PhysRAG: Enhancing Physics-Awareness in Video Generation via Retrieval-Augmented Generation

Kexu Cheng[1,2*], Zicheng Liu[1*], Mingju Gao[1*]
Chunhe Song[2†], and Hao Tang[1†]

[1] School of Computer Science, Peking University, Beijing, China
[2] Institute of AI for Industries, Chinese Academy of Sciences, Nanjing, China
[*]Equal contribution. [†]Corresponding authors: bjdxtanghao@gmail.com.

Abstract. Developing physically aware video generation models remains a significant challenge due to the difficulty in capturing diverse physical phenomena, such as thermal dynamics, mechanics, and optics. In this work, we introduce PhysRAG, a novel pipeline that enhances physical awareness in video generation through Retrieval-Augmented Generation (RAG). To address the issue of limited high-quality data, we design a two-stage data filtering pipeline based on the WISA-80K dataset, resulting in a curated set of 7K high-quality videos for training. Furthermore, we construct a physical video database and develop a mechanism to inject physical knowledge into a video diffusion model using learnable queries. Our method achieves state-of-the-art performance in both visual quality and physical rule compliance, surpassing existing models in benchmarks such as PhyGenBench and VBench. We conduct extensive ablation studies to validate the effectiveness of our key components, including the data filtering pipeline, RAG mechanism, and method for physical information extraction. To facilitate future research, our code, data, and models are prepared for release at https://github.com/sediment1024/PhysRAG.

Keywords: Video Generation · Retrieval-Augmented Generation · Physical Aware Generation

1 Introduction

With the advancement of computing power and the growing scale of training data, text-to-video (T2V) generation has made significant strides [20, 33, 34, 41, 56, 67, 71]. While the generated visuals appear highly realistic, these models often face difficulties in adhering to physical laws due to the lack of physical modeling. Accurately modeling physical dynamics in generated videos thus remains a challenging and underexplored problem [3, 32, 38, 45]. Overcoming the limitation of physical awareness in video generation models will enable them to better capture real-world phenomena, which can benefit downstream applications such as embodied AI [7, 29, 30], autonomous driving [63, 66], and more.

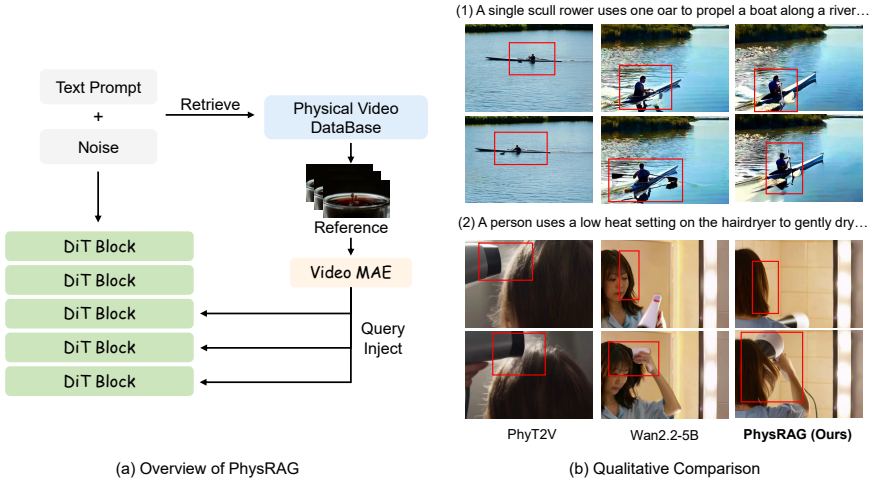

Fig. 1: Overview and Qualitative Results of PhysRAG. (a) Architecture of the proposed retrieval-augmented framework. We retrieve relevant physical videos from a database, extract features using Video MAE, and inject these explicit physical priors into the DiT blocks via learnable query injection. (b) Qualitative comparison with baseline models (Wan2.2-5B and PhyT2V). Red boxes highlight regions with complex physical interactions. Compared to the baselines, PhysRAG generates significantly more realistic physical dynamics, such as accurate water-oar interaction (top) and natural hair movement under airflow (bottom).

A significant body of prior work has focused on addressing physical dynamics in video generation, which can generally be categorized into explicit and implicit approaches. Explicit methods [5,12,22,50,54,57,65,69,75–77,80,81] rely on deterministic physical simulators or mathematical models to simulate motion directly. While these methods guarantee strict adherence to physical laws, they often struggle with generalization in complex, open-world environments that lie outside their predefined constraints (for example, simulating thermal effects is difficult). On the other hand, implicit methods [11, 15, 23, 31, 36, 39, 40, 59, 62, 79] are predominantly data-driven, focusing on designing physics-aware latent spaces or models to facilitate physically plausible generation. However, these models face a significant challenge in controllability; it remains notoriously difficult to inject precise physical guidance or enforce strict physical constraints within learned implicit representations.

Consider how humans acquire physical knowledge. We perceive and predict events in the physical world by observing scenes in our daily lives, even without consciously understanding the underlying laws of physics. We instinctively anticipate outcomes based on our past experiences with similar physical principles. For instance, we intuitively know that a ball will fall to the ground if we release it, simply by observing freefall. In this process, humans explicitly retrieve relevant scenes from memory (similar to the concept of RAG [35]) and implicitly extract the physical laws governing these scenes, applying them to new scenar-

ios. Similarly, we propose that physical awareness in video generation can be enhanced by explicitly selecting videos that adhere to the same physical laws we want to generate and implicitly injecting the physical principles from these videos into the generation process via learnable queries.

Challenge. Despite the clear motivation, developing a robust video generation model with physical awareness remains challenging. A primary obstacle is the scarcity of high-quality data that captures diverse physical laws, such as thermal dynamics, mechanics, optics, and others. While some datasets aim to train physical-aware video generation models [6, 21, 36, 55, 59, 74], these methods either suffer from a lack of high-quality data [59] or are limited to a small range of physical phenomena, such as freefall [36]. To address this limitation and better train our model, we have designed a two-stage data filtering pipeline. This pipeline first applies a coarse pre-filter to assess the quality of captions and remove those that are irrelevant to the physical phenomenon, retaining the top 10% of the videos. In the second stage, a fine grounded filter ensures the consistency between video frames and the associated prompts. This approach enables the selection of a high-quality subset that covers a wider range of physical phenomena, providing a more reliable foundation for training our physical-aware video generation model.

