PhyCo: Learning Controllable Physical Priors for Generative Motion

Sriram Narayanan[1,2] Ziyu Jiang[2] Srinivasa G. Narasimhan[1] Manmohan Chandraker[2,3]

[1]Carnegie Mellon University [2]NEC Labs America [3]UC San Diego

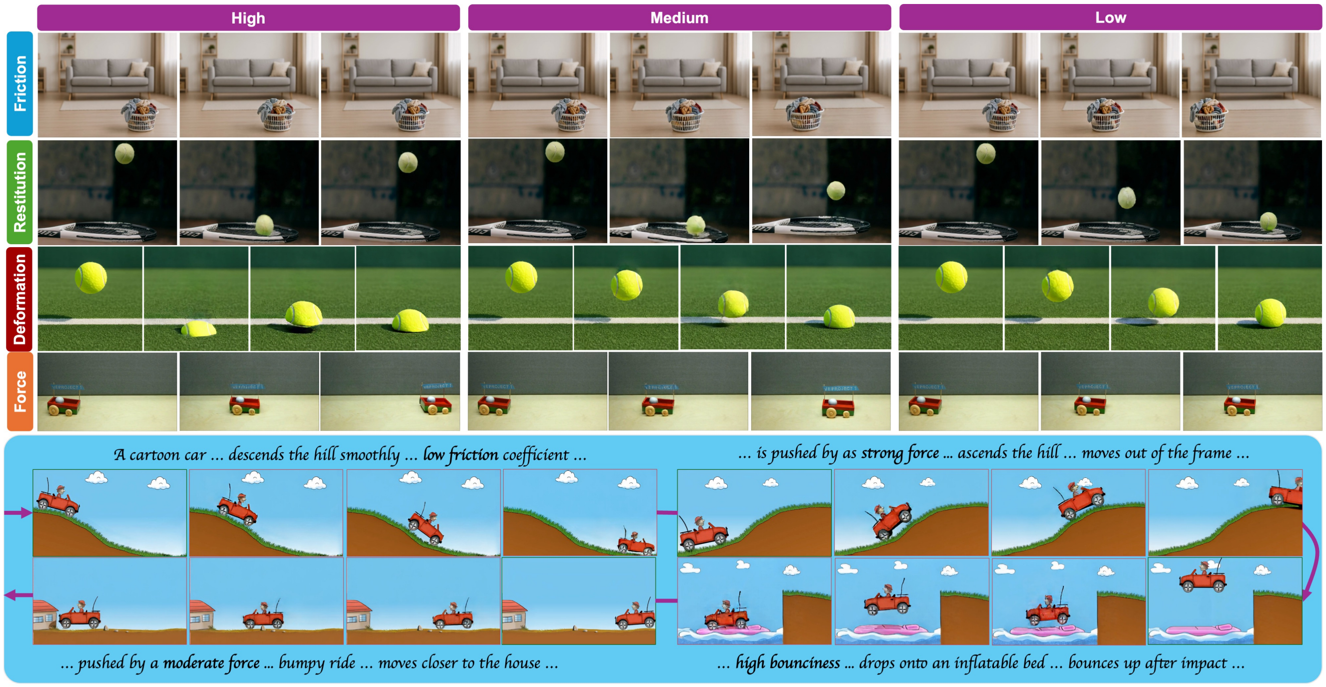

High
Medium
Low
Friction
Restitution
Deformation
Force
A cartoon car ... descends the hill smoothly ... low friction coefficient ...
... is pushed by as strong force ... ascends the hill ... moves out of the frame ...
... pushed by a moderate force ... bumpy ride ... moves closer to the house ...
... high bounciness ... drops onto an inflatable bed ... bounces up after impact ...

Figure 1. PhyCo generates videos conditioned on spatial physical property maps, producing motion consistent with real-world dynamics. Top: Continuous control over individual attributes of friction, restitution, deformation, and applied force yields smooth and physically meaningful variations. Bottom: Strong compositional generalization enables coherent motion in stylized scenes by combining multiple attributes (e.g., force + friction, restitution + deformation), even beyond the simulation domain.

Abstract

Modern video diffusion models excel at appearance synthesis but still struggle with physical consistency: objects drift, collisions lack realistic rebound, and material responses seldom match their underlying properties. We present PhyCo, a framework that introduces continuous, interpretable, and physically grounded control into video generation. Our approach integrates three key components: (i) a large-scale dataset of over 100K photorealistic simulation videos where friction, restitution, deformation, and force are systematically varied; (ii) physics-supervised fine-tuning of a pretrained diffusion model using a ControlNet conditioned on pixel-aligned physical property maps; and (iii) VLM-guided reward optimization, where a fine-tuned vision–language model evaluates generated videos with targeted physics queries and provides differentiable feedback. This combination enables a generative model to produce physically consistent and controllable outputs through variations in physical attributes—without any simulator or geometry reconstruction at inference. On the Physics-IQ benchmark, PhyCo significantly improves physical realism over strong baselines, and human studies confirm clearer and more faithful control over physical attributes. Our results demonstrate a scalable path toward physically consistent, controllable generative video models that generalize beyond synthetic training environments. Additional results and resources are available at `https://phyco-video.github.io/`.

1. Introduction

Understanding and generating physically grounded behaviors is a core challenge in building intelligent visual models. Humans effortlessly infer how objects react to forces and slide, bounce, or deform when interacting with other surfaces, but it remains challenging for video generation models. While modern diffusion-based video generators excel at synthesizing realistic textures, lighting and motion continuity, they often violate the basic laws of physics: objects hover or fall too slowly under gravity, collisions occur without rebound and soft objects fail to deform realistically [26]. Importantly, despite the large variety of data on which foundational video diffusion models are trained, it remains difficult to controllably generate variations in physical properties.

This work takes a step towards bridging the above gaps between visual and physical realism and control. Impressive advances have been made to address this gap in recent works that integrate physical simulation with generative models [22, 24, 50]. But they depend on explicit solvers, such as rigid-body dynamics in PhysGen [24] and MPM-based optimization in PhysDreamer [50] and hybrid simulation in WonderPlay [22]. While this leads to fine-grained motion coherence, the need for reconstructed 3D geometry or predefined materials during inference limits scalability and generalization. Similarly effective advances have been achieved through implicit approaches to guide motion without explicit simulation, such as PhysCtrl [37], VLIPP [45], PISA [21] and ForcePrompting [11], that embed physical cues through learned or language-driven priors using trajectory generation, vision-language reasoning, gravity-based supervision, or force-conditioned prompting. But while improving semantic consistency, they lack continuous control over a diversity of underlying physical properties.

In contrast, we present PhyCo, a framework that endows generative video models with continuous and interpretable physical property conditioning (Fig. 1). Instead of merely conditioning on external guidance, we explicitly train video diffusion models to represent and manipulate physical properties such as friction, restitution, deformation, and applied force. This enables controllable synthesis of physically consistent motion and interactions across diverse materials and contact conditions purely through generative modeling, without requiring geometric reconstruction or simulator feedback at inference. At the same time, it offers quantitative control over motion behavior and aligns naturally with representations used in physics simulators, allowing direct supervision and interpretable manipulation.

We achieve these distinctions through three novel contributions. First, we introduce a large-scale, multi-scenario dataset of 100K physically grounded simulation videos with continuous physical property annotations that disentangle visual appearance from underlying physics. The dataset – built on Kubric [12] with PyBullet [10] for physics and Blender [4] for rendering – spans diverse materials, interactions and views, covering multiple physical regimes from rigid collisions to deformable impacts, to provide a structured foundation for learning physically meaningful dynamics that generalize beyond simulation. Second, we propose physics-supervised fine-tuning of a pretrained diffusion backbone (Cosmos-Predict2 [28]) using a ControlNet [49] architecture that injects spatially aligned physical property maps. Third, we introduce VLM-guided reward optimization, where a fine-tuned vision-language model (VLM) evaluates generated videos through targeted physics questions, providing differentiable rewards that encourage physically plausible behavior. This combination of explicit conditioning and semantic feedback enables controllable, interpretable, and physically consistent video generation that generalizes from simulation-rich training to real-world scenarios without any simulator or handcrafted physical modeling at inference time.

