WISA: World Simulator Assistant for Physics-Aware Text-to-Video Generation

Jing Wang[1,2§*], **Ao Ma**[2†*], **Ke Cao**[2*], **Jun Zheng**[1], **Jiasong Feng**[2],
Zhanjie Zhang[2], **Wanyuan Pang**[3], **Xiaodan Liang**[1,4,5‡]
[1]Shenzhen Campus of Sun Yat-Sen University, [2]360 AI Research,
[3]University of Science and Technology Beijing, [4]Peng Cheng Laboratory,
[5]Guangdong Key Laboratory of Big Data Analysis and Processing,
wangj977@mail2.sysu.edu.cn

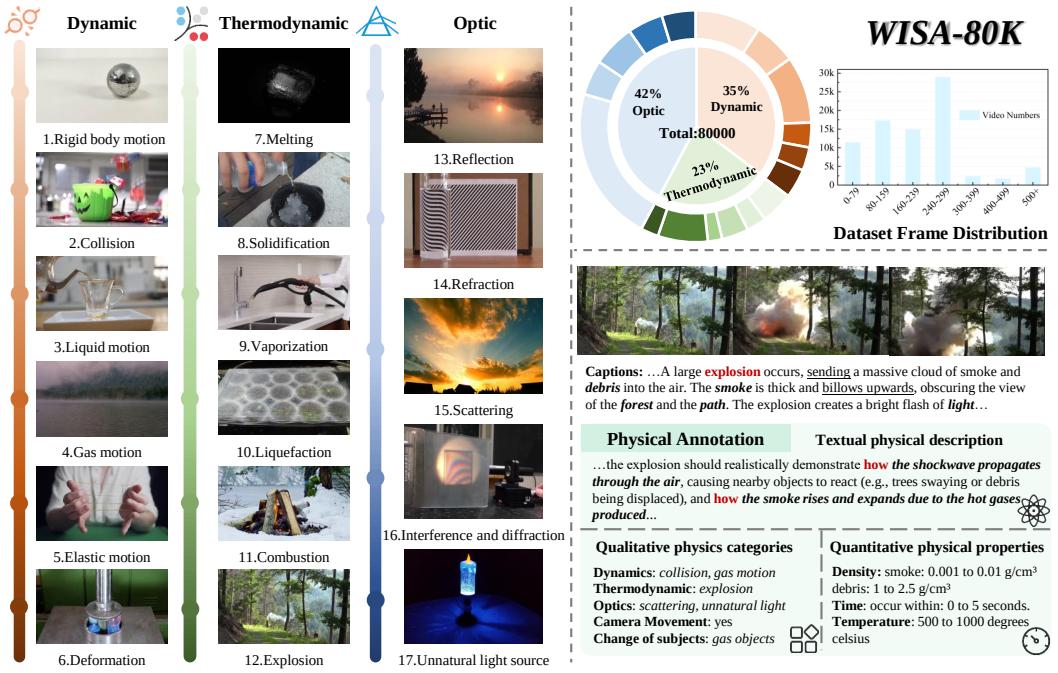

Figure 1: **Overview of our physical dataset WISA-80K**. (Left) Examples of 17 physical phenomena across three physics categories in WISA-80K. (Top right) WISA-80K consists of approximately 80,000 video clips, with 35% related to *Dynamics*, 23% to *Thermodynamics*, and 42% to *Optics*. (Top right) Distribution of frame counts across all videos in WISA-80K. (Bottom right) An example of physical annotation in WISA-80K.

Abstract

Recent advances in text-to-video (T2V) generation, exemplified by models such as Sora and Kling, have demonstrated strong potential for constructing world simulators. However, existing T2V models still struggle to understand abstract physical principles and to generate videos that faithfully obey physical laws. This limitation stems primarily from the lack of explicit physical guidance, caused by a significant gap between high-level physical concepts and the generative capabilities of current models. To address this challenge, we propose the **W**orld **S**imulator **A**ssistant (**WISA**), a novel framework designed to systematically decompose and integrate physical principles into T2V models. Specifically, WISA decomposes physical knowledge into three hierarchical levels: textual physical descriptions, qualitative physical categories, and quantitative physical properties. It then incor-

*Equal Contribution. ‡Corresponding Authors. †Project Leader. §Conducted during internship

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

porates several carefully designed modules—such as Mixture-of-Physical-Experts Attention (MoPA) and a Physical Classifier—to effectively encode these attributes and enhance the model's adherence to physical laws during generation. In addition, most existing video datasets feature only weak or implicit representations of physical phenomena, limiting their utility for learning explicit physical principles. To bridge this gap, we present **WISA-80K**, a new dataset comprising 80,000 human-curated videos that depict 17 fundamental physical laws across three core domains of physics: dynamics, thermodynamics, and optics. Experimental results show that WISA substantially improves the alignment of T2V models (such as CogVideoX and Wan2.1) with real-world physical laws, achieving notable gains on the VideoPhy benchmark. Our data, code, and models are available in the https://wisav1.github.io/WISA/.

1 Introduction

Many recent studies (e.g., Cosmos [1], Kling [14], Step-Video-T2V [21], Sora [26], and CogVideoX [41]) have endeavored to develop robust text-to-video (T2V) models for building world simulators [39, 6, 43]. While these models are capable of generating highly realistic and text-consistent videos, leveraging the scale of their data and architectures, they still face challenges in understanding abstract physical principles and producing videos that fully align with real-world physical laws [3, 23].

The substantial gap between abstract physical laws and their visual manifestations presents a significant challenge for injecting physical guidance into T2V models. Physical principles or laws are often conveyed through abstract natural language, reflecting the underlying operational logic of the real world. In contrast, generative models map textual descriptions directly to the visual appearance of objects, including their color and shape. There is a complex logical reasoning process between physical principles and the visual physical phenomena they give rise to. However, generative models, which are trained to map learned data distributions, struggle to extract appropriate physical information from a single textual instruction and translate it into a physically consistent visual representation for a specific scenario. This challenge becomes even more pronounced in video generation, where the strict temporal order of physical events must be preserved.

To this end, we propose the **W**orld **S**imulator **A**ssistant (**WISA**), which decomposes abstract physical principles into multiple categories of physical information and integrates them into T2V models to enable physics-aware generation. Specifically, it decomposes physical principles into three levels: textual physics descriptions, qualitative physics categories, and quantitative physical properties, and designs appropriate tailored injection methods for each type of information. The **Textual Physical Description** outlines the physical principles relevant to the scene, the resulting physical phenomena, and their specific visual manifestations. WISA incorporates this information by concatenating it with the caption before feeding it into the text encoder. The **Qualitative Physics Categories** indicate the types of physical phenomena that may be present in a scene. Following the focus of existing physical T2V benchmarks (e.g., VideoPhy and PhyGenBench), WISA targets 17 representative phenomena commonly encountered in video generation tasks. These span three major branches of physics (i.e., dynamics, thermodynamics, and optics) and include examples such as collision (dynamics), refraction (optics), and melting (thermodynamics). Recognizing that different physical phenomena require distinct physical features, WISA proposes **M**ixture-of-**P**hysical-Experts **A**ttention (**MoPA**), inspired by MoE [30] and MoH [12]. MoPA assigns expert attention heads to individual physics categories, activating only the relevant experts during generation to specialize in modeling the associated phenomena. When a scene involves multiple physical phenomena, MoPA dynamically activates multiple expert heads, allowing the model to effectively capture and synthesize complex physical interactions. **Quantitative Physics Properties** refer to numerical physical attributes that directly influence the physical process, such as density, duration, and temperature. WISA encodes these properties as physical embeddings and injects them into the model via AdaLN [27]. In addition, WISA employs a Physical Classifier, which is designed to recognize qualitative physics categories and assist in perceiving physical properties.

However, extracting the above physical information and subsequently understanding physical principles from general scene video in existing datasets [25, 36] is a suboptimal approach for T2V models. Firstly, general scene videos often feature the interweaving of multiple physical phenomena. Individual physical phenomena are not prominently visualized, which makes it difficult to accurately extract physical information and establish a precise connection between the physical data and its

corresponding visual manifestation. Secondly, in these datasets, only a few videos distinctly highlight specific physical phenomena as representative examples, while most videos treat physical phenomena as secondary elements. For instance, in the Figure 2, the flow of water is a secondary element. Despite having physical information guidance, the T2V models are unable to perceive the physical principles of fluid motion from this type of data.

