf{F}_s\mathbf{W}_Q)(\mathbf{F}_v\mathbf{W}_K)^\top}{\sqrt{d}}\right)(\mathbf{F}_v\mathbf{W}_V), \tag{1}$$

The aggregated features $\mathbf{H}_T$ are then fed into a learnable router, which maps them to a compact physical codebook $\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^{N_c}$. The codebook serves as a compact shared space for organizing physics-relevant information shared across different videos. By selecting a small subset of codes for each sample, the router encourages the mined information to be represented by a shared and reusable set of physical codes. Specifically, the router predicts a code distribution $\mathbf{R}_T = \mathrm{Softmax}(g(\mathbf{H}_T))$, and selects Top-$K$ physical codes to construct the teacher physical tokens $\mathbf{P}_T = \sum_{k=1}^{K} \mathbf{R}_{T,i_k}\mathbf{c}_{i_k}$, which is decoded by a physical decoder and a semantic decoder to align with the pooled video feature $\bar{\mathbf{F}}_v$ and semantic feature $\bar{\mathbf{F}}_s$ respectively. The overall training objective can be formulated as:

$$\begin{aligned}\mathcal{L}_{\text{teacher}} &= \lambda_{\text{phy}}\mathcal{L}_{\text{phy}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} \\ &= \lambda_{\text{phy}}\left[\text{SmoothL1}\left(D_{\text{phy}}(\mathbf{P}_T), \bar{\mathbf{F}}_v\right) + 1 - \cos\left(D_{\text{phy}}(\mathbf{P}_T), \bar{\mathbf{F}}_v\right)\right] \\ &\quad + \lambda_{\text{sem}}\left[\text{SmoothL1}\left(D_{\text{sem}}(\mathbf{P}_T), \bar{\mathbf{F}}_s\right) + 1 - \cos\left(D_{\text{sem}}(\mathbf{P}_T), \bar{\mathbf{F}}_s\right)\right].\end{aligned} \tag{2}$$

3.3 DISTILLING TRANSFERABLE PHYSICAL KNOWLEDGE

As discussed in Subsection 3.2, we mine physics-related information from videos and construct a compact codebook $\mathcal{C}$ storing transferable physical knowledge. The physical knowledge is then distilled into a student module, enabling the student to infer physical codes corresponding to the textual condition without access to videos. Specifically, the text prompt is firstly encoded into text features F_t, and then incorporated with semantic tokens F_s obtained from a vision language model through Transformer-based fusion, which can complement the textual condition with more specific and comprehensive descriptions of the underlying physical phenomena:

$$\mathbf{H}_S = \mathcal{F}_{\text{fusion}}(F_t, F_s) \tag{3}$$

where F_s is firstly introduced in Subsection 3.1. The fused features $\mathbf{H}_S$ are further summarized through attention pooling into a compact query q_S, which is used by the router to predict student's routing distribution $\mathbf{R}_S = \text{Softmax}(g_S(q_S))$ and select the Top-K codes from the physical codebook $\mathcal{C}$. Notably, the codebook is inherited from teacher module and kept frozen during the training of student module, effectively preserving the physical knowledge mined from videos. The selected weighted codes form the student physical token set $\widetilde{P}_S = [\mathbf{R}_{S,i_k} c_{i_k}]_{k=1}^{K} \in \mathbb{R}^{K \times d_p}$, which are subsequently fed into diffusion backbone for video generation. To facilitate knowledge distillation during training, the selected codes are further aggregated into a global student physical token $\mathbf{P}_S$, which is aligned with the teacher physical token $\mathbf{P}_T$. Meanwhile, We further distill the teacher routing behavior by aligning the complete routing distributions. The student's overall objective is:

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{global}} + \lambda_{\text{route}} \mathcal{L}_{\text{route}} = 1 - \cos(\mathbf{P}_T, \mathbf{P}_S) + \lambda_{\text{route}} D_{\text{KL}}(\mathbf{R}_T \,\|\, \mathbf{R}_S) \tag{4}$$