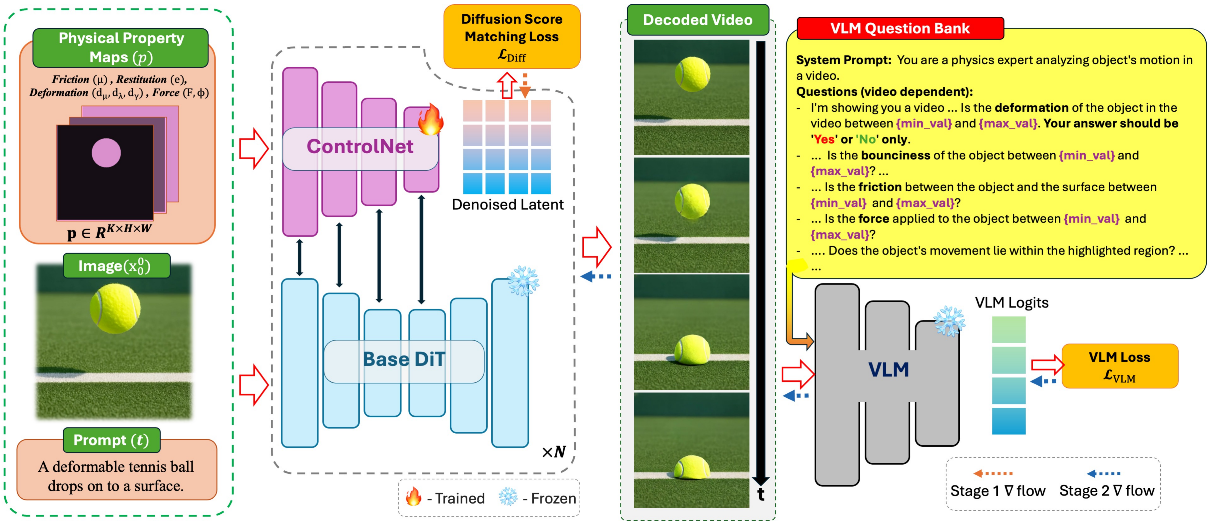
Dataset & Benchmarks	**Dataset Size**	**Photo-real**	**Object Dyn.**	**Viewpoints**	**Physical Property Annotations**
CLEVRER [47]	20k videos	✗	Multiple	✗	✗
CoPhy [1]	238k videos	✗	Single	✗	F, M, G (val only)
ComPhy [8]	8k videos	Partial (syn+real)	Multiple	✗	M, C
IntPhys [32]	15k videos	✗	Single	✗	✗
Physion [2]	16k videos	✗	Single	✗	✗
Physion++ [35]	8k videos	✗	Single	✗	F, M, R, D
ShapeStacks [13]	20k images	✗	N/A	✓	Stability only
Force Prompting [11]	38k videos	✓	Single	✓	Force (implicit)
PhyCo (Ours)	**100k videos**	✓	**Multiple**	✓	**F, M, R, D Force**

Table 1. Comparison of physics-rich datasets and benchmarks used for learning and evaluating object dynamics. F, M, R, D, and G denote friction, mass, restitution, deformation, and gravity, respectively.

In extensive evaluations on the Physics-IQ benchmark [26] and human preference studies, our approach consistently outperforms prior video models in both physical realism and controllable variation. Moreover, it generalizes to unseen materials, forces and interactions, with compositionality across variations, demonstrating that embedding physical property priors offers a scalable path toward controllable, physically consistent [23] generative world models.

2. Related Work

Physics Rich Datasets. A variety of benchmarks and datasets have been proposed to study object dynamics and to evaluate the predictive capabilities of deep models. However, most existing datasets are constrained in terms of physical property diversity, scene realism, and coverage of complex interactions. As a result, they often fall out of distribution for today's powerful generative video models. They also fall short of the requirements posed by modern generative video models, which demand richer, more diverse, and physically grounded supervision. Table 1 summarizes several representative physics-focused datasets. Although


Physical Property Maps (p)
Friction (μ) , Restitution (e), Deformation ($d_\mu, d_\lambda, d_\gamma$) , Force (F, ϕ)
$p \in R^{K\times H\times W}$
Image(x_0^0)
Prompt (ℓ)
A deformable tennis ball drops on to a surface.
ControlNet
Base DiT
×N
Diffusion Score Matching Loss
$\mathcal{L}_{\text{Diff}}$
Denoised Latent
🔥 - Trained
❄ - Frozen
Decoded Video
t
VLM Question Bank
VLM
VLM Logits
VLM Loss
$\mathcal{L}_{\text{VLM}}$
Stage 1 ∇ flow
Stage 2 ∇ flow

System Prompt: You are a physics expert analyzing object's motion in a video.
Questions (video dependent):
- I'm showing you a video ... Is the **deformation** of the object in the video between {min_val} and {max_val}. **Your answer should be 'Yes' or 'No' only.**
- ... Is the **bounciness** of the object between {min_val} and {max_val}? ...
- ... Is the **friction** between the object and the surface between {min_val} and {max_val}?
- ... Is the **force** applied to the object between {min_val} and {max_val}?
- .... Does the object's movement lie within the highlighted region? ...

...

Figure 2. Overview of the proposed PhyCo two-stage training pipeline. In Stage 1 (Physics-Supervised Fine-Tuning), the base DiT model is conditioned via a ControlNet using physically rich simulation data, enabling controllable video generation with respect to key physical properties such as friction, restitution, deformation, and applied forces. The model is optimized using a diffusion score matching loss to achieve realistic dynamics. In Stage 2 (VLM-Guided Alignment), the generated videos are analyzed by a frozen Vision-Language Model (VLM) using a curated Physics Question Bank that queries property-specific behaviors (e.g., deformation magnitude, frictional effects, motion alignment). The model receives reward gradients based on VLM logits and responses, guiding the model towards physically plausible and interpretable dynamics.

the Force-Prompting dataset [11] achieves high photorealism, it remains limited in both scene diversity and annotated physical properties. These limitations underscore the need for large-scale, physically grounded datasets that can better align generative video models with the principles of intuitive physics and enhance their ability to produce physically plausible dynamics.

Controllable Video Generation. Recent advances in video diffusion models have sparked growing interest in achieving fine-grained control and physical grounding in video generation. A large body of work has explored motion-level controllability. ATI [36] learns to generate videos from trajectory prompts by interpolating features in latent space, while Go-with-the-Flow [6] warps Gaussian noise across timesteps to precisely steer object motion. Several approaches focus on camera motion control, for example, CamI2V [51], CameraCtrl [15], and CamCo [42] encode viewpoint trajectories for accurate camera dynamics. Other efforts emphasize object-centric guidance [34, 39, 48] or box-level motion cues [38, 44]. In contrast, our goal is to enable physical property control rather than explicit motion or trajectory specification.

Physical controllability in video generation has emerged along two main directions. The first integrates explicit physics simulators with generative models to ensure physically grounded behavior. PhysGen [24] couples a 2D physics engine [5] with diffusion-based generation, while WonderPlay [22] employs material point methods (MPM) [17] to simulate interactions using Gaussian splats, leveraging diffusion models to enhance photorealism. Similar hybrid approaches [7, 33, 40, 50, 52] use MPMs or spring-mass systems to incorporate physical plausibility. PhysDreamer [50] inverts physical parameters from generated motion, and PhysAnimator [41] generates sketch-based physics cues to guide diffusion synthesis. Although these methods yield physically consistent results, they rely on complex simulation pipelines at test time, limiting their scalability and flexibility.