Therefore, we propose **PhysRAG**, a pipeline that enhances physical awareness in video generation through Retrieval-Augmented Generation. To address the issue of scarce high-quality data, we introduce a two-stage data filtering pipeline based on WISA-80K [59], which filters out videos with irrelevant captions and low-quality content, resulting in a curated dataset of 7K high-quality videos for training. To make the video diffusion model aware of physics, as illustrated in Figure 1 (a), we manually construct a physical video database and extract videos relevant to the given prompts. To inject the physical knowledge from these videos, we design a mechanism that leverages learnable queries to extract and transfer the relevant physical knowledge. As shown in Figure 1 (b), our method enables better physical consistency and awareness in video generation.

In our experiment, we evaluate our method on two benchmarks: PhyGenBench [44] and VBench [27]. Across multiple evaluation settings, our method shows improvements in both visual quality and physical rule compliance, even outperforming closely related models, such as Pika [48] and Kling [34]. Additionally, we conduct ablation studies to validate the effectiveness of our proposed modules, including the data filtering pipeline, the RAG mechanism, and the method for extracting and injecting physical information into the video diffusion model. Our contributions can be summarized as follows:

- We introduce **PhysRAG**, a novel pipeline that integrates Retrieval-Augmented Generation for enhancing physical awareness in video generation.
- We design a two-stage data filtering pipeline that significantly improves the quality and relevance of training data, yielding a high-quality curated dataset of 7K videos for training.

- We construct a manual physical video database and develop a mechanism that leverages learnable queries to inject physical knowledge into the video diffusion model.
- Our method achieves state-of-the-art performance in physical-aware video generation, demonstrating improvements in both visual quality and rule compliance over existing models.

2 Related Work

2.1 Text-to-Video Diffusion Models

Diffusion-based frameworks have emerged as the dominant paradigm for text-to-video (T2V) generation due to their superior visual quality and scalability. Foundational works established the baseline for temporally coherent synthesis [24,25,52], while subsequent latent diffusion approaches further enhanced fidelity and training efficiency through efficient adaptation and tuning strategies [8,14,64,68]. Recently, the field has converged on large-scale Transformer-based backbones (Video DiT) to improve temporal modeling, as exemplified by systems like Lumiere, Open-Sora, Vchitect-2.0, HunyuanVideo, Wan, CogVideoX, and Step-Video-T2V [4, 19, 33, 43, 56, 71, 82]. Despite their rapid progress toward world-simulator capabilities, maintaining physically plausible dynamics and consistent motion over time remains a significant challenge [3, 10, 44, 77].

2.2 Physics-Aware Video Generation

Achieving physically plausible dynamics remains a significant challenge in video generation. To address the frequent violation of basic physical laws, existing literature can be broadly categorized into explicit and implicit modeling paradigms. Explicit approaches incorporate physical knowledge through deterministic simulators, trajectory guidance, or structured mathematical constraints [22,23,50,54, 57, 59, 65, 69, 70, 76, 77, 80]. While ensuring strict adherence to predefined rules, these methods often struggle to generalize across complex, open-world scenarios where phenomena are difficult to explicitly simulate. Conversely, implicit methods are predominantly data-driven, leveraging techniques such as direct preference optimization (DPO), reinforcement learning, and feature alignment to align models with physically plausible distributions [11,15,36,39,40,62,79,80]. Although these alignment-based strategies offer broader generalization, they face significant challenges in controllability; because physical priors are typically encoded as scalar rewards or prompt-side reasoning, injecting precise physical guidance into the learned implicit representations remains notoriously difficult. Despite these advances, existing approaches lack mechanisms to exploit explicit spatiotemporal video references as physical priors. Our method addresses this gap through a retrieval-augmented conditioning design that **explicitly** retrieves relevant physical exemplars and **implicitly** injects their encoded dynamics into the latent denoising process via learnable queries.

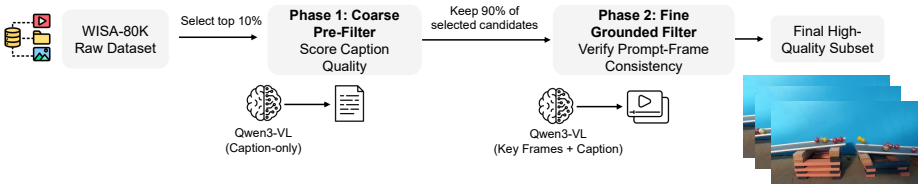

Fig. 2: Two-stage preprocessing for WISA-80K. We first use Qwen3-VL (caption-only) to select the top 10% candidates, then use Qwen3-VL (key frames + caption) to verify prompt–frame consistency and retain 90%, yielding a high-quality subset (∼7K videos).

2.3 Retrieval-Augmented Conditioning for Video Generation

Retrieval-augmented generation (RAG) has evolved from text-only external-memory frameworks [35] to multimodal systems like MuRAG and RA-CM3 that jointly retrieve images and text, highlighting the broader utility of non-parametric priors for generation and reasoning [16, 72]. In visual generation, retrieval-conditioned paradigms have further established that external exemplars can improve fidelity and diversity, as demonstrated by retrieval-augmented diffusion and text-to-image generation methods such as RDM and Re-Imagen [9, 17].

In the realm of video generation, conditioning strategies are primarily distinguished by the modality of retrieved references and the mechanism of feature injection. Approaches include cross-attention conditioning, parameter-efficient control modules like ControlNet and IP-Adapter [46, 49, 73, 78], and query-based fusion mechanisms such as Q-Former [1, 37]. Within the video domain, retrieval encoders like VideoCLIP-XL and VideoMAE V2, alongside recent frameworks including RAGME, MotionRAG, and Plug-and-Play Memory, promise exemplar-guided motion transfer [47, 53, 58, 61, 83]. However, these methods typically target global style or general dynamics. For instance, while Plug-and-Play Memory [53] modulates DiTs via frequency-filtered memory tokens, it retrieves from general-domain databases and lacks dedicated mechanisms to extract explicit physical properties, rendering it sub-optimal for physics-grounded generation. Consequently, the query-based **spatiotemporal** injection of explicit physical priors into the latent denoising process of a Video DiT backbone remains under-explored. Addressing this gap, we introduce a novel retrieval-augmented conditioning design that retrieves physics-relevant video exemplars and injects their encoded spatiotemporal features directly into Video DiT blocks via learnable-query cross-attention and feature-space conditioning.

3 Physics-Aware Video Data Construction

Constructing a reliable physics-aware video dataset is a foundational prerequisite for training trustworthy generative models. Although the existing WISA-80K dataset [59] provides a large collection of internet-sourced videos with coarse-grained cleaning, it cannot be used directly because of persistent data quality issues, as illustrated in Table 6. Such noise can hinder the model's ability to

learn underlying physical laws, making a subsequent cleaning and filtering stage essential.