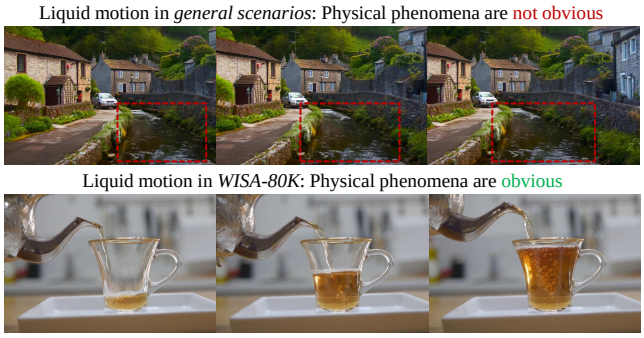

Figure 2: Comparison between general videos in Koala-36M and videos with distinct physical phenomena in WISA-80K.

To address these challenges, we collect and construct **WISA-80K**, a dataset containing **80,000** videos that represent 17 physical phenomena across three major branches of physics as shown in Figure 1, designed as a data assistant for world simulators. Specifically, based on the previously defined physics categories, we manually collect videos that clearly exhibit obvious physical phenomena corresponding to each category (e.g., as shown in the lower part of Figure 2). We then apply shot boundary detection, aesthetic quality filtering, and video captioning to the raw videos. Subsequently, we leverage GPT-4o mini to extract and decompose the physical information from the video captions into textual physics descriptions, qualitative physics categories, and quantitative physics properties for WISA.

Our contributions can be summarized as follows:

- We propose a physical principle decoupling method, bridging the gap between physical laws and generative modeling. In this method, physical principles are represented as structured physical information, encompassing textual physical descriptions, qualitative physics categories, and quantitative physical properties.
- We present the World Simulator Assistant (WISA), which guides T2V models to efficiently learn specific physical phenomena based on structured physical information, through specialized designs such as Mixture-of-Physical-Experts Attention (MoPA) and Physical Classifier.
- We manually collect 80,000 video clips that clearly showcase physical phenomena, creating the first large-scale physics video dataset, WISA-80K. It broadly covers common physical phenomena observed in the real world, encompassing 17 types of physical events (e.g., Collision, Melting, and Reflection) across three major branches of physics.
- Quantitative and qualitative experimental results demonstrate WISA and WISA-80K can greatly assist basic T2V models in producing videos that better align with real-world physical laws, while introducing only a 3.5% increase in parameter count and 5% inference time.

2 Related Work

Text-to-Video Generation Early text-to-video (T2V) generation research [10, 9, 33, 7, 8, 34, 4] primarily extend image generation models [5, 18, 17, 28, 19, 22] with temporal capabilities to enable video generation. These methods often suffered from limited realism and restricted motion dynamics. The powerful 3D spatio-temporal modeling and scalability of Diffusion Transformers [27, 15] have greatly advanced the development of visual generation models. Enabled by Diffusion Transformers, a series of recent T2V works (including OpenSora [42], Cosmos [1], Sora [26], CogVideoX [41], HunyuanVideo[x], Kling [14], Wan2.1 [31], and Step-Video-T2V [21]) significantly improve the realism and motion quality of video generation by scaling up model parameters and training data. These works are widely considered as a promising pathway towards building a World Simulator. However, they still struggle to generate videos that fully comply with real-world physical laws as they essentially fit the data distribution [13] from general-scene datasets such as Koala-36M [36] and OpenVid [25], where physical laws are not explicitly reflected and physical phenomena are not prominently presented (e.g., in the upper part of Figure 2). In contrast, our carefully curated WISA-80K dataset prioritizes the explicit presentation of typical physical phenomena as the primary criterion for video collection as presented in Figure 1. And it provides detailed and structured

physical information annotations, making it a valuable data assistant for enhancing the physical consistency of video generation.

Physical-aware Video Generation Recently, researchers [24, 3, 23, 16, 2, 20, 40, 37] have increasingly focused on improving and evaluating the physical consistency of generated videos. On the one hand, Videophy [3] and PhyGenBench [23] build test samples that reflect various physical laws, and they evaluate how well generated videos follow real-world physical laws by either training physics classification models with manual annotations or using question-answering methods based on Vision-Language models [38]. Physics-IQ[24] establishes a high-quality image-to-video benchmark designed to evaluate the ability of I2V models to generate physically consistent video sequences based on an initial state and textual instructions. On the other hand, DANO [16], MotionCraft [2], and PhysGen [20] parse objects from images and estimate their rigid motion in a differentiable manner by considering physical properties such as mass, inertia, friction, and rotation. Based on these estimations, they animate the images into videos. However, these methods are restricted to fixed physical categories (e.g., rigid motion) and static scenarios that involve only object motion, which hinders their generalizability. PhyT2V [40] leverages large language models and vision-language models to extract physical inconsistency information from generated videos. Based on the extracted physical feedback, it iteratively refines the textual description over multiple rounds, improving video generation quality. Although this approach offers generality, it introduces significant inference overhead and fails to enhance the generative model's ability to encode physical knowledge. In this paper, WISA incorporates structured physical information into the generative model, enhancing its physical perception and enabling it to handle various physical phenomena more effectively.

3 WISA-80K

3.1 Data Collection and Annotation

Physical Laws Definition: In previous physics evaluation benchmarks [3, 23], the physical phenomena emphasized in video generation tasks have primarily focused on three fundamental branches of physics: *Dynamics*, *Thermodynamics*, and *Optics*. Therefore, in this paper, we select 17 representative physical phenomena from these three core domains, excluding specialized cases such as electromagnetic phenomena.

Dynamics: We consider six common dynamic phenomena: *Collision*, *Rigid Body Motion*, *Elastic Motion*, *Liquid Motion*, *Gas Motion*, and *Deformation*. For instance, the swinging of a pendulum serves as an example of *Rigid Body Motion*.

Thermodynamics: We select six common thermodynamic phenomena: *Melting*, *Solidification*, *Vaporization*, *Liquefaction*, *Explosion*, and *Combustion*. For example, a time-lapse of melting ice cream illustrates the *Melting* phenomenon.

Optics: We define five common optical phenomena: *Reflection*, *Refraction*, *Scattering*, *Interference and Diffraction*, and *Unnatural Light Sources*.

We did not include certain physical phenomena (e.g., sublimation, condensation) due to their infrequent occurrence in real-world scenarios and the associated difficulties in collecting sufficient high-quality data. Subsequently, for each selected physical phenomenon, we manually collected videos from the Internet that clearly demonstrate the corresponding behavior, without relying on any existing video datasets for filtering or selection. During the collection process, we exclude videos with overlaid text or significant visual blur to ensure clarity and quality. As a result, we curate a dataset comprising approximately 40,000 videos.

Pre-processing and Caption: We use PySceneDetect [29] to split the raw videos into individual scene clips, followed by filtering based on aesthetic scores. This process yields approximately 80,000 high-quality clips. Then, we utilize Qwen2.5-VL [35] to generate video captions using the following prompt: {*Please describe the content of this video in as much detail as possible, including the objects, scenery, animals, and camera movements within the video.*} The caption length is limited to 256 tokens.

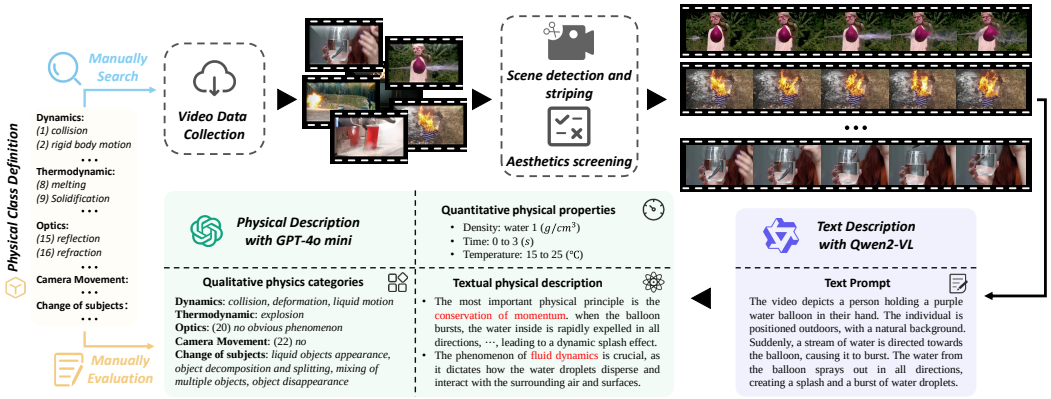

Figure 3: Pipeline of WISA-80K. We first define 17 common physical phenomena and, based on this, manually collect 80,000 video samples that clearly illustrate these phenomena. Then, we perform shot detection and aesthetic filtering on the raw videos. Text descriptions are extracted using Qwen2.5-VL, and detailed physical annotations are generated with GPT-4o mini.