The second direction pursues implicit physical control by embedding physics priors directly into diffusion models without explicit simulators. VLIPP [45] leverages vision-language models (VLMs) to plan motion trajectories for diffusion guidance, while PhysCtrl [37] introduces a learned point-cloud trajectory generator to control motion within pre-trained diffusion models. While these methods achieve controllability without simulation, they primarily emphasize kinematic guidance. Closest to our work is Force Prompting [11], which fine-tunes a video diffusion model on a 15K-video dataset annotated with force directions. Unlike their single-attribute setup, our method enables control over diverse physical attributes including friction, restitution, external force, and deformation through spatially aligned physical property maps, offering richer and more general forms of physical conditioning.

Reward Optimization for Video Generation. Reward-based optimization has become a key paradigm for aligning diffusion models with desired downstream behaviors in both image and video domains. Early methods such as ImageReward [43] and DFTM [9] learn differentiable reward functions that enable direct fine-tuning of diffusion models toward human or task-specific preferences. Prabhudesai et al. further advanced this idea to the video domain with VADER [29, 30], introducing gradient-based reward alignment for video diffusion models. More recently, Luo et al. [25] employed VLM feedback as differentiable rewards for guiding image generation, and Kumari et al. [20] demonstrated VLM-based optimization for image editing without paired supervision. In contrast, our work leverages VLM feedback specifically for physics-aware reward optimization, where targeted queries about friction, restitution, deformation, and applied forces guide a controllable video generation model toward physically consistent and interpretable outputs.

3. Method

Our goal in this work is to enable diffusion models with continuous and interpretable control over key physical properties—friction, restitution, deformation, and applied forces, while maintaining photorealistic synthesis and broad generalization. Achieving this requires both (i) a source of physically meaningful supervision that fairly generalizes beyond the training domain and (ii) a mechanism to inject such information into a pretrained generative model. We therefore introduce a unified pipeline that couples physically grounded simulation data with a two-stage training procedure: physics-supervised fine-tuning and VLM-guided reward optimization.

3.1. Physically Grounded Simulations

Although producing synthetic data with physics engines is straightforward, creating simulations that meaningfully benefit controllable video generation is far from trivial. The challenge is not the quantity of data but the quality and relevance of the physical behaviors it expresses. In practice, we find that simulations must satisfy two criteria to effectively teach controllable physics priors: (1) the physical attribute of interest must manifest clearly and unambiguously in the visual motion, and (2) the scenario must lie within the competence range of the pretrained diffusion backbone. Overly complex scenes—such as interactions involving multiple objects or cluttered dynamics—often produce outputs that current diffusion models struggle with, as also noted in recent studies such as PISA [21]. Training on such data introduces unnecessary variance and slows learning. In contrast, focused, well-structured simulations accelerate controllability by letting the model associate each physical property with its canonical visual signature, similar to the design philosophy behind Force-Prompting [11].

Following this insight, we construct a large suite of physics-rich but visually clean simulation scenarios using the Kubric framework [4, 10, 12]. Each simulation isolates or combines fundamental object–environment interactions while systematically varying four key physical parameters—friction, restitution, deformation, and applied force. These variations allow the model to observe consistent motion changes tied directly to interpretable physical attributes.

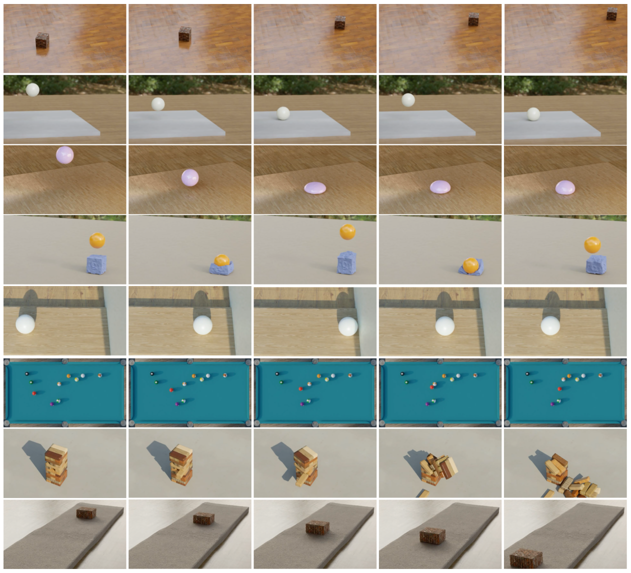


Figure 3. Example simulation videos from our dataset comprising of more than 100K videos from 8 different scenarios.

Our dataset spans six controlled scenarios designed to expose different dynamic regimes: a brick sliding on a flat plane, a ball rebounding off a wall, a vertically bouncing ball, a soft ball dropping under gravity, an object impacting a deformable body, and multiple balls colliding on a pool table. Each sequence is randomized in object color, surface material, camera placement, and HDRI illumination (50 environments following [11]), with high-quality textures from Polyhaven to enhance photorealism. This balance of controlled physics and appearance diversity helps the diffusion model disentangle visual variation from underlying dynamics. In total, we generate more than 100K videos across all scenes. The resulting dataset provides a structured yet diverse foundation for learning controllable, interpretable, and physically grounded video dynamics.

3.2. Physics Supervised Fine-tuning

The overall pipeline is illustrated in Fig. 2. We fine-tune a pre-trained Cosmos-Predict2-2B [28] diffusion backbone by conditioning the denoising process on physical property inputs using a ControlNet-style [49] architecture. Formally, the generator G_θ models the conditional distribution $p_\theta(\mathbf{x}_{1:T}|\mathbf{t}, \mathbf{x}_0^0, \mathbf{p})$, where $\mathbf{x}_{1:T}$ is the target video sequence, $\mathbf{t}$ the text prompt, $\mathbf{x}_0^0 \in \mathbb{R}^{C\times H\times W}$ the initial frame, and $\mathbf{p} \in \mathbb{R}^{K\times H\times W}$ encodes spatially aligned physical attributes.

Encoding Physical Property Maps. To enhance compactness and promote better generalization from limited training data, we represent objects as spatially aligned circular blobs. The ControlNet receives a tokenized representation of the property field $\mathbf{p}$ where each attribute is normalized between $[-1, 1]$. We decompose $\mathbf{p}$ into groups $\{\mathbf{p}^{(g)}\}_{g=1}^{G}$, where each $\mathbf{p}^{(g)} \in \mathbb{R}^{3\times H\times W}$ contains semantically related channels: (1) friction μ_f and restitution e, padded with a constant channel; (2) Neo-Hookean deformation parameters d_μ, d_λ, d_γ; and (3) force magnitude F and direction represented by $(\cos\phi, \sin\phi)$. Each group $\mathbf{p}^{(g)}$ is tokenized by the Cosmos tokenizer $\tau(\cdot)$, producing embeddings $\mathbf{z}^{(g)} = \tau(\mathbf{p}^{(g)}) \in \mathbb{R}^{L_g\times D}$, which are linearly projected to form the conditioning sequence.

Training. We finetune only the ControlNet layers while keeping the base diffusion model and tokenizer weights frozen to preserve their pretrained representations. The encoded physical property maps $\mathbf{z}^{(g)}$ are first passed through an adapter network $A(\cdot)$ that projects the tokenized property embeddings $\mathbf{h}_p = A(\mathbf{z}_p)$ to match the input dimensionality of the DiT backbone. Each semantic group $\mathbf{h}_p^{(g)}$ is processed by a separate ControlNet branch to enable faster training and compositionality of various physical properties. Each ControlNet is trained on a subset of PhyCo dataset where the corresponding input physical property manifests in the observed dynamics. The optimization follows the diffusion score-matching objective used in the Cosmos World Foundation Model [18, 19, 28], ensuring consistent noise scheduling and temporal supervision across frames.