3.4 PHYSICAL KNOWLEDGE INJECTION

The inferred physical tokens $\widetilde{P}_S \in \mathbb{R}^{K \times d_p}$ are incorporated into the diffusion backbone through a lightweight physical branch, which consists of several physical blocks. The physical blocks are initialized from $P^{(0)} = \widetilde{P}_S$, and each block is updated through self-attention and a feed-forward network with a GELU activation and a residual connection:$\hat{P}^{(j)} = P^{(j-1)} + \text{SelfAttn}(P^{(j-1)})$, $P^{(j)} = \hat{P}^{(j)} + \text{FFN}(\hat{P}^{(j)}) = \hat{P}^{(j)} + W_2\,\text{GELU}(W_1 \hat{P}^{(j)})$ The resulting physical tokens are then injected into several early DiT blocks through cross-attention. To reduce the memory and computational overhead of cross-attention over long video token sequences, we perform the interaction in a low-dimensional attention space:

$$X_l' = X_l + \alpha_l W_o\,\text{CrossAttn}(Q_l, K_j, V_j), \tag{5}$$

where α_l controls the injection strength at the l-th interaction block, and Q_l, K_j, V_j are projected to a low dimension:$Q_l = X_l W_Q, K_j = P^{(j)} W_K, V_j = P^{(j)} W_V$. Meanwhile, introducing physical tokens as an additional condition may perturb the original text-conditioned semantics. Therefore, We employ a lightweight semantic adapter to incorporate the same VLM-derived semantic tokens F_s used in Subsection 3.3 into the original text features, through low-dimensional residual cross-attention similar to physical branch:

$$F_t' = F_t + \lambda_{\text{sem}} W_o\,\text{CrossAttn}\left(F_t W_Q^s, F_s W_K^s, F_s W_V^s\right) \tag{6}$$

Where, λ_{sem} is a learnable scalar gate controlling the semantic residual. The pretrained diffusion backbone remains frozen during training, while the physical branch, physical injection modules, and semantic adapter are optimized. In this way, the physical branch injects explicit physical guidance, while the semantic adapter reinforces the original semantic condition, with the frozen backbone retaining its pretrained visual generation prior, enabling the model to generate physically plausible and high-quality videos.

4 EXPERIMENT

4.1 IMPLEMENTATION DETAILS

Training We build DiPhy upon two video generation models, CogVideoX-2B(Yang et al., 2025b) and Wan2.1-1.3B(Wang et al., 2025a). We randomly select 10k videos from WISA-80K(Wang et al., 2025c) for training, which is a large dataset containing more than 80k videos relevant to various physical phenomena, employ V-JEPA2(Mur-Labadia et al., 2026) to obtain video features and leverage the Qwen3-VL-8B-Instruct(Bai et al., 2025) to derive semantic tokens. For the training objectives in Eqs. 2 and 4, we use fixed loss weights throughout training. Specifically, we set $\lambda_{\text{phy}} = 1$ and $\lambda_{\text{sem}} = 0.25$ for the teacher loss, and $\lambda_{\text{route}} = 0.1$ for the student loss. More implementation details for training are provided in **Supplementary Material**.

Evaluation We evaluate our method on **VideoPhy2**(Bansal et al., 2026), a physics-oriented benchmark that contains 600 prompts providing coverage over various physical phenomena in the real world. This benchmark measures generated videos through Physical Commonsence(PC), Semantic Adherence(SA) and Joint pass rate(Joint). The VideoPhy2 employs automatic evaluator assigning a score from 1 to 5 for each aspect, scores of 4 or higher are regarded as PC = 1 or SA = 1, and only both PC = 1 and SA = 1 are regarded as Joint=1, implying a physically plausible and semantic aligned video. Since DiPhy requires a first-frame input, we generate the first frame offline from the corresponding prompt using FLUX (Labs, 2024), rather than using a real video frame. This avoids introducing additional visual information and ensures a fair comparison with text-to-video methods. Moreover, we evaluate the overall quaity of generated videos on **VBench**(Huang et al., 2024), which evaluate the overall generation quality from multiple quality-related dimensions.

4.2 QUANTITATIVE COMPARISONS

Table 1: Results on the VideoPhy2 benchmark. We include open-source T2V models and physics-informed methods. For physics-informed methods, the best and second-best performances are highlighted in **bold** and underlined, respectively.

	Method	PC	SA	Joint	Inference Time
Open-source T2V Models	Zeroscope	54.3	21.5	17.5	–
	OpenSora V1.1	60.5	24.8	23.0	–
	VideoCrafter2	58.3	26.5	20.8	–
Physics-Informed Methods	CogVideoX-2B	56.2	27.8	23.3	335s
	+ VideoREPA	<u>62.7</u>	27.0	23.2	360s
	+ DiPhy	61.8	**28.7**	**25.8**	365s
	Wan2.1-1.3B	58.0	26.3	21.7	870s
	+ PhysVid	62.5	25.7	22.8	960s
	+ DiPhy	**63.3**	<u>28.3</u>	<u>24.7</u>	930s