3.2 Physical Information Decompose

We believe that simple video captions are not sufficient to clearly represent the physical information and related physical phenomena in a video. As shown in the Figure 3, we further constructed structured physical annotations to analyze the physical information from multiple dimensions. Specifically, we decompose the physical information into: *textual physical descriptions*, *qualitative physics categories*, and *quantitative physical properties*.

Textual physical descriptions: Provide a detailed explanation of the physical principles to be considered and the resulting intuitive physical phenomena, while supplementing the missing physical information in the prompt. For instance, the prompt "*an antique clock swings*" corresponds to the textual physical description: "... *the amplitude of the swing gradually decreases* ...".

Qualitative physics categories: These indicate the types of physical phenomena involved in a video. Although each video is collected based on a specific physical phenomenon, it may still encompass multiple types. Therefore, for each video, we identify the presence of dynamics-related, optics-related, and thermodynamics-related phenomena to effectively handle cases where multiple physical effects are coupled. Additionally, three categories of anomalies (i.e., *No obvious dynamic phenomenon*, *No obvious thermodynamic phenomenon*, and *No obvious optical phenomenon*) are introduced to account for scenarios that do not involve dynamics, thermodynamics, or optical phenomena. Furthermore, nine categories of visual phenomena are introduced, two of which pertain to whether the shot exhibits motion, while the remaining seven correspond to changes in the state of moving entities (i.e., *Object decomposition and splitting*, *Mixing of multiple objects* ... For detailed explanations, please refer to **Supplementary Material A.5**. In total, there are 29 qualitative physics categories.

Quantitative physical properties: Three physical attributes related to multiple physical phenomena are annotated, namely the density of primary motion physics, the time range during which the physical phenomenon occurs, and the temperature range during which the physical phenomenon occurs.

Due to the significant computational overhead and cost associated with video multi-modal models, the annotation of the above physical information is carried out using GPT-4o mini based on the caption. Specifically, we conduct five rounds of annotation to label qualitative physical phenomenon categories (i.e., dynamics, thermodynamics, optics, motion, the state of objects), and three rounds to annotate quantitative physical attributes (i.e., *Density*, *Time* and *Temperature*). Detailed annotation prompts and examples are provided in the **Supplementary Material A.7 and A.8**.

To verify the reliability of the automatic annotations, we conducted a manual evaluation on a randomly sampled subset of 500 videos from our dataset. Our evaluation focused on the three types of AI-generated annotations: 1) Textual Physical Descriptions: We measured human rater satisfaction (i.e., does the description accurately reflect the video's physics?). Result: 95% satisfaction rate. 2) Qualitative Physics Categories: We compared the AI-generated labels against the ground-truth labels assigned during initial video collection. Result: 76% accuracy. 3) Quantitative Physical Properties: We again measured human rater satisfaction (i.e., are the estimated density/time/temperature values plausible?). Result: 86% satisfaction rate. While some label noise is inherent in any large-scale,

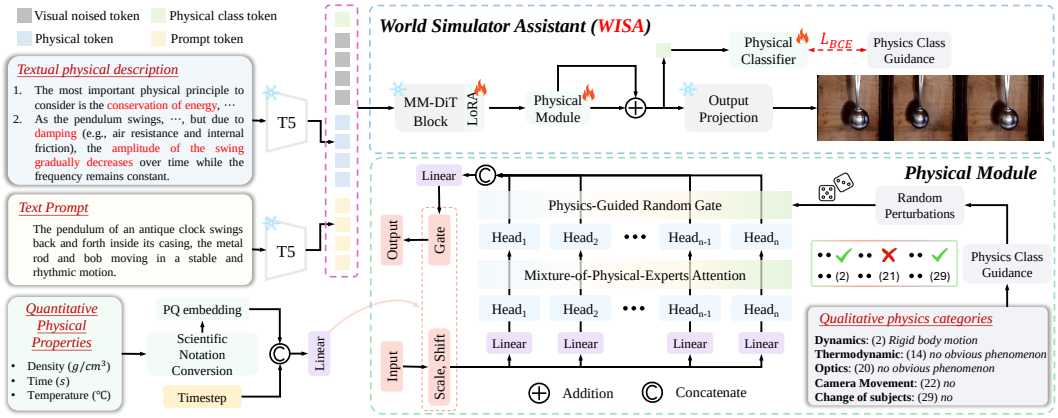

Figure 4: Overview of the proposed WISA. WISA introduces the Physical Module and Physical Classifier, which leverage structured physical annotations to guide and assist T2V models in generating physics-aware videos.

automatically annotated dataset, these results demonstrate that the overall quality of WISA-DATA-80K's annotations is high. The data is sufficiently reliable to provide a strong learning signal, as evidenced by the performance gains in experiments. More analysis of WISA-80K please refer to the **Supplementary Material A.6 and A.9**.

4 Method

4.1 Overview

Given textual physical descriptions, qualitative physical categories, and quantitative physical properties, we design the WISA framework to efficiently incorporate these conditions into existing T2V models (i.e., CogVideoX [41] or Wan2.1 [31]). To facilitate the learning of physical knowledge while preserving the model's original capabilities with limited video data, we design three distinct condition injection methods tailored to each of the three categories of physical information, as illustrated in Figure 4. Specifically, for the textual physical descriptions, we concatenate them with the video caption and leverage the generative model's inherent semantic understanding to generate visual phenomena described in text (Such as "amplitude of the swing gradually decreases over time" in Figure 4). For qualitative and quantitative physical conditions, WISA introduces the Physical Module. In this module, we propose a Mixture-of-Physical-Experts Attention (MoPA), which assigns expert heads to each physics category to model category-specific features. Quantitative physical quantities are encoded as physical embeddings and then integrated into the denoising feature within the module using AdaLN. Additionally, we introduce a qualitative Physical Classifier to help the model understand the physical conditions. Due to the significant computational and parameter cost introduced by MoPA, only one physical module is inserted after all the Diffusion transformer blocks to accelerate training and reduce the overall burden. Detailed explanations and elaborations of the Physical Module and Physical Classifier are provided in Sec. 4.2 and Sec. 4.3.

4.2 Physical Module

Most videos from real-world scenes involve the coupling of multiple physical phenomena. Even when decomposed into distinct physical categories in WISA-80K, it remains challenging for T2V models to comprehend the abstract qualitative physical categories and accurately model specific types of physical phenomena. To address this challenge, we propose a Mixture-of-Physical-Experts Attention within the Physical Module. Inspired by MoH [12], this mechanism assigns each head in the multi-head self-attention to a specific class of physical phenomena and activates the output of the relevant head only when the corresponding phenomenon is present. This approach treats each head as an expert in its domain, enabling it to independently model the properties of a particular physical phenomenon. In the presence of coupled physical phenomena, multiple corresponding expert heads are activated to effectively model the interactions among them.

Specifically, qualitative physical categories are encoded as $P_c \in \mathbb{R}^C$, where C denotes the number of defined physical phenomena (i.e., 29). Here, $P_c^i = 0$ indicates that the corresponding category is not activated, and $P_c^i = 1$ indicates that the corresponding category is activated, with i being the category index. Physical categories cannot be absolutely correct and may contain noise, such as incorrect activations or suppressions. To mitigate the impact of these noises on training, we employ a random perturbation operation, where the positions with $P_c^i = 1$ are set to 0, and the positions with $P_c^i = 0$ are set to 1.0 with a certain probability (i.e., 0.2), resulting $\hat{P}_c$. After the multi-head self-attention operation, the denoising feature $F_h \in \mathbb{R}^{N \times d \times h}$ (where h presents the number of head and $h = C$, and d denotes head dimension) will interact with $\hat{P}_c$ to activate and suppress the experts corresponding to different physical phenomena. The feature dimension is then restored through concatenation and a linear layer. The mathematical representation of this process is as follows:

$$\begin{aligned} &\hat{P}_c = \mathrm{Random}(P_c),\ F_h = \mathrm{MHSA}(F), \\ &F_o = \mathrm{Linear}(\mathrm{Reshape}(F_h \odot \hat{P}_c)) \end{aligned} \tag{1}$$

where Random denotes random perturbations operation, MHSA represents multi-head self-attention, and $\odot$ denotes element-wise multiplication.