3.3. VLM Reward Optimization

The second stage of our training leverages feedback from a foundational Vision–Language Model (VLM) to refine both the controllability and physical alignment of the generated videos. While supervised fine-tuning on physics-rich simulation datasets yields visually coherent and physically plausible results, it alone does not guarantee strong control fidelity. To bridge this gap, we employ a VLM as a generalized critique model that evaluates physical consistency through targeted, physics-aware queries. The VLM outputs token logits that are then converted into a differentiable reward signal, guiding the generator toward physically interpretable and controllable behaviors.

Approach. Standard diffusion training with score matching involves denoising a noised version of the ground-truth video. However, such single-step reconstructions are unsuitable for VLM evaluation due to two main reasons: (i) the visual details and object boundaries remain blurry, and (ii) they already encode the global physical trajectory from the conditioning signal (e.g., applied force direction), which masks the true inference-time behavior of the model. For instance, a partially noised ground-truth video of an object under an applied force still shows its dominant motion direction, even though the model might fail to reproduce it at inference time.

To obtain a faithful proxy of the inference-time generation process, we instead perform an N-step denoising rollout to generate a predicted latent $\hat{\mathbf{z}}_0$ given the initial frame $\mathbf{x}_0^0$, text prompt $\mathbf{t}$, and physical property maps $\mathbf{p}$. The latent is decoded to a video $\hat{\mathbf{x}}_0 \in \mathbb{R}^{T\times C\times H\times W}$, which serves as the input to the VLM along with a structured set of physics queries. Our formulation is inspired by recent VLM-based optimization works such as [20, 25], but differs in that our feedback focuses on physical controllability of videos rather than semantic or aesthetic alignment for image editing.

VLMs for Physical Question Answering. While off-the-shelf VLMs excel at recognizing explicit visual cues, they often struggle with implicit physical reasoning such as motion consistency, restitution, or frictional response. To improve reliability, we fine-tune Qwen2.5-VL-3B for 200 steps using our synthetic simulation dataset, where each clip is paired with multiple physics-related questions (e.g., "Does the object move in the intended direction of force?"). This short adaptation yields high accuracy ($\approx 85\%$) within 100 iterations, enabling robust reward computation.

Physical Alignment Reward. We structure each query as a binary ("Yes" / "No") question following prior work in vision-language reward learning [20, 25]. For each generated video $\hat{\mathbf{x}}_0$, we compute a reward by comparing VLM logits corresponding to the correct and incorrect answer tokens, guided by ground-truth property ranges $\mathbf{p} \in [0, 1]$. Each physical attribute (e.g., friction, deformation, restitution, force magnitude or direction) is probed with multiple thresholded questions over $\{\textit{min\_val}, \textit{max\_val}\}$ to obtain dense feedback signals. To assess directional consistency, we overlay a blue angular sector on the video and query whether the motion lies within this region.

We compute the VLM alignment loss as a binary cross-entropy over the logit difference between correct and incorrect answer tokens:

$$\mathcal{L}_{\text{VLM}} = -\sum_i \log\sigma(\zeta_+^{(i)} - \zeta_-^{(i)}), \tag{1}$$

where $\zeta_+^{(i)}$ and $\zeta_-^{(i)}$ are the logits for the correct and incorrect responses to the i-th question, respectively.

Training. We fine-tune only the ControlNet layers of our diffusion model corresponding to the physical property maps $\mathbf{p}$ using the VLM-based reward loss $\mathcal{L}_{\text{VLM}}$, excluding the diffusion score-matching objective. This focused optimization yields more stable and physically consistent generations compared to joint training with score matching. For meaningful feedback, we perform a 10-step denoising rollout, decode the latent into a video, and backpropagate the VLM loss end-to-end through the VLM, tokenizer, and DiT backbone.

Model	Solid Mechanics (↑)	Fluid Dynamics (↑)	Optics (↑)	Magnetism (↑)	Thermodynamics (↑)	IQ Score (↑)
SVD-XT [3]	21.9	20.5	6.8	8.4	17.1	19.1
LTX-Video-I2V [14]	30.2	29.8	15.9	13.2	8.4	26.8
SG-I2V [27]	34.6	31.2	15.9	13.1	8.4	29.7
Cogvideo-I2V-5B [16]	30.4	29.8	16.7	13.3	8.5	27.1
Cosmos-Predict2-2B [28]	31.7	25.2	26.2	9.1	16.9	27.7
VLIPP[45]	42.3	34.1	16.9	13.4	8.8	34.6
Test time extrapolated generation: 120 frames @ 24FPS						
Ours (Text only)	36.5	28.9	18.9	12.6	32.0	30.9
Ours (ControlNet)	42.3	30.7	19.3	12.6	40.1	35.3
Ours (ControlNet + VLM Loss)	44.1	31.2	20.1	17.2	33.1	36.3
Train-time conditions: 57 frames @ 24FPS + last-frame repetition						
Ours (Text only)	43.9	38.5	17.5	21.7	26.8	36.5
Ours (ControlNet)	49.7	37.8	16.3	19.9	18.2	38.9
Ours (ControlNet + VLM Loss)	53.1	44.3	20.3	20.8	35.9	43.6

Table 2. Quantitative results of physically plausible video generation on Physics-IQ Benchmark.

4. Experimental Results

In this section, we evaluate our approach across quantitative metrics, qualitative comparisons, user preferences, and ablation studies. Our experiments aim to answer four core questions about our method: (i) generating plausible physical dynamics, (ii) fine-grained controllability over physical attributes (iii) VLM reward optimization to enhance physical fidelity and (iv) generalization across scenarios and new objects. To faithfully answer these questions we compare our method against several state-of-the art video generation models across physics benchmarks and other carefully designed controlled experiments and studies.

Baselines. Our primary baselines include text-conditioned image-to-video world models such as Cosmos-Predict2 [28], CogVideoX-I2V-5B [46], SVD-XT [3] and LTX-Video-I2V [14]. We also evaluate a text-only fine-tuned variant of Cosmos-Predict2 trained on our proposed PhyCo dataset, to measure the benefit of physics-aware supervision. We further compare against two baselines closest to our approach, Force-Prompting [11] and VLIPP [45]. Force-Prompting improves controllability and physics awareness through force-specific supervision, whereas VLIPP uses a VLM to extract coarse motion trajectories (with bounding boxes) that guide the diffusion model toward physically consistent outputs.

Ours vs. Methods	Friction	Restitution	Deformation	Force
Force Prompting [11]	–	–	–	71.7%
CogVideoX-I2V-5B [46]	95.5%	100.0%	82.2%	91.1%
Cosmos-Predict2B [28]	100.0%	93.2%	91.3%	86.4%
Ours (Text only)	90.9%	67.4%	56.8%	58.7%

Table 3. User study results evaluated on the physical realism axis for different physical properties. Physical realism reports the percentage of pairwise preferences (%) for our method over the baselines in a 2AFC human evaluation, where scores above 50% indicate a preference for our ControlNet generations.

Method	FM	Fric.	FD (°)	Res.	Def.
Base Model (zero-shot) [28]	0.38	0.33	91.87	0.40	0.45
Text-only (finetuned)	0.31	0.30	40.35	0.31	0.14
ControlNet (–VLM)	0.33	0.24	38.05	0.28	0.14
ControlNet (+VLM)	**0.28**	**0.20**	**22.53**	**0.16**	**0.10**

Table 4. Ablation study on synthetic data across five controllable properties: FM = force magnitude error, Fric. = friction error, FD = angular deviation in force direction (lower is better), Res. = restitution error, and Def. = deformation error.

Quantitative Evaluation on Physics-IQ Benchmark. We first evaluate our method on the Physics-IQ [26] benchmark, which measures the physical realism of generated videos across five domains—Solid Mechanics, Fluid Dynamics, Optics, Magnetism, and Thermodynamics. The benchmark computes a *Physics-IQ* score by comparing the timing and spatial alignment of key actions in generated videos against real-world reference sequences (396 videos). Results on this benchmark are presented in Tab. 2. While the Physics-IQ benchmark evaluates 5-second videos (120 frames at 24 FPS), our model is trained on 57-frame sequences. Nevertheless, even under this train and test time mismatch, our model significantly outperforms several state-of-the-art open-source video generation systems across all evaluated categories. For completeness, we additionally report results obtained when we adhere to our training-time conditions by generating 57 frames and repeating the final frame to match the benchmark duration. Both these results paint a consistent picture: our physics-aware conditioning leads to more realistic and physically coherent dynamics, underscoring the benefit of the proposed PhyCo dataset and the robustness of our approach even when tested beyond the training domain.