To address this issue, we introduce a **two-stage** data preprocessing pipeline to extract a reliable subset from the raw WISA-80K dataset, as illustrated in Fig. 2. The pipeline progressively filters out low-quality and physically inconsistent samples, yielding a cleaner and more trustworthy dataset for subsequent model training.

Phase 1: Coarse Pre-Filter. In the first phase, filtering is performed at the text level only. We use Qwen3-VL-4B [2] to score the quality and physics relevance of the text descriptions, and retain the top 10% of candidates. This text-only screening efficiently removes a large fraction of irrelevant samples at low cost, but it inevitably introduces false positives, i.e., samples whose captions appear physically meaningful while the corresponding video content is weakly related or mismatched.

Phase 2: Fine Grounded Filter. To further remove text–video misaligned samples, we introduce a multimodal grounding verification step. For each candidate video retained after Phase 1, we uniformly sample key frames and feed them, together with the corresponding text prompt, into Qwen3-VL-4B [2] to assess text–visual consistency. Samples that fail this grounding check are discarded, while the remaining samples are kept for the final subset.

Design Rationale. We adopt a two-phase filtering strategy to balance data quality and computational efficiency. Phase 1 uses low-cost text-only screening to remove a large portion of irrelevant samples, while Phase 2 applies a stricter multimodal grounding check to eliminate false positives from the first stage. In Phase 2, we use uniformly sampled key frames instead of the full video sequence as a cost-effective proxy since processing all frames jointly with the caption using Qwen3-VL-4B is prohibitively expensive at the dataset scale (approximately 2000 gpu hours on NVIDIA H20 devices). The resulting high-quality subset contains approximately **7K** videos from the original **80K** videos in WISA-80K.

4 The Proposed Method

4.1 Overview

Figure 3 presents the overall pipeline of our framework. The key idea is to adopt a RAG paradigm that explicitly guides video synthesis with real-world physical dynamics by retrieving reference videos with similar underlying physics from a curated video database. To this end, we first construct an offline Physical Video Database (Section 4.2), where videos are manually curated and categorized by physical phenomena (e.g., collision, combustion, and explosion). During inference, given a text prompt, we retrieve the most physically relevant video from this database using the VideoCLIP-XL model [58]. The retrieved video is then encoded by a pre-trained VideoMAE V2 encoder. The resulting latent representations are injected into the original diffusion transformer blocks via our **Query Inject** module (Section 4.3), in which learnable query tokens attend to the encoded video latents to capture physically relevant dynamics.

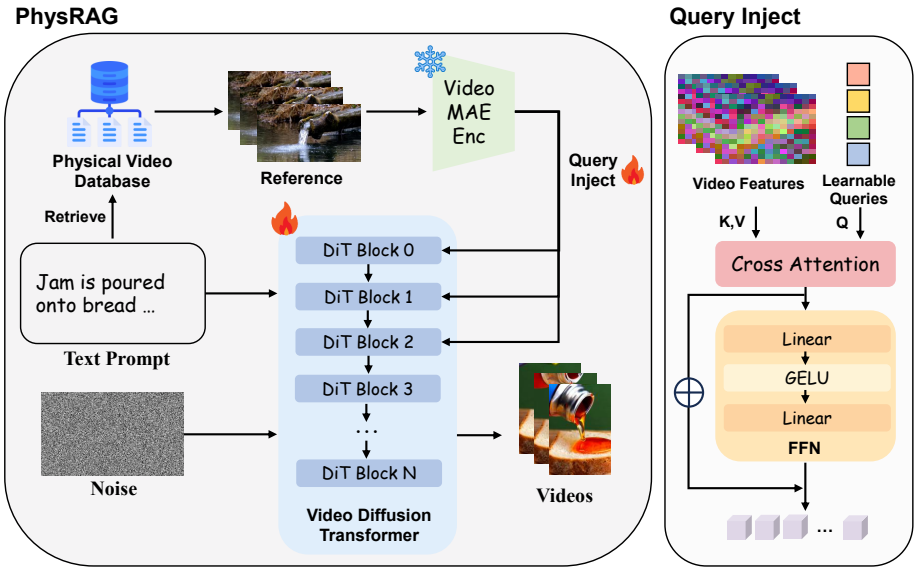

Fig. 3: Overview of our retrieval-augmented video diffusion framework. A Video Diffusion Transformer (DiT) generates videos from text and noise while retrieving relevant physical videos from a database. Retrieved features are encoded by Vide MAE and injected into DiT blocks through the Query Inject module.

4.2 PhysRAG Database Construction

The effectiveness of our retrieval-augmented generation framework critically depends on the quality and organization of the physical video database. As discussed earlier, videos that depict the same physical phenomenon often share similar underlying dynamics. Therefore, constructing a structured database with representative examples of diverse physical phenomena is essential for providing reliable physical priors to guide the generative model.

To this end, we build the **PhysRAG Database**, a manually curated physical video corpus designed for retrieval at inference time. We first identify common real-world physical phenomena that frequently appear in open-domain video generation scenarios, and systematically group them into **17 categories**. For each category, we manually collect and curate **10 high-quality videos** that clearly demonstrate the corresponding physical characteristics, resulting in a database of **170 videos** in total. Each video is then assigned to its category and stored under the corresponding directory, forming a simple yet effective hierarchical organization for retrieval. This structured design offers two key benefits. First, it improves retrieval reliability by constraining candidate videos to physically meaningful exemplars. Second, it provides diverse but consistent references within each category, which helps the model extract transferable physical priors instead of overfitting to a single video instance.

4.3 Physical Prior Injection

Motivation. In order to explicitly incorporate **physical priors** into the video generation process, our primary goal is to integrate physical knowledge derived from real-world videos into the Video Diffusion Transformer (DiT). While the pre-trained VideoMAE V2 [61] encoder is proficient at extracting rich spatio-temporal features from these retrieved videos, directly embedding such dense latent representations into the DiT presents a significant challenge, as shown in Table 3. This is because our aim is to selectively preserve only the relevant **physical dynamics** while excluding any irrelevant or extraneous information. If these non-relevant features are incorporated, they could introduce undesirable biases, ultimately degrading the quality of the generated content.

To address this challenge, we introduce the **Query Inject** module, as depicted in Figure 3. This module employs a set of learnable queries, which act as an information bottleneck. These queries selectively attend to the video latents, extracting the physical priors that are critical to the task. After the physical priors are distilled, they are injected into the DiT blocks by concatenating them with the original video features. To ensure consistency in the token dimensions of the DiT model, we then use a projector to map the concatenated features back to the original token count. This approach effectively integrates the extracted physical information, enriching the model's output with essential physical dynamics while avoiding the introduction of irrelevant features or biases. As a result, the generated videos exhibit significantly improved quality and realism.