Due to the large variations in the time and temperature spans of different physical phenomena, we first represent the temperature and time in quantitative information using scientific notation, with coefficients and exponents. These values $P_p \in \mathbb{R}^n$ are mapped through a linear layer, concatenated with the timestep embedding $T_e \in \mathbb{R}^t$, and injected by AdaLN. The mathematical representation of this process is as follows:

$$\begin{aligned} &\alpha, \beta, \gamma = \mathrm{Chunk}(\mathrm{Linear}(\mathrm{Concat}(\mathrm{Linear}(P_p), T_e)), \mathrm{dim} = -1) \\ &F = F * (1 + \alpha) + \beta,\ F_o = F_o * \gamma \end{aligned} \tag{2}$$

Generative models often consist of multiple transformer blocks with large feature dimensions, inserting the Physical Module after every block would lead to an explosion in both parameters and computational complexity. Therefore, we insert the Physical Module only after the final transformer block, achieving efficient physical information guidance while mitigating the aforementioned issues.

4.3 Physical Classifier

To guide the generative model in understanding abstract physical categories and modeling physical properties, we introduce a Physical Classifier after the Physical Module to predict qualitative physical categories. We introduce a learnable embedding vector, which we call the [PHYSICS_TOKEN]. This token is concatenated with the noisy visual tokens and text prompt tokens and is processed by the entire MM-DiT and MoPA architecture. The final hidden state of the [PHYSICS_TOKEN] $F_c \in \mathbb{R}^C$. is fed into a simple MLP classification head. This head outputs logits corresponding to the 29 qualitative physical categories, performing a multi-label classification task over the categories defined in our work. This output is used only to compute the multi-label binary cross-entropy loss for training.

$$L_{pc} = \sum_{i=1}^{C} (P_c^i \log(f_c^i) + (1 - P_c^i) \log(1 - f_c^i)), \tag{3}$$

where C is the number of physical categories, and $f_c \in \mathbb{R}^C$ represents the predicted probabilities, which are obtained by passing F_c through the sigmoid function. For each video, the model predicts which of these phenomena are present. During inference, the output from the classifier head is entirely discarded and has no influence on the video generation process. Its sole purpose is to serve as an auxiliary training signal, guiding the model to better learn and represent physical concepts.

To balance the introduced classification loss L_{pc} and the diffusion loss $L_{diffusion}$, we adopt the following loss function to optimize the physics-aware generative model.

$$L = L_{diffusion} + \lambda L_{pc} / (1 + L_{pc}.\mathrm{detach}), \tag{4}$$

where λ is balance coefficient.

Table 1: Quantitative evaluation using VideoCon-Physics conduct on the Videophy and PhyGenBench prompt lists. The best performing metrics are highlighted in **bold**.

Method	Inference	Prompts from VideoPhy [3]				Prompts from PhyGenBench [23]			
	Time (s)	IS (↑)	CLIPSIM (↑)	SA (↑)	PC (↑)	IS (↑)	CLIPSIM (↑)	SA (↑)	PC (↑)
VideoCrafter2 [7]	-	-	-	0.47	0.36	-	-	-	-
OpenSora [42]	-	28.72	0.2638	0.21	0.35	-	-	-	-
HunyuanVideo [17]	-	-	-	0.46	0.28	-	-	-	-
Cosmos-Diffusion-7B [1]	600	25.58	0.2444	0.52	0.27	20.17	0.1956	0.41	0.24
CogVideoX-5B [41]	210	30.17	0.2714	0.57	0.41	26.49	0.2590	0.34	0.42
CogVideoX-5B + PhyT2V [40]	1800	-	-	0.59	0.42	-	-	0.38	0.42
CogVideoX-5B-WISA	220	34.62	**0.2822**	**0.62**	**0.45**	27.31	**0.2813**	0.39	**0.45**
Wan2.1-14B [31]	900	36.52	0.2686	0.54	0.31	33.41	0.2488	0.39	0.28
Wan2.1-14B-WISA	960	**38.18**	0.2813	0.60	0.36	**37.62**	0.2725	**0.42**	0.33

5 Experiments

Training Setting: We select the current representative open-source T2V model, CogVideoX-5B and Wan2.1-14B, as the base T2V models to validate the effectiveness of WISA. More training detail, please refer to **Supplementary Material A.3**.

Evaluation: We select VideoCon-Physics from Videophy [3] to evaluate the physical law consistency (PC) and semantic coherence (SA) of the generated videos. We use 160 carefully crafted prompts from PhyGenBench [23] and 344 prompts from Videophy, designed to reflect various physical principles, for testing. Following VideoCon-Physics[2], we compute SA and PC by averaging the predicted results. Additionally, we adopt the Inception Score (IS) to evaluate the perceptual quality of generated videos and employ CLIP similarity (CLIPSIM) [11] to measure text-video alignment. More evaluation detail, please refer to **Supplementary Material A.3**.

5.1 Quantitative comparison

We select five general text-to-video generation models (i.e., VideoCrafter2, OpenSora, HunyuanVideo, CogVideoX-5B and Cosmos-Diffusion-7B) and PhyT2V, a method specifically designed to enhance physical properties, for quantitative comparison, as shown in Table 1.

VideoPhy: WISA achieves state-of-the-art performance on both SA and PC metrics while maintaining high efficiency. Compared to the CogVideoX-5B, CogVideoX-5B-WISA improves SA and PC scores by 0.05 and 0.04, respectively, demonstrating that our proposed method significantly enhances the realism of generated videos. PhyT2V improves its performance by iteratively analyzing physical errors in generated video captions and adjusting the input prompts based on feedback from VideoCon-Physics scores. However, its cumbersome pipeline, which involves multiple rounds of Tarsier-34B [32] inference for video generation, introduces extremely long inference time—approximately 9 times longer than the original generation model. Cosmos exhibits poor performance due to the disordered physical processes and inconsistent temporal sequences. Furthermore, Wan2.1-14B-WISA also achieves performance improvements on Wan2.1, showing advantages in metrics such as IS and CLIPSIM. However, since Wan2.1 is not among the nine generative models used to construct the training data for VideoCon-Physics (whereas CogVideoX is included), it shows certain disadvantages in SA and PC. Despite this, Wan2.1 demonstrates superior video quality, achieving better performance in the IS.

PhyGenBench: We also evaluate WISA using prompts from PhyGenBench, observing significant improvements on both CogVidoX and Wan2.1, which demonstrates the generalizability of WISA.

5.2 Qualitative comparison

We further provide a qualitative comparison with existing methods to demonstrate the advantages of WISA. As shown in the Figure 5, for the example of the rope supports a wooden swing, CogVideoX-

[2]https://github.com/Hritikbansal/videophy/issues/5

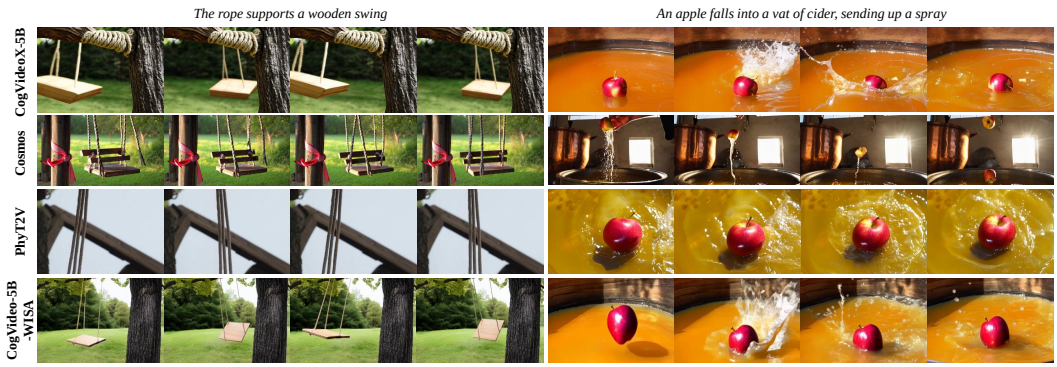

Figure 5: Qualitative comparison between CogVideoX-5B-WISA and existing T2V methods. CogVideoX-5B-WISA exhibits better alignment with real-world physical laws.