User Study. We conduct a 2AFC study with 16 participants, each comparing 39 video pairs differing in a single physical attribute. Across 98 generated videos per method, users con-

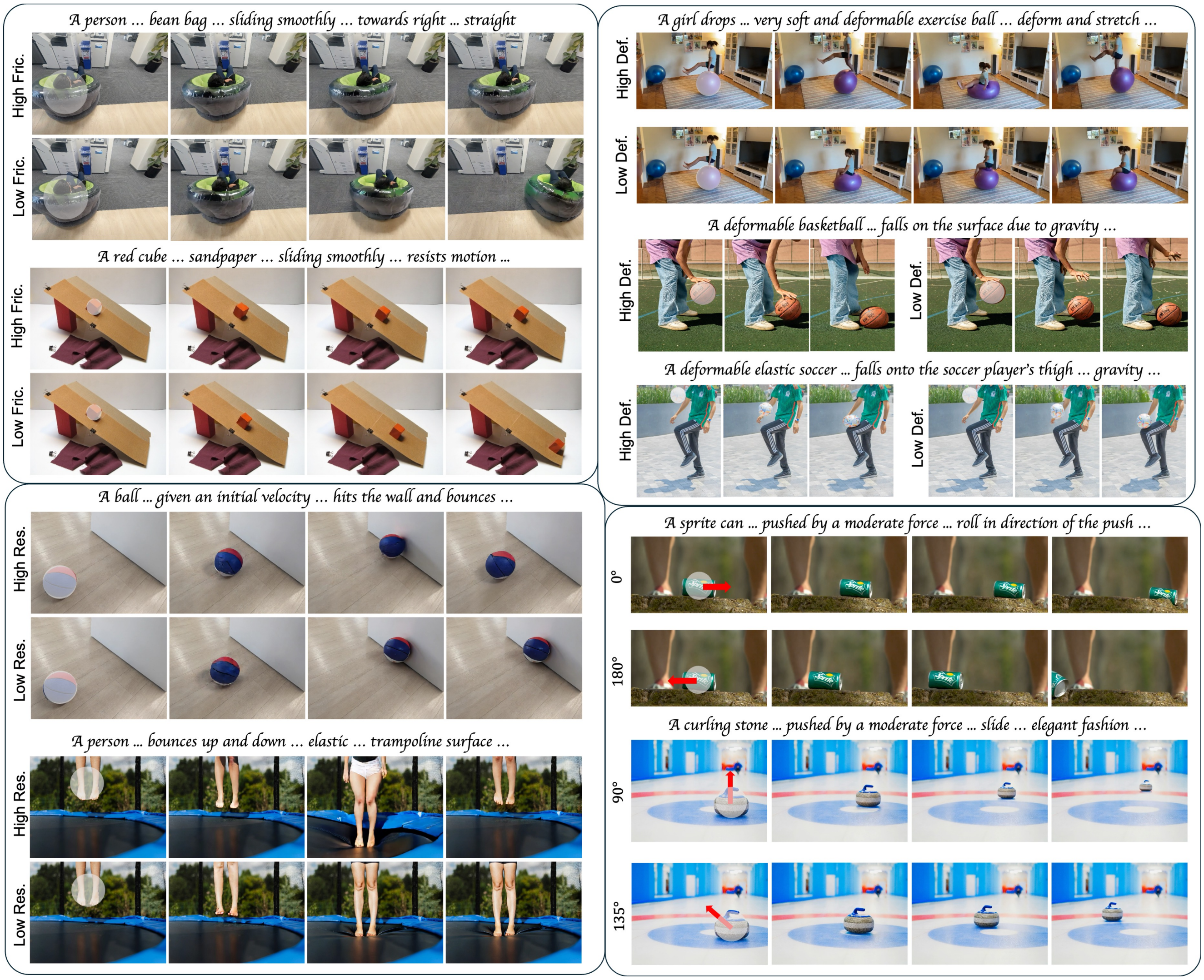

A person ... bean bag ... sliding smoothly ... towards right ... straight
High Fric.
Low Fric.
A red cube ... sandpaper ... sliding smoothly ... resists motion ...
High Fric.
Low Fric.
A girl drops ... very soft and deformable exercise ball ... deform and stretch ...
High Def.
Low Def.
A deformable basketball ... falls on the surface due to gravity ...
High Def.
Low Def.
A deformable elastic soccer ... falls onto the soccer player's thigh ... gravity ...
High Def.
Low Def.
A ball ... given an initial velocity ... hits the wall and bounces ...
High Res.
Low Res.
A person ... bounces up and down ... elastic ... trampoline surface ...
High Res.
Low Res.
A sprite can ... pushed by a moderate force ... roll in direction of the push ...
0°
180°
A curling stone ... pushed by a moderate force ... slide ... elegant fashion ...
90°
135°

Figure 4. Shows representative frames from our generated outputs, illustrating controllable behavior across key physical attributes including friction, restitution, deformation, and external forces. White blobs in the figure highlights the locations of spatially aligned pixel property inputs. These results highlights the ability of our model to generalize beyond the training domain for controllable physically consistent generation. For instance, the girl hopping on the exercise ball exhibits a noticeably higher bounce when the ball's deformability is increased, compared to the lower-deformation setting.

sistently prefer PhyCo in Table 3, indicating more realistic physical behavior and clearer variations. This confirms that our conditioning improves both controllability and perceived physical plausibility.

Ablations on PhyCo Data. To quantitatively assess how well generated videos adhere to the input physical properties, we conduct evaluations on an in-domain simulation test set consisting of 100 videos spanning all attributes. A fine-tuned Qwen2.5-VL-3B model [31] predicts physical properties from generated outputs, which we compare to the ground-truth conditioning inputs. Results of this analysis are presented in Tab. 4. As shown, models trained with explicit VLM-based reward optimization achieves significantly better alignment with the intended input properties, confirming that our reward formulation strengthens controllability and ensures more faithful adherence to physical conditioning during generation (Fig. 7).

Force Direction Adherence. We assess force-conditioned controllability on 25 real-world videos by applying random force directions and measuring the angular deviation between intended and observed motion. Our model achieves a substantially lower mean directional error (15.2°) compared to Force-Prompting (40.5°), indicating more reliable and precise control over induced dynamics. As shown in

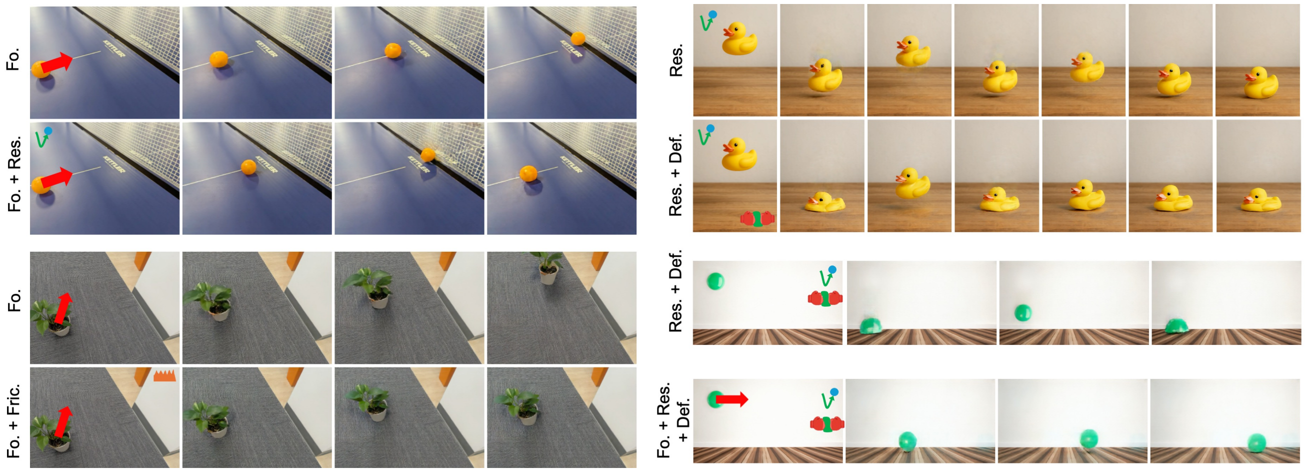


Figure 5. Qualitative results illustrating compositional control over multiple physical attributes within the same scene. Force, friction, restitution, and deformation are denoted as Fo, Fric, Res, and Def, respectively.