We compare our DiPhy with several open-source T2V models including Zeroscope(cerspense, 2023), OpenSora V1.1(Zheng et al., 2024), VideoCrafter2(Chen et al., 2024), CogVideoX-2B(Yang et al., 2025b) and Wan2.1-1.3B(Wang et al., 2025a) , and Physics-Informed Methods such as VideoREPA(Zhang et al., 2025) and PhysVid(Pathak et al., 2026). As illustrated in Table 1, DiPhy achieves excellent performance among the compared physics-informed methods. Compared with the corresponding baselines, DiPhy improves PC by approximately 9% and increases the Joint score by 10.7% and 13.8% respectively. These results demonstrate that our DiPhy effectively enhances the physical consistency of generated videos while maintaining strong semantic adherence. On V100 GPUs, with video features and semantic tokens pre-computed, DiPhy only adds a lightweight physical branch and semantic adapter during denoising, resulting in modest inference overhead over the base T2V model. Moreover, as shown in Table 2, DiPhy maintains competitive overall video quality on VBench. In particular, it achieves higher scores in *Motion Smoothness* and *Dynamic Degree* than the corresponding baselines and other physics-informed methods, reflecting smoother motion and

more plausible dynamic behavior. Meanwhile, its competitive performance across other quality-related dimensions indicates that such physical guidance does not compromise the generation capability of pretrained diffusion models. Consequently, DiPhy improves the physical plausibility of generated videos while preserving high visual fidelity.

Table 2: Results on VBench quality score. For each backbone, the best performance among the compared methods is highlighted in **bold**.

Method	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	**Quality Score**
CogVideoX-2B	90.2	**93.8**	95.9	97.6	80.1	45.0	59.2	77.8
+ VideoREPA	**91.8**	93.3	95.5	97.6	78.3	**46.6**	62.3	78.4
+ DiPhy	91.4	93.5	**96.0**	**97.8**	**80.5**	44.7	**63.1**	**78.7**
Wan2.1-1.3B	90.0	93.4	95.8	97.2	80.6	**49.5**	66.6	79.3
+ PhysVid	91.6	**93.6**	**96.2**	97.5	71.3	48.6	**69.5**	79.6
+ DiPhy	**91.7**	92.3	95.8	**97.6**	**81.0**	48.9	68.3	**80.0**