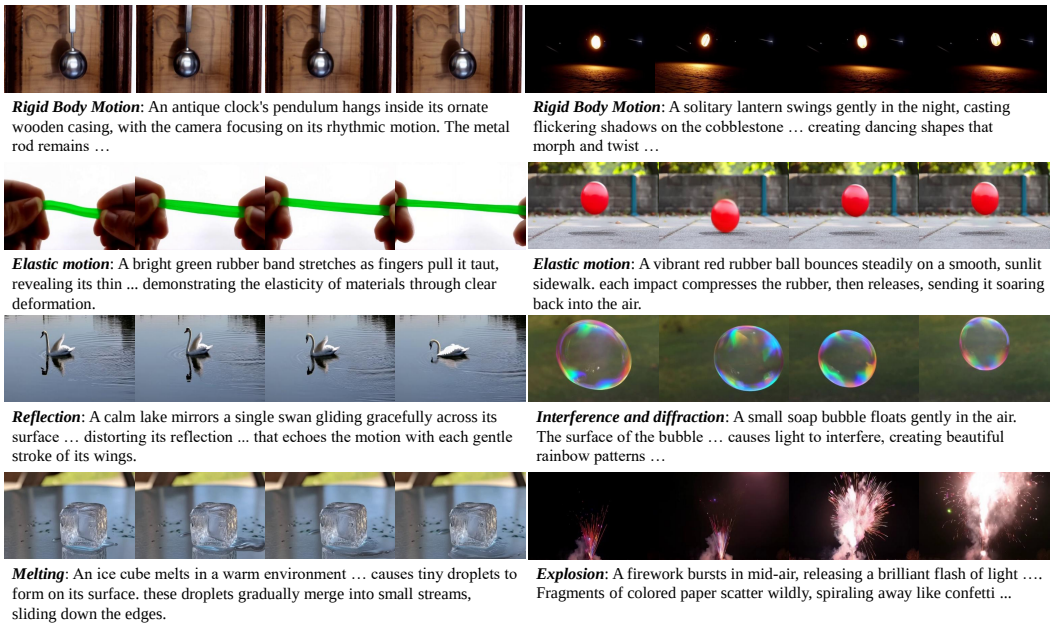

Figure 6: More samples generated by CogVideoX-5B-WISA, covering additional physical phenomena.

5B-WISA generates a video where the rope suspending the wooden seat swinging back and forth in accordance with physical laws. In contrast, CogVideoX-5B produces unstable swing motion; PhyT2V fails to generate the swinging behavior of the wooden seat; and Cosmos generates a physically inconsistent scene where the rope breaks while the wooden seat remains horizontally suspended. In the example on the right, WISA successfully simulates the process of an apple falling into water: the water surface remains calm before the apple enters, splashes form as the apple impacts the water, and the apple experiences buoyant force after submersion. However, CogVideoX-5B generates chaotic water and apple movements, PhyT2V omits the falling process, and Cosmos mistakenly generates two apples at the end. Additional videos generated by CogVideoX-5B-WISA, demonstrating various physical phenomena, are also presented in the Figure 6. All aforementioned videos, along with comparisons on Wan2.1, are provided in the Project Page.

5.3 Ablation Study

We conduct ablation studies on VideoPhy using VideoCon-Physics to verify the effectiveness of key components in our method, as shown in the Table 2. The baseline is CogVideoX-5B. As expected, removing MoPA results in a performance drop due to the absence of qualitative physical information as guidance. Similarly, the inclusion of the Physical Classifier aids the generative model in perceiving and modeling physical properties, thereby enhancing both semantic relevance and consistency with physical laws. Notably, the evaluation model VideoCon-Physics [3] is trained on samples generated

by nine different T2V models, leading to a distribution shift when compared to the real-world videos in WISA-80K. Consequently, relying solely on LoRA yields only limited improvement. To further investigate the impact of clearly-defined physical phenomena data versus general scene data on physical perception, we fine-tune LoRA on 80,000 videos from an open-source video dataset. This results in only a slight performance decline, indicating that the physically grounded videos in WISA-80K provide substantial value for modeling physical properties.

Table 2: Ablation study on the key components of WISA. "PC" denotes the Physical Classifier.

Setting	Data	Textual Physical Descriptions	Qualitative Physics Categories	Quantitative physical properties	SA (↑)	PC (↑)
Baseline	-	-	-	-	0.57	0.41
only LoRA	General Data	-	-	-	0.57	0.40
only LoRA	WISA-80K	✓	-	-	0.58	0.43
w/o MoPA	WISA-80K	✓	-	✓	0.59	0.43
w/o MoPA	WISA-80K	-	-	✓	0.57	0.42
w/o PC	WISA-80K	-	✓	-	0.60	0.44
w/o PC	WISA-80K	✓	✓	✓	0.61	0.44
WISA	WISA-80K	✓	✓	✓	0.62	0.45

5.4 User Preference

The physical consistency of generated videos is abstract and difficult to quantify directly. Therefore, we conduct a human evaluation to assess the effectiveness of WISA. Specifically, we selected three representative models for comparison. The evaluation considered two aspects: semantic consistency and physical alignment. Each candidate model is ranked in both aspects, receiving a score based on its ranking: 3 points for first place, 2 points for second, and 0 points for last. We collected preference results from 100 participants. As shown in the Figure 7 demonstrates that WISA achieves a significant advantage in physical alignment, while also maintaining strong semantic consistency.

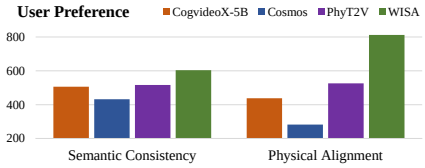

Figure 7: User Preference on VideoPhy prompts.

5.5 Attention Map Analysis

We further conduct a visual analysis of the Mixture-of-Physical-Experts attention maps, aiming to investigate whether different physical experts focus on the regions corresponding to distinct physical phenomena. As shown in the Figure 8, the rigid body motion expert perfectly focuses on the swing region, while the non-dynamics expert attends to the static background with no apparent motion. This demonstrates that the MoPA effectively models and captures the corresponding physical attributes.

6 Conclusion

In this paper, we present WISA framework, which decomposes physical principles into structured physical information, including textual physical descriptions, qualitative physical categories, and quantitative physical properties. To help T2V models learn these physical aspects effectively, WISA incorporates two key components: the Mixture-of-Physical-Experts Attention and the Physical Classifier. Building on this, we construct WISA-80K, a dataset containing 80,000 video clips that cover 17 physical phenomena across three fundamental categories of physics, providing a high-quality data foundation. Experimental results show that WISA and WISA-80K can effectively help produce videos that better align with real-world physical laws, while the additional computational overhead is under 5%. We hope that WISA can provide valuable insights into the research on building powerful world simulators. We further discuss the limitation of this paper in the Supplementary Material A.2.

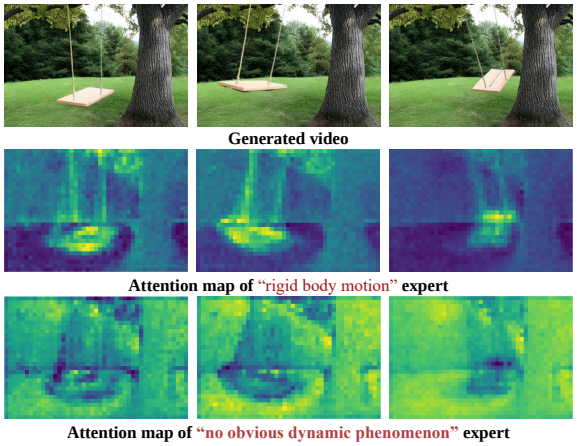

Figure 8: Attention maps of different physical experts.

ACKNOWLEDGEMENTS

This work is supported by Scientific Research Innovation Capability Support Project for Young Faculty (No.ZYGXQNJSKYCXNLZCXM-I28), National Natural Science Foundation of China (NSFC) under Grants No.62476293 and General Embodied AI Center of Sun Yat-sen University.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[2] Luca Savant Aira, Antonio Montanaro, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. *arXiv preprint arXiv:2405.13557*, 2024.