Formulation. Given a retrieved physical reference video, we extract its offline VideoMAE-V2 features as $F_v \in \mathbb{R}^{L \times C_v}$, where L is the token sequence length and C_v is the feature dimension. We further maintain a set of learnable query tokens $Q \in \mathbb{R}^{N \times C_h}$, where $N = 128$ is the number of query tokens and C_h is the hidden dimension of the query adapter.

The query tokens interact with the physical features through cross-attention:

$$H_{ca} = \mathrm{CrossAttn}(Q, F_v, F_v) \in \mathbb{R}^{N \times C_h}, \tag{1}$$

where Q serves as the query, while F_v serves as both the key and value. This operation adaptively aggregates the physical dynamics most relevant to the current generation process.

We then refine the attended features using a standard feed-forward network with a GELU activation and a residual connection:

$$H_{ffn} = H_{ca} + \mathrm{FFN}(H_{ca}), \tag{2}$$

where

$$\mathrm{FFN}(x) = W_2\,\mathrm{GELU}(W_1 x). \tag{3}$$

Afterward, the refined features are projected to a compact physical-prior token space:

$$H_{out} = H_{ffn} W_o \in \mathbb{R}^{N \times C_p}, \tag{4}$$

where C_p denotes the physical-prior token dimension, and W_1, W_2, and W_o are learnable projection matrices.

Let the intermediate DiT tokens be $H_{\mathrm{DiT}} \in \mathbb{R}^{T \times C_d}$, where T is the DiT token length and C_d is the DiT hidden dimension. Before injection, we align the physical-prior tokens to the DiT token space using a lightweight alignment operator $\mathcal{A}(\cdot)$, which projects the channel dimension and adjusts the token sequence to length T:

$$\tilde{H}_p = \mathcal{A}(H_{out}) \in \mathbb{R}^{T \times C_d}. \tag{5}$$

We then concatenate the aligned physical-prior tokens with the DiT tokens along the sequence dimension and apply a lightweight fusion projector:

$$H_{\mathrm{fuse}} = \phi\Big(\mathrm{Concat}(H_{\mathrm{DiT}}, \tilde{H}_p)\Big) \in \mathbb{R}^{T \times C_d}. \tag{6}$$

Finally, we inject the fused physical prior through a gated residual connection:

$$H'_{\mathrm{DiT}} = H_{\mathrm{DiT}} + \alpha H_{\mathrm{fuse}}, \tag{7}$$

where α is a learnable scalar gate. The resulting H'_{DiT} is fed into subsequent DiT blocks, enabling the denoising process to exploit retrieved physical priors while preserving the original DiT interface.

5 Experiments and Results

5.1 Settings

Data. Our training data is sourced from WISA-80K [59]. Following the filtering pipeline outlined in Sec. 3, we curate a subset of approximately 7K high-quality videos. This subset is used for all SFT and RAG-based training experiments below.

Evaluation. We evaluate our method on two benchmarks, following [53]

- **PhyGenBench** [44] for evaluating physical commonsense, assessed according to the official protocol outlined in [44]. The evaluation is conducted using GPT-4o [28] as the MLLM judge. PhyGenBench is designed to evaluate the physical reasoning capabilities of models by testing their ability to handle a range of physical scenarios, including mechanics, optics, and other physical phenomena. The protocol measures how well a model predicts and adheres to physical commonsense principles in these contexts, providing a comprehensive benchmark for assessing the integration of physical knowledge into generative models.
- **VBench** [27], following the per-dimension protocol on a pre-registered subset that balances temporal and perceptual/semantic factors. In this evaluation, we focus on low-level metrics, such as color, style, object classification, and their average, as well as high-level metrics, including subject consistency, spatial relationships, multi-object handling, and their average.

Baselines. We select baselines, including both open-source and closed-source models, to comprehensively evaluate our method. The closed-source commercial

Table 1: Quantitative comparison with state-of-the-art video generation models on PhyGenBench. Our proposed PhysRAG achieves the highest overall average score of 0.58, demonstrating superior physical correctness and outperforming both leading closed-source commercial systems and open-source baselines.

Source	Method	Size	Mechanics(↑)	Optics(↑)	Thermal(↑)	Material(↑)	Average(↑)
Closed	Pika [48]	–	0.35	0.56	0.43	0.39	0.44
	Gen-3 [51]	–	0.45	0.57	0.49	<u>0.51</u>	0.51
	Kling [34]	–	0.45	0.58	<u>0.50</u>	0.40	0.49
Open	CogVideoX [71]	2B	0.38	0.43	0.34	0.39	0.39
	CogVideoX [71]	5B	0.39	0.55	0.40	0.42	0.45
	Open-Sora V1.2 [82]	1.1B	0.43	0.50	0.44	0.37	0.44
	Lavie [64]	860M	0.30	0.44	0.38	0.32	0.36
	Vchitect 2.0 [19]	2B	0.41	0.56	0.44	0.37	0.45
	DiT-Mem [53]	5B	0.56	**0.74**	0.48	0.47	<u>0.56</u>
	Wan 2.2 [56]	5B	<u>0.58</u>	0.60	<u>0.50</u>	0.48	0.54
	Wan 2.2 + PhysRAG	5B	**0.59**	<u>0.66</u>	**0.54**	**0.53**	**0.58**

models include Pika [48], Gen-3 [51], and Kling [34]. For open-source models, we compare against prominent text-to-video frameworks such as CogVideoX [71], Open-Sora [82], Lavie [64], Vchitect 2.0 [19], ModelScope [60], and VideoCrafter [13]. Furthermore, we benchmark against DiT-Mem [53], a recent method that modulates DiTs via frequency-filtered memory tokens. However, since it retrieves from general-domain databases, it lacks explicit physical priors compared to our approach. Finally, Wan 2.2 (5B) [56] serves as our direct baseline to highlight the improvements introduced by our PhysRAG integration.

Implementation Details. We implement our method using PyTorch, building upon the foundational text-to-video (T2V) model Wan2.2-5B [56]. The model is fine-tuned on four NVIDIA H20 GPUs for two days with an effective batch size of 128 (a micro-batch size of 16 per GPU and 2 gradient accumulation steps). We apply a learning rate of 1×10^{-6} and a weight decay of 0.01, training for 20 epochs at a video resolution of $49 \times 704 \times 480$. We jointly train the Wan backbone and learnable queries and inject the learnable queries into the 0,1,2 layers of the DiT blocks. For the PhysRAG module, reference videos are processed using VideoCLIP-XL [58] to extract features, while FAISS [18] is employed for the efficient retrieval of video latents that are similar to the input user prompt. Prior to fine-tuning, we pre-encode videos into VAE latents and prompts into T5 embeddings, which reduces preprocessing overhead and improves training throughput. To optimize memory efficiency, we utilize BF16 mixed precision, gradient checkpointing, and DeepSpeed ZeRO-3 with CPU offloading. We employ the AdamW [42] optimizer with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and save checkpoints every 400 steps.