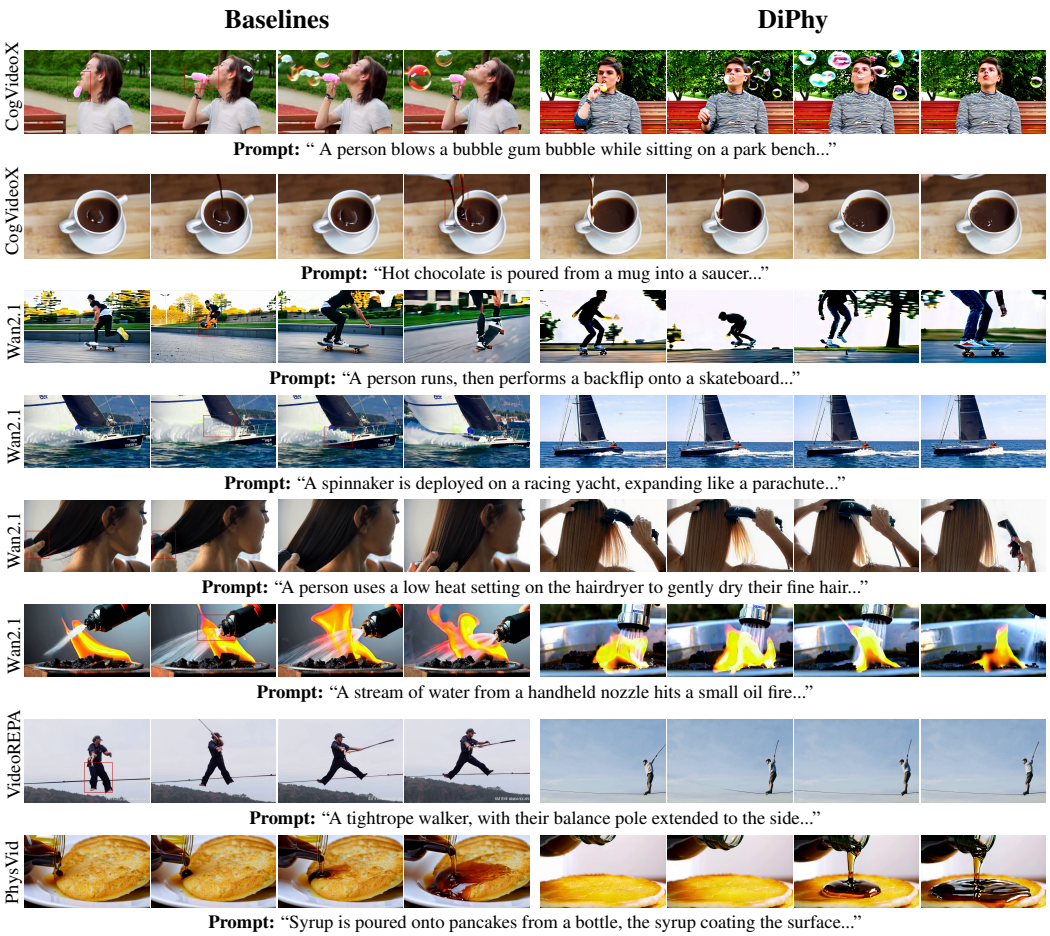

Figure 3: Qualitative comparisons among open-source T2V models, physics-informed methods and DiPhy across diverse physical scenarios. For each example, four frames are uniformly sampled from the generated video.

4.3 Qualitative Comparisons

As shown in Fig. 3, DiPhy better captures physically meaningful object interactions and motion dynamics. In the bubble blowing example, the baseline places the bubble at an implausible location, while DiPhy models the local interaction more accurately. In the liquid pouring example, the baseline generates two pouring streams simultaneously, whereas DiPhy produces a more coherent single stream. Meanwhile, DiPhy also shows advantages across diverse physical phenomena, including rigid body dynamics, fluid dynamics, and gas dynamics. In the tightrope-walking example, the baseline produces an implausible suspended body posture, while DiPhy better preserves the expected kinematic relationship with the rope. In the water–fire interaction, DiPhy generates a more appropriate contact region between the water stream and the flame. For the sailing example, DiPhy also produces a more reasonable wave position relative to the moving boat. Overall, these results suggest that DiPhy improves the physical plausibility of both local interactions and broader dynamic behaviors.

4.4 Ablation Study

Table 3: Ablation study on different components of DiPhy. The best and second-best results are highlighted in **bold** and underlined, respectively.

Settings	PC	SA	Joint
baseline	56.2	27.8	23.3
w/o physical branch	57.3	**29.0**	24.2
w/o semantic adapter	<u>61.2</u>	27.7	<u>25.0</u>
w/o teacher module	58.7	28.3	24.7
DiPhy	**61.8**	<u>28.7</u>	**25.8**

We conduct ablation studies on VideoPhy2(Bansal et al., 2026) Benchmark, using CogVideoX-2B(Yang et al., 2025b) as the backbone and focusing on two aspects: component contributions and physical knowledge injection strategies. We first evaluate the semantic adapter, physical branch, and teacher module. For the ablation of teacher module, we remove the teacher routing and global physical supervision, randomly initialize the physical codebook and supervise it only with pooled semantic features, while keeping the downstream generation pipeline unchanged. As shown in Table 3, removing the teacher module degrades PC clearly, confirming that the physical knowledge mined from videos provides useful guidance for improving the physical consistency of generated videos. Meanwhile, Removing the semantic adapter decreases SA, indicating that it helps provide global semantic constraints during generation. In addition, removing the physical branch leads to a clear drop in PC yet a slightly increase in SA, suggesting that knowledge incorporated through physical branch can enhance the physical plausibility of videos, while the injection may introduce a mild trade-off with semantic adherence.

Table 4: Ablation on different strategies for injecting physical knowledge.

Settings	PC	SA	Joint
global injection	60.7	28.0	25.2
physical cross-attention	**61.8**	**28.7**	**25.8**



(a) global injection (b) physical cross-attention

Figure 4: Ablation on different strategies for injecting physical knowledge.

We further compare two injection strategies: physical cross-attention and global injection. In the latter, the selected physical codes are aggregated into a single global token, projected to the DiT hidden dimension, and broadcast as a shared residual. Unlike global injection, physical cross-attention explicitly models interactions between latent tokens and individual physical codes, leading to improved physical plausibility (Table 4) and accurate local interactions (Fig. 4).

4.5 PHYSICAL CODEBOOK

The physical codebook is a key component of DiPhy for organizing transferable physical knowledge. We extensively analyze it from two perspectives: the intrinsic properties of the codebook and its usage patterns across different physical phenomena.