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

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[7] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024.

[8] Jiasong Feng, Ao Ma, Jing Wang, Bo Cheng, Xiaodan Liang, Dawei Leng, and Yuhui Yin. Fancyvideo: Towards dynamic and consistent video generation via cross-frame textual guidance. *arXiv preprint arXiv:2408.08189*, 2024.

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

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[12] Peng Jin, Bo Zhu, Li Yuan, and Shuicheng Yan. Moh: Multi-head attention as mixture-of-head attention. *arXiv preprint arXiv:2410.11842*, 2024.

[13] 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. *arXiv preprint arXiv:2411.02385*, 2024.

[14] Kuaishou. Kling. `https://klingai.kuaishou.com/`, 2024. Accessed: 2024-09-03.

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

[16] Simon Le Cleac'h, Hong-Xing Yu, Michelle Guo, Taylor Howell, Ruohan Gao, Jiajun Wu, Zachary Manchester, and Mac Schwager. Differentiable physics simulation of dynamics-augmented neural objects. *IEEE Robotics and Automation Letters*, 8(5):2780–2787, 2023.

[17] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

[18] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[19] Shanyuan Liu, Dawei Leng, and Yuhui Yin. Bridge diffusion model: bridge non-english language-native text-to-image diffusion model with english communities. *arXiv preprint arXiv:2309.00952*, 2023.

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

[21] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025.

[22] Yuhang Ma, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Hico: Hierarchical controllable diffusion model for layout-to-image generation. *Advances in Neural Information Processing Systems*, 37:128886–128910, 2024.

[23] 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. *arXiv preprint arXiv:2410.05363*, 2024.

[24] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.

[25] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

[26] OpenAI. Sora. `https://openai.com/`, 2024. Accessed: 2024-09-03.

[27] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[29] PySceneDetect Developers. Pyscenedetect. `https://www.scenedetect.com/`.

[30] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

[31] Wan Team. Wan: Open and advanced large-scale video generative models. 2025.

[32] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024.

[33] Jing Wang, Ao Ma, Jiasong Feng, Dawei Leng, Yuhui Yin, and Xiaodan Liang. Pt-t2i/v: An efficient proxy-tokenized diffusion transformer for text-to-image/video-task. In *The Thirteenth International Conference on Learning Representations*.

[34] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[35] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[36] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. *arXiv preprint arXiv:2410.08260*, 2024.

[37] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pages 1–20, 2024.

[38] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024.

[39] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

[40] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. *arXiv preprint arXiv:2412.00596*, 2024.

[41] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[42] 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, March 2024.

[43] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly articulate the main contributions of the paper, including the proposed WISA and WISA-80K (3), the design of the Physical Module (4.2), and the overall performance of WISA in experiments (5). These claims are well supported by the methods and experimental results presented in the main paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attainable by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the proposed method have been clearly discussed in the paper (A.2).

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides all necessary implementation details, including model architecture, training settings, hyperparameters, and dataset preprocessing steps. We will release the full code and trained models upon publication.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: To preserve anonymity during the review process, we have not released the code and data yet. However, we commit to open-sourcing both the code and data, along with detailed instructions for reproducing the experiments, upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all relevant experimental details in Supplementary Material A.3. These details ensure the clarity and reproducibility of our experimental results.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: While we do not report explicit error bars or confidence intervals, we follow the standard evaluation protocols from prior related work to ensure fair comparisons.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed information about the compute resources used for our experiments, including the type of GPUs and other relevant training settings (A.3). These details are included to support reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper fully adheres to the NeurIPS Code of Ethics. We have ensured that all ethical guidelines regarding data usage, privacy, and research integrity have been strictly followed.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

Justification: The paper does not explicitly discuss societal impacts. However, since the work focuses on text-to-video generation, we acknowledge potential misuse, such as generating misleading or harmful content. We plan to include safeguards and responsible usage guidelines when releasing models to mitigate such risks.

Guidelines:

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

11. **Safeguards**

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

Answer: [Yes]

Justification: Since our work focuses on text-to-video generation, we recognize its potential for misuse, such as in generating misleading or fake images. To mitigate this, we intend to release our model under a research-only license with clear usage restrictions. The dataset used is public and curated to avoid harmful content.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use several publicly available datasets and pretrained models (e.g., CogVideoX, Wan2.1) in our experiments. All datasets and models are properly cited in the main text and/or appendix. We ensured that their licenses are respected, and no data under restrictive or ambiguous licenses was used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: While we plan to release the code and model after the review process, at the time of submission no new assets are publicly released. Upon release, we will ensure that comprehensive documentation, licensing terms, and usage guidelines are provided in accordance with the NeurIPS guidelines.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve any crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were only used for writing and editing assistance. They were not involved in any part of the core methodology, experimental design, or scientific contribution.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Technical Appendices and Supplementary Material

A.1 Broader Impacts and Ethical Considerations

The development of powerful generative models, particularly those capable of creating realistic video content, carries a significant responsibility. The work presented in this paper, while aimed at advancing scientific understanding and technical capability in physics-aware video generation, is not exempt from potential societal impacts and ethical challenges. This section transparently discusses these issues and outlines the steps we are taking to mitigate potential harms.

Potential for Misuse and Societal Impact: We recognize the dual-use nature of our work. The same technology that can be used for creative applications, scientific simulation, or special effects could also be exploited to create compelling but fabricated content for malicious purposes. The enhanced physical realism of videos generated by a model like WISA could amplify the verisimilitude of synthetic media, increasing its potential to be used for misinformation, disinformation (e.g., "deepfakes"), or propaganda. Such content could erode public trust, be used as false evidence, or create realistic depictions of accidents or violence to incite fear.

Dataset Ethics and Release Strategy: The creation and distribution of any large-scale dataset require careful consideration of privacy, copyright, and consent. The WISA-80K dataset is constructed using video clips from publicly available channels on YouTube.

- Copyright and Terms of Service: To respect the rights of content creators and to comply with platform Terms of Service, we have adopted a metadata-only release strategy. We will not host or distribute any video files, clips, or raw pixel data. The released dataset will contain only the public YouTube video IDs, the relevant start and end timestamps of the physical phenomena, and our corresponding physical annotations. This is a standard and widely accepted practice in the research community (e.g., AudioSet) that enables reproducible research while avoiding copyright infringement.
- Responsible Data Access: To further ensure responsible use, the WISA-80K metadata will not be available for direct public download. Instead, we will implement a gated access mechanism. Researchers wishing to use the dataset must submit a request outlining their institutional affiliation and research purpose. They will be required to agree to a Data Usage Agreement (DUA), which will stipulate that: (1) the dataset is to be used for non-commercial research purposes only; (2) users are responsible for their own adherence to YouTube's Terms of Service when accessing the original videos; and (3) the metadata and any derived video content may not be redistributed.
- Content Creator Rights: We will provide a clear and accessible opt-out mechanism on our future project page. Any content creator can request the removal of their video's ID from our dataset at any time, and we are committed to promptly honoring all such requests.

Annotation Process: Quality and Bias: The physical annotations in WISA-80K were generated using a large language model (GPT-4o-mini). We acknowledge that this automated process may introduce biases or factual inaccuracies ("hallucinations").

- Human Validation: To quantify the quality and reliability of these annotations, we conducted a human validation study on a randomly sampled subset of 500 videos. The results (detailed in Appendix X) indicate a high degree of quality, with 95% satisfaction for textual descriptions and 86% satisfaction for the plausibility of quantitative estimates. The accuracy for qualitative category classification was 76% against human-assigned labels.
- Dataset Positioning: These results confirm that while some label noise is inherent, the annotations provide a strong and reliable learning signal. Nonetheless, WISA-80K should be understood as a large-scale, weakly-supervised dataset rather than a gold-standard, error-free resource. We encourage future work to further refine and build upon these annotations.

Mitigation Strategies for the Generative Model: To address the risks of misuse associated with the generative model itself, we commit to the following safeguards for any future public release:

- Visible Watermarking: All video outputs generated by our released model will be programmatically embedded with a clear and persistent visible watermark to identify them as synthetic.
- Responsible AI License: We plan to release the model under a Responsible AI License (e.g., a CreativeML Open RAIL-M license). Such licenses contractually prohibit users from employing the model for malicious, deceptive, illegal, or unethical purposes, including the generation of harmful misinformation.