5.2 Quantitative Comparisons

We compare our method with state-of-the-art methods across two comprehensive benchmarks, as shown in Tab. 1 (PhyGenBench) and Tab. 2 (VBench). Over-

Table 2: Quantitative evaluation on VBench. Integrating PhysRAG into Wan 2.2 consistently improves both low-level and high-level metrics, achieving the best Low-Avg (65.48%) and High-Avg (82.88%) among compared methods.

Method	Low-level Metrics				High-level Metrics			
	Color	Style	Obj. Class	Low-Avg	Subj. Consist.	Spatial Rel.	Multi-Obj.	High-Avg
OpenSora V1.1 [82]	74.56%	**23.50**%	86.76%	61.61%	92.35%	52.47%	40.97%	61.93%
ModelScope [60]	81.72%	23.39%	82.25%	62.45%	89.87%	33.68%	38.98%	54.18%
VideoCrafter [13]	78.84%	21.57%	**87.34**%	62.58%	86.24%	36.74%	25.93%	49.64%
CogVideo [26]	79.57%	22.01%	73.40%	58.33%	92.19%	18.24%	18.11%	42.85%
DiT-Mem [53]	93.67%	21.22%	80.00%	64.96%	95.67%	78.42%	74.38%	82.82%
Wan 2.2 [56]	85.92%	21.26%	79.12%	62.10%	95.51%	**78.91**%	69.12%	81.18%
Wan 2.2 + PhysRAG	**93.95**%	22.00%	80.50%	**65.48**%	**96.31**%	77.33%	**75.00**%	**82.88**%

all, PhysRAG demonstrates superior capability in generating physics-grounded videos while maintaining high visual quality.

Results on PhyGenBench. Tab. 1 evaluates the physical correctness of generated videos. Our method achieves the highest average score of **0.58**, setting a new state-of-the-art that outperforms both leading open-source models and closed-source commercial systems like Kling (0.49). Compared to the foundational Wan-2.2 baseline, PhysRAG brings substantial performance gains in complex physical simulations, with notable absolute improvements of **0.07** in Thermal and **0.08** in Material metrics. Furthermore, our approach surpasses the recent memory-guided method DiT-Mem (0.56 average), validating that our explicit physical feature retrieval is more effective for physics-grounded generation than general-domain memory tokens.

Results on VBench. Tab. 2 reports the evaluation of general video generation quality and semantic alignment. Integrating PhysRAG into Wan 2.2 consistently improves performance across multiple metrics. In particular, the Low-Avg increases from **62.10%** to **65.48%**, achieving the best result among all compared models, while the High-Avg improves from 81.18% to 82.88%. Our method also achieves the highest Color score of 93.95% and strong subject consistency and spatial relation performance. These results demonstrate that injecting explicit physical priors into the DiT backbone enhances generation fidelity without compromising semantic alignment.

5.3 Qualitative Comparisons

As illustrated in Figure 4, we qualitatively evaluate the physical consistency of our method against Wan2.2-5B and PhyT2V across different physical scenarios from PhyGenBench [44]. In the first example depicting a robotic arm etching a circuit board, the baseline models fail to comprehend the underlying mechanical interaction, resulting in largely static generations where the tool fails to execute the continuous motion, such as lowering and lifting, required for the etching process. PhysRAG, however, accurately captures the fine-grained, dynamic spatial interaction necessary for this precise mechanical task. In the second example involving fluid dripping onto a surface, the baselines exhibit a severe lack of understanding regarding fluid dynamics and temporal causality. Specifically,

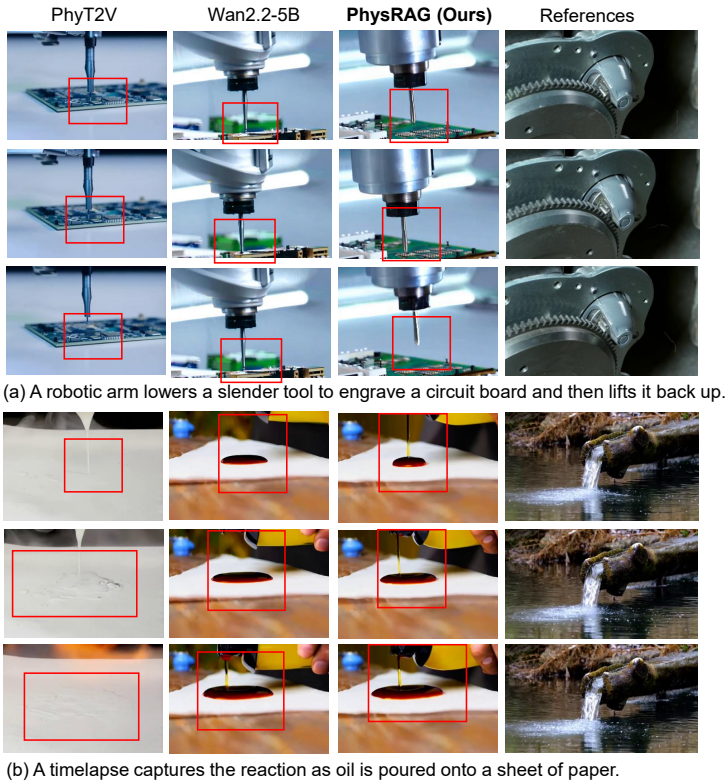

Fig. 4: Qualitative comparison of physical consistency across different scenarios from PhyGenBench [44]. PhysRAG outperforms Wan2.2-5B and PhyT2V by accurately modeling physical interactions. In the robotic arm etching scenario, PhysRAG captures the continuous motion required for the task, while baseline models fail to do so. In the fluid dripping example, PhysRAG correctly preserves fluid dynamics and material properties, whereas the baselines exhibit non-causal effects and incorrect behavior.

Wan2.2-5B generates a non-causal physical effect where the pool of liquid on the surface expands before the falling droplets actually make contact. PhyT2V suffers from a similar unprompted volume increase while also completely failing to render the correct material properties of the fluid. In contrast, PhysRAG successfully preserves the accurate material characteristics and enforces strict causal physics, ensuring that the liquid volume only accumulates sequentially after the fluid physically reaches the surface.