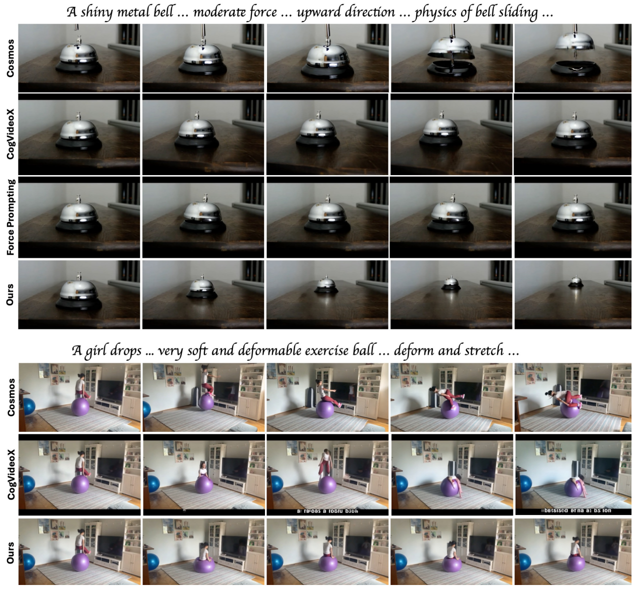

A shiny metal bell ... moderate force ... upward direction ... physics of bell sliding ...
Cosmos
CogVideoX
Force Prompting
Ours
A girl drops ... very soft and deformable exercise ball ... deform and stretch ...
Cosmos
CogVideoX
Ours

Figure 6. Qualitative comparison between our proposed method and other baselines. Note, without force inputs, Force-Prompting [11] is same as CogVideoX [46] model.

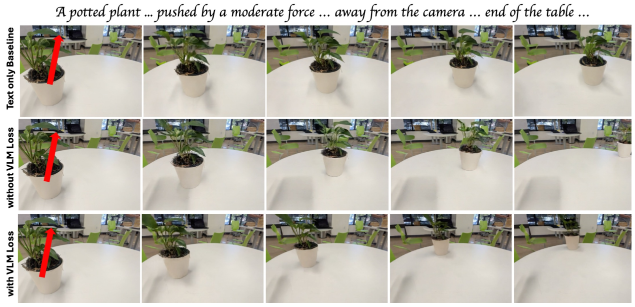

A potted plant ... pushed by a moderate force ... away from the camera ... end of the table ...
Text only Baseline
without VLM Loss
with VLM Loss

Figure 7. Ablation on explicit physical property conditioning using a ControlNet with additional VLM feedback.

Fig. 6, Force-Prompting fails to produce the desired motion—particularly in scenes where large or visually dominant objects are present, while our method consistently generates motion in the correct direction.

Qualitative Results. Representative snapshots highlighting controllability and compositionality over several physical attributes is shown in Figures 1, 4 and 5 respectively. Despite being trained solely on synthetic simulations, the model generalizes effectively to new object categories, motion types and appearance shifts (Fig. 1), exhibiting stable and coherent behavior under varying physical properties. For example, a model trained only on simple bouncing-ball simulations can generalize to a person jumping on a trampoline—where low restitution settings result in no rebound after impact. Likewise, training on simple blocks sliding across flat surfaces extends naturally to more complex object and surface configurations. These results demonstrate the strong generalization capabilities of our approach compared to the baselines (Fig. 6) and underscores the value of our simulated dataset in enabling physically consistent and controllable video generation.

5. Conclusion

We presented PhyCo, a controllable video generation framework that injects physically grounded priors into diffusion models through explicit property conditioning and VLM-guided reward optimization. Our approach enables continuous control over key dynamics such as friction, restitution, deformation and force, without requiring simulators at inference. Extensive evaluations show clear gains in both physical realism and controllability, with strong generalization beyond synthetic training data. These results point to a scalable path toward physically consistent, controllable video models that generalize reliably to real-world dynamics.

Acknowledgements: This work was partially conducted during Sriram's internship at NEC Labs America, and was supported in part by NSF grants IIS-2107236 and IIS-2513219. We also thank Kausik Sivakumar and Yug Ajmera for their insightful discussions.

References

[1] Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. *arXiv preprint arXiv:1909.12000*, 2019. 2

[2] Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021. 2

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 6

[4] Blender Online Community. Blender – a 3d modelling and rendering package. `http://www.blender.org`, 2018. 2, 4

[5] Victor Blomqvist. Pymunk. `https://pymunk.org`. 3

[6] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 13–23, 2025. 3

[7] Chuhao Chen, Zhiyang Dou, Chen Wang, Yiming Huang, Anjun Chen, Qiao Feng, Jiatao Gu, and Lingjie Liu. Vid2sim: Generalizable, video-based reconstruction of appearance, geometry and physics for mesh-free simulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26545–26555, 2025. 3

[8] Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Comphy: Compositional physical reasoning of objects and events from videos. *arXiv preprint arXiv:2205.01089*, 2022. 2

[9] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards, 2024. 4

[10] Erwin Coumans. Pybullet, a python module for physics simulation for games, robotics and machine learning. In *Proceedings of the ACM SIGGRAPH 2016 Talks*. ACM, 2016. 2, 4, 1

[11] Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals, 2025. 2, 3, 4, 6, 8

[12] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022. 2, 4

[13] Oliver Groth, Fabian B Fuchs, Ingmar Posner, and Andrea Vedaldi. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *Proceedings of the european conference on computer vision (eccv)*, pages 702–717, 2018. 2

[14] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. 6

[15] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation, 2024. 3

[16] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 6

[17] Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. The material point method for simulating continuum materials. In *ACM SIGGRAPH 2016 Courses*, New York, NY, USA, 2016. Association for Computing Machinery. 3

[18] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, pages 26565–26577. Curran Associates, Inc., 2022. 5, 1

[19] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24174–24184, 2024. 5, 1

[20] Nupur Kumari, Sheng-Yu Wang, Nanxuan Zhao, Yotam Nitzan, Yuheng Li, Krishna Kumar Singh, Richard Zhang, Eli Shechtman, Jun-Yan Zhu, and Xun Huang. Learning an image editing model without image editing pairs. *arXiv preprint arXiv:*, 2025. 4, 5

[21] Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop, 2025. 2, 4

[22] Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. Wonderplay: Dynamic 3d scene generation from a single image and actions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2025. 2, 3

[23] Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos, 2024. 2

[24] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. *PhysGen: Rigid-Body Physics-Grounded Image-to-*

Video Generation, page 360–378. Springer Nature Switzerland, 2024. 2, 3

[25] Grace Luo, Jonathan Granskog, Aleksander Holynski, and Trevor Darrell. Dual-process image generation, 2025. 4, 5

[26] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles?, 2025. 2, 6

[27] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B. Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation, 2025. 6

[28] NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. 2, 4, 5, 6, 1

[29] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation, 2024. 4

[30] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients, 2024. 4

[31] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. 7, 2

[32] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018. 2

[33] Xiyang Tan, Ying Jiang, Xuan Li, Zeshun Zong, Tianyi Xie, Yin Yang, and Chenfanfu Jiang. Physmotion: Physics-grounded dynamics from a single image, 2024. 3