A.2 Limitation

Although our approach significantly improves the ability of existing T2V models to generate videos that align with real-world physical laws, it still has the following limitations: 1) **Limited physical categories**: We collect 80,000 videos in WISA-80K, covering 17 types of physical phenomena. However, due to constraints in time and manpower, the dataset does not include all physical phenomena encountered in the real world, such as corrosion or vacuum environments. 2) **Limited physical information guidance**: WISA primarily provides high-level semantic guidance and lacks detailed constraints at the physical mechanism level (e.g., energy conservation, Newton's laws). However, introducing more detailed physical principle constraints currently requires modeling object motion based on image or 3D information, which suffers from poor generalization and can only handle limited categories and scenarios. How to incorporate physical principle constraints into text-to-video generation while maintaining generalization remains an area worth further research. 3) **Expanding WISA to Unseen Classes**: We conducted additional qualitative evaluations on unseen physical categories such as corrosion and electromagnetism, as shown in Figure 9. As expected, both our WISA-enhanced model and the base model struggle to generate physically plausible videos for these categories. This limitation primarily stems from the absence of relevant concepts and visual examples in the WISA-80K training set, underscoring the current challenge of generalizing to entirely out-of-distribution physical phenomena. However, we argue that WISA's modular design, particularly the Mixture-of-Physical-Experts Attention (MoPA) mechanism, makes it inherently more scalable than monolithic architectures. When introducing a new physical category, traditional fine-tuning methods (e.g., applying LoRA on the base model) typically require retraining large portions of the network to assimilate the new knowledge. In contrast, MoPA allows for a more targeted and efficient expansion: we can simply add and train a new expert head dedicated to the novel phenomenon while keeping the existing experts largely frozen, thus preserving prior knowledge and facilitating incremental learning.

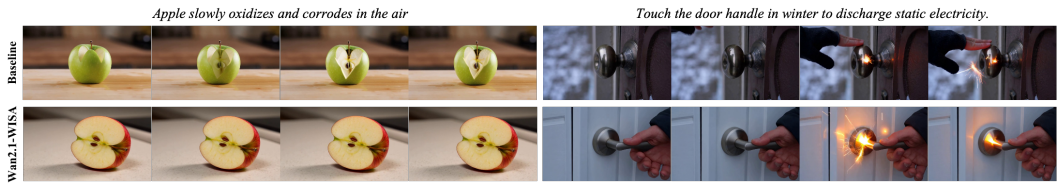

Figure 9: Visual comparison on unseen categories.

Further work: In future research, we plan to explore physically consistent generation for the image-to-video (I2V) task. In principle, WISA can be applied to I2V as well. The main distinction between I2V and text-to-video (T2V) lies in the input: I2V models receive a clean latent code for the first frame, whereas the rest of the input text and model architecture remain the same. With explicit initial frame information, the I2V model primarily learns to infer subsequent states, generating intermediate and final frames that follow plausible physical processes.

Compared with T2V, I2V is particularly relevant in embodied intelligence and robotics, where agents often observe a concrete initial state and must predict or plan the outcomes of actions in a physically consistent way. While T2V remains a more challenging task—requiring the model to imagine the entire physical process from scratch based solely on abstract text—the I2V task provides a complementary and practically important setting to study and improve physical consistency in video generation. We aim to extend WISA to I2V in future work, leveraging its modular design to generate videos that maintain accurate physical dynamics starting from a known initial state.

A.3 Training and Evaluation Detail

Training: We choose two representative open-source T2V models—CogVideoX-5B and Wan2.1-14B—as the base models to validate the effectiveness of the proposed WISA. WISA is trained on our constructed WISA-80K dataset for 8,000 steps, using a learning rate of 2e-5 and a batch size of 16. For CogVideoX-5B, the video resolution is set to 480×720 with 49 frames per video, while for Wan2.1-14B, the resolution is 480×832 with 81 frames. We adopt LoRA with a rank of 128 and an alpha of 16. During training, only the physical module, physical classifier, and LoRA parameters are updated, resulting in a total of 187 million learnable parameters for CogVideoX-5B and 587 million for Wan2.1-14B. All experiments are conducted on 8 A100 GPUs, each equipped with 80 GB of memory.

Evaluation: VideoCon-Physics was trained by collecting videos generated from nine different models, which were manually annotated for adherence to real-world physical laws and strong semantic consistency. Using this data, a vision-language model (VLM) was fine-tuned to serve as a reward model. During inference, the generated video and corresponding text prompt are fed into VideoCon-Physics, which outputs scores ranging from 0 to 1 for both metrics.

A.4 Inference without Annotation

For any given user prompt, we use a large language model (GPT-4o) with a set of predefined instructions to generate the required physical annotations (textual description, qualitative categories, quantitative properties). Crucially, this process uses only the input text prompt, with no access to any visual information, thus preventing any unfair information leakage. The cost is also minimal (approx. 2000 tokens per prompt). The detailed instructions are provided in the Figure 12, Figure 13, and Figure 14.

A.5 The Definition of Physical Categories

We define a total of 29 qualitative physical categories, organized into 5 major classes. The physical categories within each class, along with their corresponding category IDs, are listed as follows:

Dynamics: *1. Collision*, *2. Rigid Body Motion*, *3. Elastic Motion*, *4. Liquid Motion*, *5. Gas Motion*, *6. Deformation*, and *7. No obvious dynamic phenomenon*

Thermodynamics: *8. Melting*, *9. Solidification*, *10. Vaporization*, *11. Liquefaction*, *12. Explosion*, *13. Combustion* and *14. No obvious thermodynamic phenomenon*

Optics: *15. Reflection*, *16. Refraction*, *17. Scattering*, *18. Interference and Diffraction*, *19. Unnatural Light Sources*, and *20. No obvious optical phenomenon*

Camera motion: *21. Yes*, *22. No*

The state of object: *23. Liquids Objects Appearance*, *24. Solid Objects Appearance*, *25. Gas Objects Appearance*, *26. Object decomposition and splitting*, *27. Mixing of Multiple Objects*, *28. Object Disappearance* and *29. No Change*

Specifically, *Liquids objects appearance*: new liquids appear from the camera over time and due to external forces, such as water squeezed out of a towel. *Solid objects appearance*: new solids appear from the camera over time and due to external forces, such as Chemical reaction that produces precipitates, or cars drive in from outside the camera. *Gas objects appearance*: new gas appears from the camera over time and due to external forces. *Object decomposition and splitting*: Over time and under the action of external forces, an object is broken into multiple sub-parts: such as fruits and vegetables being cut in half. *Mixing of multiple objects*: Over time and with the action of external forces, two objects of the same state mix together, such as two solutions mixing. *Object disappearance*: As time passes and external forces act, objects disappear from the camera. *No change*: No change in the state of the object

A.6 Dataset Property Analysis

We visualize the distribution of different physics categories and video frame counts in WISA-80K, as shown in the paper Figure 1. Dynamics frequently occur in daily life, accounting for the

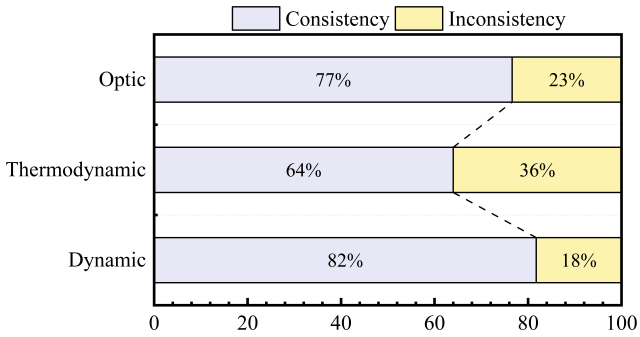

Figure 10: Accuracy of qualitative physical category annotations.

largest proportion at 47%. Optics and thermodynamics, which typically require specific temperature or environmental conditions, account for 29% and 24%, respectively. The proportions of each subcategory are shown in the outer ring of the Figure Based on the labels of the manually collected videos, we evaluate the accuracy of the qualitative physical category annotations. The results are shown in the Figure 10, where the accuracy for dynamics, optics, and thermodynamics reaches 84%, 71%, and 64%, respectively, with an overall accuracy of 75%.