Codebook Properties. We first examine the diversity and utilization of the inferred physical codes. The average pairwise cosine similarity is 0.280, with a maximum of 0.552 and a minimum of 0.017, indicating clear differences among codes while preserving certain similarities that may related to shared physical characteristics. We further analyze code selection frequency using Top-1 and Top-8 coverage, which count the codes selected as the highest-probability code and those appearing in the Top-8 set respectively. We also compute the normalized entropy of the selection distribution as selective diversity, ranging from 0 to 1, where higher values indicate more diverse usage. The Top-1 and Top-8 selections cover **48** and **49** codes respectively, with a selective diversity of **0.89**. These results indicate broad codebook utilization without severe selection collapse, while several frequently selected codes may encode physical knowledge shared across various scenarios.

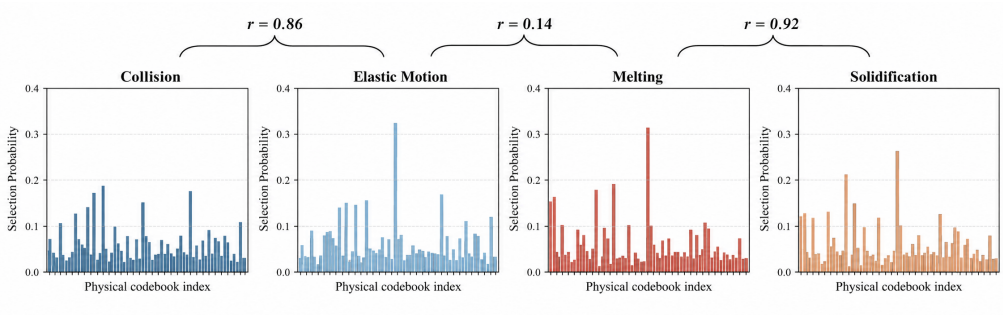

Figure 5: Code selection distributions for representative physical categories. r represents the Pearson correlation coefficient calculated between different distributions.

Relationship between Codes and Physical Phenomena. We further investigate whether different physical phenomena exhibit structured code selection patterns. For each physical category, we compute the average code selection distribution and measure the *Pearson correlation* between categories. As shown in Fig. 5, physics-related categories tend to have more similar selection patterns. For example, collision and elastic motion achieve a correlation of $r = 0.86$, while melting and solidification reach $r = 0.92$. In contrast, elastic motion and melting show a much lower correlation of $r = 0.14$. These results indicate that the codebook can capture physical regularities across different phenomena, where related physical categories tend to have similar code selection patterns while distinct categories show more differentiated routing behaviors. This suggests that the codes are not selected arbitrarily, but instead organize transferable physical knowledge according to the underlying characteristics of different scenarios.

5 CONCLUSION

In this paper, we propose DiPhy, which mines and distills transferable physical knowledge for physics-aware video generation. A JEPA-based teacher module firstly mines physics-related cues from videos and organizes them into a codebook. These cues is then distilled into a student module through the shared codebook, enabling the student to infer physical codes for different phenomena without access to videos. The inferred physical codes are subsequently injected into the diffusion backbone through a physical branch to improve the physical plausibility of generated videos. Extensive experiments on physics-oriented benchmarks demonstrate that DiPhy consistently improves physical plausibility while maintaining semantic adherence and overall video quality.

Limitations. Despite these improvements, DiPhy mainly learns physical knowledge from video features without explicitly incorporating physical laws. Moreover, the inferred physical codes lack deterministic physical definitions, which limits the interpretability of codebook. Future work will integrate physical laws with data-driven learning for more interpretable modeling of object interactions and complex physical activities.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research methodology and experimental design, code development and debugging, the writing and revision of the manuscript. We also used generative AI tools for literature search, language polishing, and the preparation and refinement of scientific figures. All AI-assisted code, analyses, and written content were reviewed and verified by the authors, and experimental results were obtained from our own implementations and evaluations. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on improving the physical plausibility of generated videos and does not involve human-subject studies or the collection of personal data. Our experiments are conducted using existing video datasets, benchmarks, and pretrained open-source T2V models. As with other video generation technologies, improved realism may potentially be misused for generating misleading synthetic content. We therefore encourage responsible use of the proposed method and appropriate safeguards when deploying video generation systems. When misused without labels and watermarks, it may produce fake videos and lead to the spread of misinformation.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the DiPhy architecture, training objectives, and physical knowledge injection mechanism in Section 3, with implementation and evaluation settings reported in Section 4. Additional training details, and experimental analyses are provided in the appendix. The supplementary material includes representative generated videos and code for the main components and core framework to facilitate reproduction of our results.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 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, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=9D2QvO1uWj.