[34] Maham Tanveer, Yang Zhou, Simon Niklaus, Ali Mahdavi Amiri, Hao Zhang, Krishna Kumar Singh, and Nanxuan Zhao. Motionbridge: Dynamic video inbetweening with flexible controls. *arXiv preprint arXiv:2412.13190*, 2024. 3

[35] Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Joshua B. Tenenbaum, Daniel LK Yamins, Judith E Fan, and Kevin A. Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties, 2023. 2

[36] Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any trajectory instruction for controllable video generation, 2025. 3

[37] Chen Wang*, Chuhao Chen*, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. Physctrl: Generative physics for controllable and physics-grounded video generation. In *NeurIPS*, 2025. 2, 3

[38] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024. 3

[39] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024. 3

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

[41] Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang. Physanimator: Physics-guided generative cartoon animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 10793–10804, 2025. 3

[42] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 3

[43] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, pages 15903–15935. Curran Associates, Inc., 2023. 4

[44] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3

[45] Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, and Xu Jia. Vlipp: Towards physically plausible video generation with vision and language informed physical prior, 2025. 2, 3, 6

[46] 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. 6, 8

[47] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 2

[48] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023. 2, 4

[50] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. PhysDreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*. Springer, 2024. 2, 3

[51] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. 3

[52] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In *European Conference on Computer Vision*, pages 407–423. Springer, 2024. 3

PhyCo: Learning Controllable Physical Priors for Generative Motion

Supplementary Material

A. Video Results on Webpage

All video results are available in the supplementary webpage. Please open *index.html* using a Chromium-based browser. The webpage includes both the examples shown in the main paper and additional results not included due to space constraints. We organize the content as follows:

Generalization Across Artistic Styles. We present multi-style storylines where four key frames guide the model to generate physically consistent sequences. Figure 8 shows the first frames used in these examples.

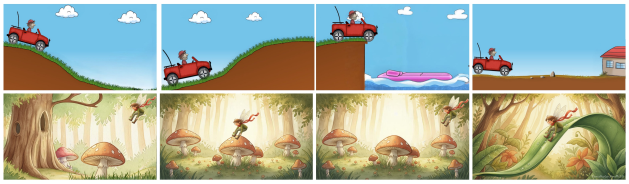


Figure 8. Shows the first frames that were used in the generated video storyline for two different examples

Fine-Grained Control of Physical Properties. We provide videos generated under three distinct levels (low, medium, high) for each physical attribute to demonstrate smooth and continuous control.

Compositionality of Multiple Attributes. We showcase combinations such as force+friction, force+bounciness, and bounciness+deformation. Even though some parameter pairings (e.g., restitution and deformation) are difficult to simulate accurately in current physics engines [10], our model still achieves visually convincing behavior.

Baseline Comparisons. Side-by-side videos compare our approach with recent video diffusion models. We also include interactive control over force direction with pre-generated results.

B. PhyCo Dataset Details

The PhyCo dataset consists of physically simulated scenes in PyBullet and photorealistic rendering in Blender. Each sample is a 4-second video (24 FPS) at 432×768 resolution. Alongside RGB frames, we provide synchronized depth maps, per-frame segmentation masks, and structured metadata describing scene geometry, object material properties, and all applied physical parameters. We also include a standardized scene-level text description for every video to support multi-modal supervision and VLM-based evaluation.

Sampling Physical Properties. We vary the key physical parameters that govern object dynamics and deformation. For rigid bodies, friction and restitution coefficients are uniformly sampled between 0 and 1 in PyBullet [10], covering behaviors from smooth sliding to high resistance, and from inelastic impacts to highly bouncy interactions.

For deformable bodies, we use a FEM-based Neo-Hookean model, varying the Lamé coefficients μ and λ along with a damping coefficient γ to span materials from nearly rigid to highly deformable. We also simulate a range of external forces, where low to high magnitudes map to gentle interactions and strong impacts. These forces are projected from 3D world coordinates onto the rendered 2D frame using the camera parameters, enabling direct correspondence between physical actions and visual outcomes.

C. Implementation Details

This section summarizes implementation specifics for the proposed PhyCo model, including ControlNet fine-tuning and VLM-based supervision, as well as Qwen2.5-VL fine-tuning used for evaluation.

C.1. PhyCo Implementation Details

ControlNet Training. We fine-tune only the ControlNet layers, while keeping the base video diffusion model and tokenizer weights frozen to preserve the pretrained dynamics learned by the Cosmos World Foundation Model [28]. Training is performed using 4×H100 GPUs for 10k optimization steps (approximately half a day) per attribute-specific ControlNet branch. We supervise 57 frames per sequence at 24 FPS, employing a per-device batch size of 1 with 2 steps of gradient accumulation, yielding an effective batch size of 8. We use a learning rate of $2^{-14.5}$ and maintain a peak memory footprint of approximately 45 GB per GPU. Optimization follows a standard diffusion score-matching loss [18, 19] with consistent noise scheduling and temporal supervision.

VLM-Based Reward Optimization. To incorporate physics-aware perceptual supervision, each ControlNet branch is further trained with a VLM-guided reward loss for 100 iterations (roughly 70 minutes). Videos are spatially downsampled to half resolution and temporally subsampled to a maximum of 16 frames before being fed to the VLM. This configuration uses 8×H200 GPUs with an effective batch size of 4 and requires up to 115 GB VRAM. Reducing the number of generated frames or directly aligning DiT latents with the VLM input would further lower memory needs—an avenue we leave for future work.

C.2. Qwen2.5-VL Fine-tuning for Evaluation

We adapt Qwen2.5-VL-3B [31] on the PhyCo dataset to robustly infer physical properties from video inputs. Training samples are generated using the physics-focused queries listed in Fig. E. For all binary (Yes/No) questions, we ensure a balanced set of responses. Videos are temporally subsampled to at most 16 frames. For force-direction queries, we add a highlighted blue sector overlay indicating the target force angle (see Fig. 13). We fine-tune for 200 iterations using LoRA with rank and α set to 64, across 4×H100 GPUs with an effective batch size of 128 and learning rate of 2×10^{-4}.

D. Additional Results

Motion Consistency Evaluation. We further evaluate motion consistency using the Fréchet Video Motion Distance (FVMD) [23] on Physics-IQ benchmark videos (Table 5). FVMD measures the distributional distance between generated and reference motion features, capturing temporal dynamics independently of appearance quality, with lower values indicating more realistic motion.

Our method achieves the best or second-best FVMD scores across most domains, demonstrating improved temporal coherence and physically plausible motion. Notably, incorporating VLM-based reward optimization consistently yields further gains over ControlNet without VLM, particularly in solid mechanics, fluid dynamics, and magnetism. These trends closely mirror the Physics-IQ results, indicating strong alignment between perceptual physics reasoning and motion statistics. Minor deviations are observed in thermodynamics, likely due to the limited number of evaluation scenes in this domain.

Methods	S.M.	F.D.	Opt.	Mag.	Therm.
Base Model (zero-shot)	4676.9	3277.9	3200.0	1586.8	**1618.9**
Text-only (finetuned)	3565.8	1782.6	4486.4	1164.0	1736.6
ControlNet(-VLM)	2340.0	1223.9	**2991.0**	646.2	2164.0
ControlNet(+VLM)	**2337.7**	**1223.1**	3032.9	**643.7**	2132.6

Table 5. FVMD-based comparison [23] of the proposed methods against base model on Physics-IQ benchmark videos.

Generalization Across Backbones. To demonstrate that our dataset enables physically consistent generation beyond a specific architecture, we finetune a Wan2.2 video model using only text conditioning. As shown in Table 6, finetuning the Wan2.2 base model on the proposed PhyCo dataset yields a 4.6% improvement in average on Physics-IQ score. This gain highlights the effectiveness of PhyCo in imparting physically meaningful priors, even without explicit conditioning mechanisms such as ControlNet. Figure 10 further illustrates qualitative examples, where the finetuned Wan2.2 model exhibits controllable physical behavior.