5.4 Ablation Study

Ablation on Injection Methods. Table 3 presents a comparison of different mechanisms for integrating retrieved features into the Video DiT backbone, evaluating their impact on the generation of physically realistic videos. The **concat** method, which directly concatenates the retrieved features with the video

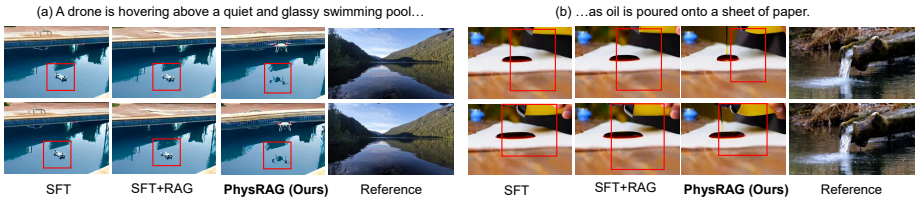

Fig. 5: Ablation on PhyGenBench [44]. We compare PhysRAG (ours) with SFT and SFT+RAG on (a) drone-over-pool reflection and (b) oil pouring. In (a), SFT and SFT+RAG miss the water-surface reflection, while PhysRAG preserves it. In (b), SFT and SFT+RAG show non-causal liquid growth without coherent downward flow; PhysRAG produces temporally consistent pouring and accumulation.

Table 3: Ablation on Injection Methods.

Method	Mechanics	Optics	Thermal	Material	Avg
Concat	0.541	0.633	0.488	0.500	0.540
Cross Attention	0.541	0.633	0.522	0.516	0.553
PhysRAG (Ours)	**0.586**	**0.660**	**0.540**	**0.526**	**0.578**

features in the backbone, struggles to handle unaligned spatiotemporal dynamics, yielding the lowest performance with a score of 0.540. The **direct cross attention** method introduces a mechanism that allows the model to attend to the retrieved features while aligning them with the current video features, which improves the performance to a score of 0.553. However, this approach still underperforms due to the introduction of dense reference features that bring significant visual noise, such as irrelevant background details and appearance cues, which hinder the model's ability to extract pure physical priors. In our approach, the **learnable query-based mechanism** serves as an information bottleneck, selectively focusing on the essential physical dynamics while filtering out disruptive noise. This targeted integration enables the model to explicitly distill the most important physical dynamics, leading to a substantial improvement in performance, with our method achieving a state-of-the-art score of 0.578.

Physical Prior Probing. To further verify whether Query Inject preserves physically meaningful information, we conduct a linear probing experiment on a held-out WISA-80K probing set with 17 physical categories. As shown in Table 4, the adapter hidden tokens achieve performance close to the raw retrieved features, with only a small decrease in accuracy and Macro-F1. This suggests that the learnable-query bottleneck effectively distills physical-category information from retrieved videos while filtering out irrelevant visual details.

Ablation on Training Strategies. Table 5 and Figure 5 evaluate the necessity of joint training. Standard **SFT** on curated data yields a baseline of 0.546, proving that gains stem significantly from the RAG mechanism rather than just data quality. The **SFT+RAG** variant (which involves an initial SFT phase, followed by freezing the backbone to train only queries) achieves only 0.551, indicating that joint optimization is crucial for the model to effectively

Table 4: Physical prior probing on 17 physical categories.

Representation	Acc. (%)	Macro-F1
Raw retrieved feat.	67.65	64.25
Adapter tokens	67.06	63.14

Table 5: Ablation on Training Strategies.

Method	Mechanics	Optics	Thermal	Material	Avg
SFT	0.575	0.626	0.511	0.475	0.546
SFT+RAG	0.566	0.626	0.511	0.500	0.551
PhysRAG (Ours)	**0.586**	**0.660**	**0.540**	**0.526**	**0.578**

internalize retrieved physical priors. As visualized in Figure 5, **PhysRAG** with joint training produces superior physical consistency—such as accurate reflections and causal fluid dynamics—compared to frozen-backbone variants. These results validate that the synergy between a trainable DiT backbone and our RAG module is essential for enforcing strict physical laws.

Ablation on Quality of Training Data. Table 6 evaluates the effectiveness of our data filtering pipeline by comparing models trained on raw data versus our curated physical dataset. Comparing **Random-SFT** with **Filtered-SFT**, the filtered data yields a baseline improvement from 0.539 to 0.546, demonstrating that removing low-quality samples provides a cleaner training signal. More importantly, when integrating our RAG module, **PhysRAG + Random Data** achieves only 0.544, even underperforming compared to the pure Filtered-SFT baseline. In contrast, our full method, **PhysRAG + Filtered Data**, reaches a state-of-the-art score of 0.578. These results show that RAG benefits substantially from high-quality, physically consistent data.

Ablation on Injection Layers. Table 7 investigates the impact of injecting physical features into different layers of the Video DiT backbone. We compare three single-layer injection variants—**RAG-Front** (layer 0), **RAG-Mid** (layer 15), and **RAG-Back** (layer 24)—against our proposed **Multi-layer** approach (layers 0, 1, and 2). The results show that single-layer injection, particularly at the extreme front or back, is insufficient to capture the full complexity of physical dynamics. In contrast, our multi-layer strategy achieves the highest performance (0.578) across all metrics. This superiority stems from two factors: first, injecting into the initial layers allows the DiT to prioritize learning high-level physical representations, early in the generation process; second, the multi-layer design facilitates a more comprehensive integration of physical priors across different levels of feature abstraction, ensuring better consistency in the resulting video.

Computational Overhead. Table 8 reports the additional cost introduced by PhysRAG under the same inference setting. Compared with Wan2.2-5B, PhysRAG adds only 114.25M parameters (+2.28%) and increases the inference latency by 0.82s per sample (+1.24%). The FAISS retrieval itself takes only about 0.0065s, which is negligible compared with the diffusion denoising process. These

Table 6: Ablation of Quality of Training Data.

Method	Mechanics	Optics	Thermal	Material	Avg
Random-SFT	0.575	0.606	0.500	0.475	0.539
Filtered-SFT	0.575	0.626	0.511	0.475	0.546
PhysRAG + Random Data	0.558	0.640	0.511	0.466	0.544
PhysRAG + Filtered Data (Ours)	**0.586**	**0.660**	**0.540**	**0.526**	**0.578**

Table 7: Ablation on Query Inject Layers.

Method	Mechanics	Optics	Thermal	Material	Avg
RAG-Front	0.508	0.500	0.511	0.393	0.478
RAG-Mid	0.533	0.653	0.500	0.500	0.546
RAG-Back	0.541	0.580	0.500	0.464	0.521
Multi-layer (Ours)	**0.586**	**0.660**	**0.540**	**0.526**	**0.578**

Table 8: Computational overhead.