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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HA8KSQW7SO.

cerspense. Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023. Hugging Face model repository.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 7310–7320. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00698. URL https://doi.org/10.1109/CVPR52733.2024.00698.

Kexu Cheng, Zicheng Liu, Mingju Gao, Chunhe Song, and Hao Tang. Physrag: Enhancing physics-awareness in video generation via retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2606.26916.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, June 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Lin Geng Foo, Mark He Huang, Alexandros Lattas, Stylianos Moschoglou, Thabo Beeler, and Christian Theobalt. Physical simulator in-the-loop video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4301–4311, June 2026.

Alejandro Castañeda Garcia, Jan Warchocki, Jan van Gemert, Daan Brinks, and Nergis Tomen. Learning physics from video: Unsupervised physical parameter estimation for continuous dynamical systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27924–27933, June 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 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.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/kang25g.html.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *CoRR*, abs/2412.03603, 2024. doi: 10.48550/ARXIV.2412.03603. URL https://doi.org/10.48550/arXiv.2412.03603.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

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

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 (CVPR) Workshops*, pp. 1430–1440, June 2024.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation, 2025a. URL https://openreview.net/forum?id=6rMHcLWxl4.

Siwei Meng, Yawei Luo, and Ping Liu. Grounding creativity in physics: A brief survey of physical priors in AIGC. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pp. 10593–10602. ijcai.org, 2025b. doi: 10.24963/IJCAI.2025/1176. URL https://doi.org/10.24963/ijcai.2025/1176.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2026, Tucson, AZ, USA, March 6-10, 2026*, pp. 948–958. IEEE, 2026. doi: 10.1109/WACV61042.2026.00099. URL https://doi.org/10.1109/WACV61042.2026.00099.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Saurabh Pathak, Elahe Arani, Mykola Pechenizkiy, and Bahram Zonooz. Physvid: Physics aware local conditioning for generative video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41847–41858, June 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, and Ruimao Zhang. Worldsimbench: Towards video generation models as world simulators. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/qin25f.html.

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 (CVPR)*, pp. 10684–10695, June 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*, abs/2503.20314, 2025a. doi: 10.48550/ARXIV.2503.20314. URL https://doi.org/10.48550/arXiv.2503.20314.

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 Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/f53fd88a4340063ecd258c0ae9948b40-Abstract-Conference.html.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. Wisa: World simulator assistant for physics-aware text-to-video generation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 5388–5416. Curran Associates, Inc., 2025c. doi: 10.52202/085713-0191. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0856bc553d3e3b9827e5140d0ad3bf8d-Paper-Conference.pdf.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560, June 2023.

Zixuan Wang, Yixin Hu, Haolan Wang, Feng Chen, Yan Liu, Wen Li, and Yinjie Lei. Chain of event-centric causal thought for physically plausible video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38122–38131, June 2026.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 18826–18836. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01754. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xue_PhyT2V_LLM-Guided_Iterative_Self-Refinement_for_Physics-Grounded_Text-to-Video_Generation_CVPR_2025_paper.html.

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, and Xu Jia. 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 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=LQzN6TRFg9.

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H. Chan. Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rJ6N6sunaU.

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 Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/b1d4973f5d21708abef3cd6f17d842c8-Abstract-Conference.html.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *CoRR*, abs/2412.20404, 2024. doi: 10.48550/ARXIV.2412.20404. URL https://doi.org/10.48550/arXiv.2412.20404.

Appendix

A Appendix Introduction

This appendix provides additional details, results and analyses for our DiPhy, mainly including details of physical knowledge mining(Section B), knowledge distilling(Section C), knowledge injection(Section D).

B More Details of the Knowledge Mining

B.1 Different Designs of the Teacher Module

As discussed in Section 3, the physical knowledge from teacher module should not only be informative, but also transferable to the text-conditioned student. During the development of the teacher module, we modify the construction of physical codebook and supervision for the teacher physical tokens, eventually summarizing a suitable design for knowledge mining and distilling. Variant 1 utilizes only video features to construct the codebook and supervise the physical tokens. Based on Variant 1, Variant 2 derives semantic tokens from a vision language model, which are then fused with video features through cross-attention. Compared to Variant 2, a semantic decoder is further added in the final DiPhy, providing the teacher with both physical and semantic reconstruction objective.

Table 5: Comparison of different teacher module designs.