A.7 More Examples and Annotation

Following the proposed physical information annotation pipeline, we construct the WISA-80K dataset. Several example videos and their corresponding annotations are shown in the Figure 11. This pipeline enables accurate and detailed annotation of physical information, ensuring that each video is comprehensively labeled with its relevant physical properties and phenomena.

A.8 Annotation Prompts

The detailed prompt used for physical information annotation is illustrated in the Figure 12, Figure 13, and Figure 14.

A.9 Word Cloud

We conducted a word frequency analysis on the textual physical description in the dataset and generated the word cloud shown in Figure 15. To filter out irrelevant words, we retained only nouns and selected them based on their frequency, from highest to lowest. Notably, physical terms such as 'motion,' 'phenomenon,' and 'light' appear more frequently, highlighting the strong physical relevance of the dataset.

A.10 Discussion of Quantitative Evaluation

During the quantitative evaluation, we observe several misjudgments in VideoCon-Physics, as shown in the Figure 16. Specifically, WISA generates a physically plausible process where the object enters the water first, followed by the splash, aligning well with real-world physical laws. However, this sample only receives a low score of 0.08 from VideoCon-Physics. We further conduct a simple test using Qwen2.5-VL for evaluation, and the model also struggles to distinguish the correct or incorrect sequence of physical events. These findings show the limitations of existing video-based physics evaluation metrics, indicating that future research into more reliable physical property assessments for videos is necessary.

Captions: The video depicts a close-up view of a mechanical device in operation. The device appears to be a type of press or hydraulic machine, characterized by its metallic components and the presence of a yellow and black striped cylinder at the top. The cylinder has a warning label, indicating caution or danger, which is typical for industrial machinery. The machine is in the process of compressing or forming a metal component, as evidenced by the movement of the upper part of the device, which is descending and exerting pressure on the lower part. The lower part of the machine consists of a circular base and a cylindrical component that is being compressed. The base has a handle attached to it, which is likely used for manual operation or adjustment. The metal components show signs of wear and use, with visible scratches and marks, suggesting that this is a well-used piece of equipment. The background is blurred, focusing attention on the machine, but it appears to be an industrial setting, possibly a workshop or a factory floor.
Textual physical description : when generating a video of the described mechanical device, it is crucial to consider the principle of force and pressure, particularly how hydraulic systems operate based on pascal's principle, which states that pressure applied to a confined fluid is transmitted undiminished in all directions. additionally, attention should be paid to the physical phenomenon of deformation, as the metal component being compressed will experience changes in shape and possibly yield stress, reflecting the material's properties under load.
Qualitative physics categories:{
Dynamics: collision, deformation,
Thermodynamics: no obvious thermodynamic phenomenon,
Optics: no obvious optical phenomenon,
Camera motion: no,
The state of object: 1. no change 2. object decomposition and splitting (if the metal component is being cut or broken) 3. solid objects appearance (if new metal components are formed during the process),
}
Quantitative physical properties: {
Density: mechanical device: 7.5 to 8.0 g/cm³ cylindrical component: 7.5 to 8.0 g/cm³,
Time: 0 to 5 seconds,
Temperature: 20 to 100 degrees celsius,
}

Captions: The video begins with a serene forest scene, showing a dirt path winding through a dense area of trees. The trees are tall and green, indicating a healthy forest environment. The path is flanked by the trunks of these trees, and the ground is covered with a layer of fallen leaves and small plants. The sky is not visible, suggesting that the camera is focused on the ground level. As the video progresses, there is a sudden and dramatic change in the scene. A large explosion occurs, sending a massive cloud of smoke and debris into the air. The smoke is thick and billows upwards, obscuring the view of the forest and the path. The explosion creates a bright flash of light, which is visible even through the smoke. The force of the explosion is so intense that it appears to shake the camera, causing it to vibrate slightly. The explosion is the focal point of the video, and it dominates the scene. The smoke and debris are the only visible elements.
Textual physical description : when generating a video of an explosion in a forest scene, it's crucial to consider the principles of conservation of momentum and energy, as well as the behavior of gases and smoke in response to rapid changes in pressure. the explosion should realistically demonstrate how the shockwave propagates through the air, causing nearby objects to react (e.g., trees swaying or debris being displaced), and how the smoke rises and expands due to the hot gases produced, following the laws of fluid dynamics.
Qualitative physics categories:{
Dynamics: collision, gas motion, deformation,
Thermodynamics: explosion,
Optics: scattering, unnatural light source,
Camera motion: yes,
The state of object: 1. gas objects appearance 2. object decomposition and splitting 3. object disappearance,
}
Quantitative physical properties: {
Density: smoke: 0.001 to 0.01 g/cm³ debris: 1 to 2.5 g/cm³ ,
Time: occur rapidly after the explosion. the main physical phenomena, including the explosion and the subsequent rise of smoke and debris, would typically take place within a very short time frame. \n\nbased on the description, the explosion itself and the immediate effects would likely occur within: 0 to 5 seconds.,
Temperature: 500 to 1000 degrees celsius,
}

Figure 11: The video data and its detailed annotations in WISA-80K.

Figure 12: Prompts for annotating textual physical descriptions and quantitative physical properties

Figure 13: Prompts for annotating qualitative physics categories

"model": "GPT-4o-mini" "role": "system" "role": "user"

content: You are a physicist with expertise in optics. Your task is to analyze the following text carefully and determine which of the listed optical phenomena are most likely represented based solely on the clear and explicit descriptions provided in the text. Base your judgment strictly on principles of light behavior, wave properties, and interactions with matter. Avoid making assumptions about phenomena not clearly implied or described. Select one or more options from the list below and provide your answer by outputting only the corresponding names, separated by commas (if multiple apply):\
Reflection (e.g., light bouncing off a surface)
Refraction (e.g., light bending as it passes through a medium)
Scattering (e.g., light dispersed in various directions, such as in fog or smoke)
Interference and diffraction (e.g., light waves overlapping, forming patterns or bending around obstacles)
Unnatural light source (e.g., artificial or unexpected light emissions, such as lasers or LEDs)
No obvious optical phenomenon.

content: caption

content: "You are a cinematographer and physics expert. Your task is to analyze the provided text description and determine whether the camera is in motion within the scene. Consider factors such as panning, tilting, tracking, or any other type of camera movement. If the text implies camera motion, output 'Yes' If the text indicates a stationary camera or lacks information about camera motion, output 'No' Only output 'Yes' or 'No'."

content: caption

content: "You are an ordinary residents. Please provide the possible phenomena's name in the following 7 types (i.e., liquids objects appearance; solid objects appearance; object decomposition and splitting; mixing of multiple objects, object disappearance, no change) without any explanation.
The types of phenomena:\
liquids objects appearance: new liquids appear from the camera over time and due to external forces, such as water squeezed out of a towel.
solid objects appearance: new solids appear from the camera over time and due to external forces, such as Chemical reaction produces precipitate or car drives in from outside the camera.
Gas objects appearance: new gas appear from the camera over time and due to external forces.
Object decomposition and splitting: Over time and under the action of external forces, an object is broken into multiple sub-parts: such as fruits and vegetables being cut in half.
Mixing of multiple objects: Over time and with the action of external forces, two objects of the same state mix together, such as two solutions mixing.
Object disappearance: As time passes and external forces act, objects disappear from the camera.
No change: No change in the state of the object."

content: Which of the above phenomena are most likely to occur in the text description: {caption}

Figure 14: Prompts for annotating qualitative physics categories

Figure 15: Word cloud generated from textual physical description, where larger words indicate higher frequencies in the dataset text

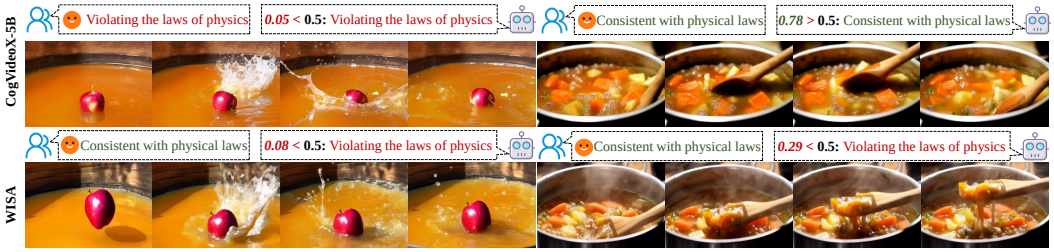

Figure 16: Human and machine evaluation results do not fully align.