Method	Params	Peak Mem.	Latency	Retrieval
Wan2.2-5B	4.998B	19.65GB	65.68s	–
PhysRAG	5.114B	21.46GB	66.50s	0.0065s
Overhead	+2.28%	+1.81GB	+1.24%	negligible

results indicate that the gains of PhysRAG mainly come from effective physical-prior retrieval and injection rather than substantially increasing model scale or inference cost.

6 Conclusion

In this work, we presented **PhysRAG**, a novel retrieval-augmented generation framework designed to enhance physical awareness in video synthesis. By curating a high-quality physical dataset through a two-stage filtering pipeline and developing a learnable query-based injection mechanism, we successfully bridge the gap between explicitly retrieved videos and implicit physics modeling. Our extensive experiments on PhyGenBench and VBench demonstrate that PhysRAG achieves state-of-the-art performance, significantly improving physical consistency in complex scenarios such as fluid dynamics and mechanical interactions while maintaining high visual fidelity. These results highlight the potential of utilizing external physical exemplars to guide diffusion models toward a deeper understanding of real-world physical laws.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities, Peking University.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Bińkowski, M.a., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 23716–23736. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. 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. arXiv preprint arXiv:2503.06800 (2025)
4. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
5. Battaglia, P., Pascanu, R., Lai, M., Jimenez Rezende, D., et al.: Interaction networks for learning about objects, relations and physics. Advances in neural information processing systems **29** (2016)
6. Bear, D.M., Wang, E., Mrowca, D., Binder, F.J., Tung, H.Y.F., Pramod, R., Holdaway, C., Tao, S., Smith, K., Sun, F.Y., et al.: Physion: Evaluating physical prediction from vision in humans and machines. arXiv preprint arXiv:2106.08261 (2021)
7. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
8. 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. pp. 22563–22575 (2023)
9. Blattmann, A., Rombach, R., Oktay, K., Müller, J., Ommer, B.: Retrieval-augmented diffusion models. Advances in Neural Information Processing Systems **35**, 15309–15324 (2022)
10. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
11. Cai, Y., Li, K., Jia, M., Wang, J., Sun, J., Liang, F., Chen, W., Juefei-Xu, F., Wang, C., Thabet, A., et al.: Phygdpo: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation. arXiv preprint arXiv:2512.24551 (2025)
12. Chang, M.B., Ullman, T., Torralba, A., Tenenbaum, J.B.: A compositional object-based approach to learning physical dynamics. arXiv preprint arXiv:1612.00341 (2016)

13. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
14. 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. pp. 7310–7320 (2024)
15. Chen, H.H., Huang, H., Chen, Q., Yang, H., Lim, S.N.: Hierarchical fine-grained preference optimization for physically plausible video generation. arXiv preprint arXiv:2508.10858 (2025)
16. Chen, W., Hu, H., Chen, X., Verga, P., Cohen, W.: Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 5558–5570 (2022)
17. Chen, W., Hu, H., Saharia, C., Cohen, W.W.: Re-imagen: Retrieval-augmented text-to-image generator. arXiv preprint arXiv:2209.14491 (2022)
18. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library (2025), https://arxiv.org/abs/2401.08281
19. Fan, W., Si, C., Song, J., Yang, Z., He, Y., Zhuo, L., Huang, Z., Dong, Z., He, J., Pan, D., et al.: Vchitect-2.0: Parallel transformer for scaling up video diffusion models. arXiv preprint arXiv:2501.08453 (2025)
20. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)
21. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3749–3761 (2022)
22. Guen, V.L., Thome, N.: Disentangling physical dynamics from unknown factors for unsupervised video prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11474–11484 (2020)
23. Hao, Y., Chen, C., Mian, A.S., Xu, C., Liu, D.: Enhancing physical plausibility in video generation by reasoning the implausibility (2025), https://arxiv.org/abs/2509.24702
24. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
25. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
26. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
27. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
28. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

29. Intelligence, P., Amin, A., Aniceto, R., Balakrishna, A., Black, K., Conley, K., Connors, G., Darpinian, J., Dhabalia, K., DiCarlo, J., Driess, D., Equi, M., Esmail, A., Fang, Y., Finn, C., Glossop, C., Godden, T., Goryachev, I., Groom, L., Hancock, H., Hausman, K., Hussein, G., Ichter, B., Jakubczak, S., Jen, R., Jones, T., Katz, B., Ke, L., Kuchi, C., Lamb, M., LeBlanc, D., Levine, S., Li-Bell, A., Lu, Y., Mano, V., Mothukuri, M., Nair, S., Pertsch, K., Ren, A.Z., Sharma, C., Shi, L.X., Smith, L., Springenberg, J.T., Stachowicz, K., Stoeckle, W., Swerdlow, A., Tanner, J., Torne, M., Vuong, Q., Walling, A., Wang, H., Williams, B., Yoo, S., Yu, L., Zhilinsky, U., Zhou, Z.: $\pi^*_{0.6}$: a vla that learns from experience (2025), https://arxiv.org/abs/2511.14759
30. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Galliker, M.Y., Ghosh, D., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., LeBlanc, D., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Ren, A.Z., Shi, L.X., Smith, L., Springenberg, J.T., Stachowicz, K., Tanner, J., Vuong, Q., Walke, H., Walling, A., Wang, H., Yu, L., Zhilinsky, U.: $\pi_{0.5}$: a vision-language-action model with open-world generalization (2025), https://arxiv.org/abs/2504.16054
31. 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)
32. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. arXiv preprint arXiv:2411.02385 (2024)
33. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
34. Kuaishou Technology: Kling ai: High-fidelity text-to-video generation. https://www.kuaishou.com/kling (2024)
35. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
36. Li, C., Michel, O., Pan, X., Liu, S., Roberts, M., Xie, S.: Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. arXiv preprint arXiv:2503.09595 (2025)
37. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 19730–19742. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/li23q.html
38. Liu, D., Zhang, J., Dinh, A.D., Park, E., Zhang, S., Mian, A., Shah, M., Xu, C.: Generative physical ai in vision: A survey. arXiv preprint arXiv:2501.10928 (2025)
39. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Xia, M., Wang, X., et al.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025)
40. Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8009–8019 (2025)

41. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177 (2024)
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), https://arxiv.org/abs/1711.05101
43. Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., Zhou, Y., Sun, D., Zhou, D., Zhou, J., Tan, K., An, K., Chen, M., Ji, W., Wu, Q., Sun, W., Han, X., Wei, Y., Ge, Z., Li, A., Wang, B., Huang, B., Wang, B., Li, B., Miao, C., Xu, C., Wu, C., Yu, C., Shi, D., Hu, D., Liu, E., Yu, G., Yang, G., Huang, G., Yan, G., Feng, H., Nie, H., Jia, H., Hu, H., Chen, H., Yan, H., Wang, H., Guo, H., Xiong, H., Xiong, H., Gong, J., Wu, J., Wu, J., Wu, J., Yang, J., Liu, J., Li, J., Zhang, J., Guo, J., Lin, J., Li, K., Liu, L., Xia, L., Zhao, L., Tan, L., Huang, L., Shi, L., Li, M., Li, M., Cheng, M., Wang, N., Chen, Q., He, Q., Liang, Q., Sun, Q., Sun, R., Wang, R., Pang, S., Yang, S., Liu, S., Liu, S., Gao, S., Cao, T., Wang, T., Ming, W., He, W., Zhao, X., Zhang, X., Zeng, X., Liu, X., Yang, X., Dai, Y., Yu, Y., Li, Y., Deng, Y., Wang, Y., Wang, Y., Lu, Y., Chen, Y., Luo, Y., Luo, Y., Yin, Y., Feng, Y., Yang, Y., Tang, Z., Zhang, Z., Yang, Z., Jiao, B., Chen, J., Li, J., Zhou, S., Zhang, X., Zhang, X., Zhu, Y., Shum, H.Y., Jiang, D.: Step-video-t2v technical report: The practice, challenges, and future of video foundation model (2025), https://arxiv.org/abs/2502.10248
44. Meng, F., Liao, J., Tan, X., Shao, W., Lu, Q., Zhang, K., Cheng, Y., Li, D., Qiao, Y., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. arXiv preprint arXiv:2410.05363 (2024)
45. Meng, S., Luo, Y., Liu, P.: Grounding creativity in physics: A brief survey of physical priors in aigc. arXiv preprint arXiv:2502.07007 (2025)
46. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
47. Peruzzo, E., Xu, D., Xu, X., Shi, H., Sebe, N.: Ragme: Retrieval augmented video generation for enhanced motion realism. In: Proceedings of the 2025 International Conference on Multimedia Retrieval. p. 1081–1090. ICMR '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3731715.3733417, https://doi.org/10.1145/3731715.3733417
48. Pika Labs: Pika: Text-to-video generation platform. https://pika.art (2024)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
50. Romero, D., Bermudez, A., Li, H., Pizzati, F., Laptev, I.: Learning to generate rigid body interactions with video diffusion models. arXiv preprint arXiv:2510.02284 (2025)
51. Runway ML: Runway gen-3 alpha. https://research.runwayml.com/gen3 (2024)
52. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
53. Song, S., Xu, Z., Zhang, Z., Zhou, K., Guo, J., Qin, L., Huang, B.: Learning plug-and-play memory for guiding video diffusion models (2025), https://arxiv.org/abs/2511.19229
54. Toth, P., Rezende, D.J., Jaegle, A., Racanière, S., Botev, A., Higgins, I.: Hamiltonian generative networks. arXiv preprint arXiv:1909.13789 (2019)

55. Tung, H.Y., Ding, M., Chen, Z., Bear, D., Gan, C., Tenenbaum, J., Yamins, D., Fan, J., Smith, K.: Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. Advances in Neural Information Processing Systems **36**, 67048–67068 (2023)
56. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
57. Wang, C., Chen, C., Huang, Y., Dou, Z., Liu, Y., Gu, J., Liu, L.: Physctrl: Generative physics for controllable and physics-grounded video generation. arXiv preprint arXiv:2509.20358 (2025)
58. Wang, J., Wang, C., Huang, K., Huang, J., Jin, L.: Videoclip-xl: Advancing long description understanding for video clip models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 16061–16075 (2024)
59. Wang, J., Ma, A., Cao, K., Zheng, J., Zhang, Z., Feng, J., Liu, S., Ma, Y., Cheng, B., Leng, D., et al.: Wisa: World simulator assistant for physics-aware text-to-video generation. arXiv preprint arXiv:2503.08153 (2025)
60. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
61. 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 (June 2023)
62. Wang, P., Wang, W., Li, Q.: Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection. arXiv preprint arXiv:2511.03997 (2025)
63. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)
64. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision **133**(5), 3059–3078 (2025)
65. Watters, N., Zoran, D., Weber, T., Battaglia, P., Pascanu, R., Tacchetti, A.: Visual interaction networks: Learning a physics simulator from video. Advances in neural information processing systems **30** (2017)
66. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6902–6912 (2024)
67. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
68. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7623–7633 (October 2023)
69. 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. pp. 18826–18836 (2025)

70. Yang, X., Li, B., Zhang, Y., Yin, Z., Bai, L., Ma, L., Wang, Z., Cai, J., Wong, T.T., Lu, H., et al.: Vlipp: Towards physically plausible video generation with vision and language informed physical prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12360–12370 (2025)
71. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
72. Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., Yih, W.t.: Retrieval-augmented multimodal language modeling. arXiv preprint arXiv:2211.12561 (2022)
73. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
74. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning (2020), https://arxiv.org/abs/1910.01442
75. Yuan, Y., Wang, X., Wickremasinghe, T., Nadir, Z., Ma, B., Chan, S.H.: Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. arXiv preprint arXiv:2509.21309 (2025)
76. Zhang, H., Huang, T., Wan, Z., Jin, X., Zhang, H., Li, H., Zuo, W.: Physchoreo: Physics-controllable video generation with part-aware semantic grounding. arXiv preprint arXiv:2511.20562 (2025)
77. Zhang, K., Xiao, C., Xu, J., Mei, Y., Patel, V.M.: Think before you diffuse: Infusing physical rules into video diffusion. arXiv preprint arXiv:2505.21653 (2025)
78. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
79. Zhang, Q., Gong, B., Tan, S., Zhang, Z., Shen, Y., Zhu, X., Li, Y., Yao, K., Shen, C., Zou, C.: Physrvg: Physics-aware unified reinforcement learning for video generative models. arXiv preprint arXiv:2601.11087 (2026)
80. 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 (2025), https://arxiv.org/abs/2505.23656
81. Zhao, Y., Li, H., He, X., Wu, B.: Phyrpr: Training-free physics-constrained video generation. arXiv preprint arXiv:2601.09255 (2026)
82. 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)
83. Zhu, C., Wu, Y., Wang, S., Wu, G., Wang, L.: Motionrag: Motion retrieval-augmented image-to-video generation (2025), https://arxiv.org/abs/2509.26391