Settings	Code Diversity	Pairwise Similarity	Code Usage	Physical Reconstruction	Semantic Reconstruction	Teacher-Student Alignment
Variant 1	0.91	0.33	52/64	0.93	–	0.77
Variant 2	0.89	0.30	49/64	0.93	–	0.79
DiPhy	0.89	0.28	49/64	0.92	0.98	**0.93**

We evaluate different teacher designs using several metrics. *Code Diversity* measures the normalized entropy of code selection and ranges from 0 to 1, where a higher value indicates more diversity. *Pairwise Similarity* is the average cosine similarity between physical codes, reflecting their overall similarity. *Code Usage* reports the number of physical codes activated by Top-8 routing over the dataset. *Physical Reconstruction* and *Semantic Reconstruction* measure the reconstruction quality of video and semantic features respectively, with values ranging from 0 to 1 and higher values indicating better reconstruction. *Teacher-Student Alignment* measures the alignment between physical tokens produced by the teacher and student, where a higher value indicates better knowledge transfer.

Table 6: Comparison of module designs on VideoPhy2. The best performance is highlighted in **bold**.

Settings	PC	SA	Joint
Baseline (CogVideoX-2B)	56.2	27.8	23.3
Variant 1	59.0	27.5	22.7
Variant 2	59.7	28.2	24.3
DiPhy	**61.8**	**28.7**	**25.8**

Analyses of Different Designs. As illustrated in Table 5, all variants can obtain well-behaved codebooks with comparable code diversity, pairwise similarity and utilization. Every variant can effectively reconstruct their corresponding targets. However, their transferability to the student differs significantly. The first two variants achieve only about 77% and 79% on teacher-student alignment

respectively, whereas DiPhy improves this value to about 91%. These results suggest that code diversity and reconstruction quality alone are insufficient for designing the teacher module. Moreover, these results may suggest a gap between semantic information and physical knowledge. While the model can readily infer plausible visual content from text, capturing the underlying physical relationships may remain challenging. This motivates our use of semantic guidance and semantic reconstruction in the teacher module, where semantic information and video-derived knowledge jointly shape the codebook. Such a design makes the mined knowledge more accessible from text conditions, facilitating the student to predict corresponding physical information without videos. Meanwhile, as shown in Table 6, DiPhy achieves higher PC, SA, and Joint scores than the other variants, supporting the effectiveness of our final teacher design.

B.2 Different loss weights

We further explore the influence of the physical and semantic reconstruction losses in the teacher module. Given physical knowledge injected into the video generation model, the physical reconstruction objective should be assigned a larger weight. As shown in Table 7, increase in the physical reconstruction weight achieves a relatively higher score on PC, indicating the incorporated knowledge is effective and valuable. Meanwhile, increasing the semantic reconstruction weight generally improves teacher-student alignment, but does not necessarily improve physical consistency. For example, $\lambda_{\text{phy}} : \lambda_{\text{sem}} = 0.1 : 1$ achieves the highest 0.96, while PC decreases to 59.3 and only $40/64$ codes are activated. This suggests that excessive semantic reconstruction may make knowledge transfer easier, but also make the knowledge less valuable for physical guidance. We therefore choose $\lambda_{\text{phy}} = 1$ and $\lambda_{\text{sem}} = 0.25$, which can mine and distill both **valuable** and **transferable** knowledge, to assist diffusion models for physically-plausible video generation.

Table 7: Effect of different loss weights in the teacher module. The best and second-best performance is highlighted in **bold** and underlined.

Setting	Code Usage	Teacher-Student Alignment	PC	SA	Joint
Baseline (CogVideoX-2B)	–	–	56.2	27.8	23.3
$\lambda_{\text{phy}} : \lambda_{\text{sem}} = 1 : 1$	48/64	0.91	60.7	28.2	25.0
$\lambda_{\text{phy}} : \lambda_{\text{sem}} = 0.5 : 1$	46/64	0.94	59.8	**29.0**	24.7
$\lambda_{\text{phy}} : \lambda_{\text{sem}} = 0.1 : 1$	40/64	0.96	59.3	28.3	24.2
$\lambda_{\text{phy}} : \lambda_{\text{sem}} = 1 : 0.5$	50/64	0.91	<u>61.7</u>	27.8	<u>25.5</u>
$\lambda_{\text{phy}} : \lambda_{\text{sem}} = 1 : 0.1$	61/64	0.88	61.3	<u>28.7</u>	25.3
$\lambda_{\text{phy}} : \lambda_{\text{sem}} = 1 : 0.25$	49/64	0.93	**61.8**	<u>28.7</u>	**25.8**

B.3 Semantic Guidance from VLMs

Given the text prompt and the first frame, we employ a vision language model to obtain corresponding physical description, assisting the model to focus on visible objects, motion, interactions, and physically relevant relationships. The complete prompt template is shown in Table 8. Notably, when performing evaluation on VideoPhy2(Bansal et al., 2026), we generate the first frame from the prompt using FLUX(Labs, 2024). However, prompts from VideoPhy2(Bansal et al., 2026) often describe dynamic events, and the generated first frame may not perfectly reflect the intended initial state. Therefore, we slightly modify the prompt template to prioritize textual information and reduce the influence of visual errors.

Table 8: Prompt template used for generating physical descriptions with the VLM.

Physics Description Prompt

You are a physics-aware video reasoning expert. Given the first frame of a video and its original text caption, infer the most plausible physical evolution of the scene. Your goal is not to simply restate the caption, but to explicitly reason about the physical process that is likely to occur.

Analyze the scene from the following aspects:

1. Identify the main physical objects and environmental elements that actively participate in the process, while excluding irrelevant background content.
2. Describe the initial physical state, including relative positions, contact or support relations, visible motion tendencies, and material or structural properties only when supported by the image or caption.
3. Identify the most relevant physical laws, principles, or mechanisms. Possible examples include gravity, inertia, momentum conservation, collision and contact mechanics, friction, elasticity, rigid-body motion, fluid mechanics, buoyancy, pressure, deformation, energy transfer, thermodynamics, phase transitions, magnetism, and optics.
4. Explain how these physical mechanisms affect the main objects and how the objects interact with one another.
5. Infer an ordered sequence of the major physical events expected to occur. Focus on causal event progression rather than exact frame indices or precise trajectories.
6. For each event, describe the expected physical state changes, such as motion direction, acceleration tendency, contact state, deformation, flow, dispersion, energy transfer, temperature-related changes, or phase transitions.
7. Explain the causal relationships between consecutive events and why one event leads to the next according to physical principles.
8. State the global physical constraints that should remain valid throughout the video, including temporal continuity, plausible interaction responses, and conservation-related behavior when applicable.
9. Do not invent precise numerical quantities unless explicitly provided, and do not introduce unsupported objects, materials, or interactions.

Original video caption: {caption}

Produce a compact but information-rich physical description in 2–4 sentences. The final description should preserve:

- the main physical objects,
- the relevant physical laws or mechanisms,
- the ordered physical event sequence,
- the major state changes,
- and the causal relationships between events.

Output only the final physical description. Do not output headings, bullet points, JSON, or intermediate reasoning.

C More Details of the Knowledge Distilling

Training Details. The student module takes pre-computed semantic tokens and text features as input, which are generated offline for all training samples. The physical codebook learned by the teacher module is kept frozen during the training of student, while the Transformer-based semantic fusion module and the student's router are optimized. Specifically, the physical codebook is configured with 64 codes, each of dimension 512. The Top-K physical codes are selected according to the routing probabilities with $K = 8$, and the weighted codes are formed as $\widetilde{P}_S \in \mathbb{R}^{8\times512}$ for subsequent knowledge incorporation. These weighted codes are further summed into a global physical token $\mathbf{P}_S \in \mathbb{R}^{512}$, which is aligned with the teacher physical tokens $\mathbf{P}_T \in \mathbb{R}^{512}$ through the global supervision loss. We set the learning rate as 2e-5 and a weight decay of 0.01, training the model on 8 V100 GPUs with FP16 precision.

D MORE DETAILS OF THE KNOWLEDGE INJECTION

Training Details. During training, a lightweight physical branch is constructed to interact with the early stages of the DiT backbone, specifically blocks 0, 1, and 2, where the latent representation is still being formed and remains sensitive to additional physical guidance. The physical branch contains three blocks corresponding to these interaction stages, so that progressively updated physical tokens can be injected into different DiT blocks. Notably, the self-attention layers across these physical blocks share the same parameters, while the physical cross-attention modules responsible for injecting physical information are independently parameterized. Maintaining this lightweight design can reduce additional parameter overhead while allowing different DiT blocks to receive individual physical information from different physical states.

Inference Details. During inference, the early DiT features are still close to noisy latent states, overly strong physical injection may disturb visual content formation. Therefore, in addition to the injection coefficient α_l defined in Eq. (5), which controls the injection strength at the l-th interaction block, we further introduce stage-wise scaling factors of 0.5, 1.0, and 1.5 for the three interaction stages respectively, balancing physical guidance with visual generation and overall video quality.