Qualitative Results on Physics-IQ Dataset. Figure 9 presents qualitative comparisons from the Physics-IQ [26] benchmark, demonstrating that our method produces dynamics more consistent with real-world physical behavior. For instance, when a ball momentarily occludes behind a bag, it reappears with a coherent trajectory and speed, reflecting improved temporal consistency in motion prediction. Likewise, scenes with contact-induced deformation show physically plausible responses—such as a pillow compressing noticeably under the weight of a kettlebell while remaining largely unaffected by a lightweight paper object. These examples highlight how explicit physical conditioning leads to more realistic and physically interpretable video synthesis across diverse scenarios.

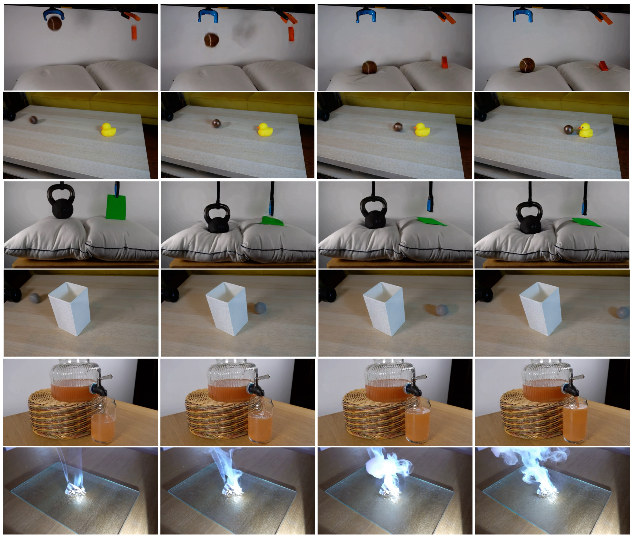


Figure 9. Qualitative results from Physics-IQ [26] benchmark.

Methods	S.M.	F.D.	Opt.	Mag.	Therm.	Avg.
Wan2.2 (zero-shot)	34.3	35.2	18.1	10.7	**36.0**	30.5
PhyCo finetuned	**42.1**	**37.6**	**21.9**	**12.2**	22.1	**35.1**

Table 6. Quantitative evaluation (120f @ 24FPS) on Physics-IQ benchmark with text only LoRA finetuning on Wan2.2 base model.

Results from VLM fine-tuning. We test the ability of the fine-tuned VLM to predict intrinsic physical physical property values from videos by testing it on a held out PhyCo test set of 100 samples. We find the mean absolute error across all four attributes (friction, restitution, deformation and force) to be 0.14 . Further, the accuracy of prediction for binary responses to be 84.8% across all four attributes.

Analysis of Flickering Artifacts. We observe that flickering primarily arises in regions with rapid per-frame motion, especially for thin or high-frequency structures. Increasing the training frame rate significantly mitigates these artifacts (Fig. 11), as it provides denser temporal supervision and reduces abrupt motion discontinuities. In addition, stronger

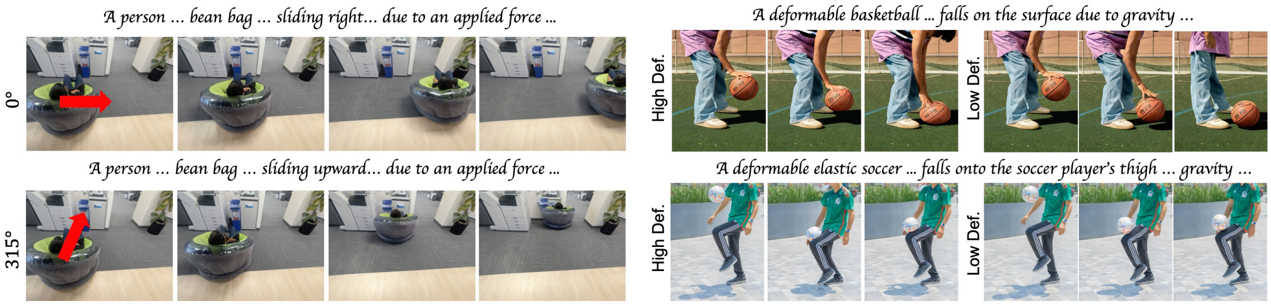

A person ... bean bag ... sliding right... due to an applied force ...
0°
A person ... bean bag ... sliding upward... due to an applied force ...
315°
A deformable basketball ... falls on the surface due to gravity ...
High Def.
Low Def.
A deformable elastic soccer ... falls onto the soccer player's thigh ... gravity ...
High Def.
Low Def.

Figure 10. Controllability results from Wan2.2 LoRA model trained with text-only conditioning using the proposed PhyCo dataset.

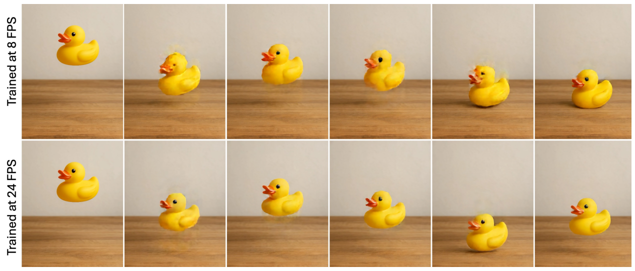
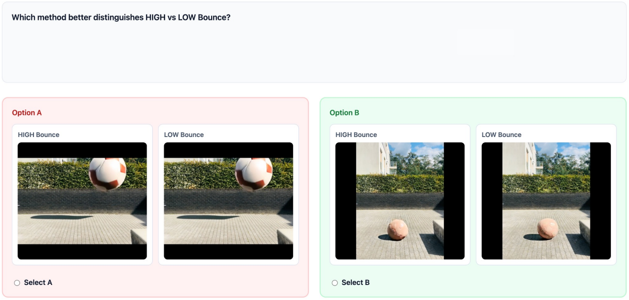


Figure 11. Illustration highlighting the impact of training FPS towards flickering artifacts on Cosmos-Predict base model.


Which method better distinguishes HIGH vs LOW Bounce?

Option A

HIGH Bounce

LOW Bounce

○ Select A

Option B

HIGH Bounce

LOW Bounce

○ Select B

Figure 12. User study interface. Participants were shown two videos for each question and asked to choose which method better expresses the specified physical property.

video diffusion backbones such as Wan2.2 exhibit markedly reduced flickering (Fig. 10), suggesting that improved temporal modeling further enhances stability.

These observations are consistent with our quantitative results: improvements in motion quality are reflected in lower FVMD scores (Tab. 5), indicating better temporal coherence. Overall, both higher-FPS training and advances in backbone architectures play a complementary role in reducing flickering and improving motion consistency.

Details on User-study. Illustration of the interface used in conducting the user-study is shown in Fig. 12

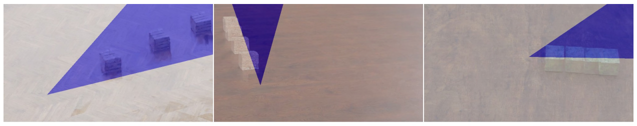


Figure 13. Figure illustrating denoised network outputs with a blue sector overlay indicating the direction of the applied force. The visualization shows the temporal average across video frames.

E. Limitations

Although our approach improves controllability and physical consistency over existing video diffusion models, the generated dynamics are still an approximation of real physics rather than an accurate reproduction. Our physical priors primarily capture simplified rigid and soft-body behaviors in controlled settings, and more complex interactions—such as articulated motion, fluid-structure coupling, or multi-contact dynamics—remain partially modeled. Additionally, while spatial property maps provide interpretable control, they do not enforce strict adherence to underlying conservation laws (e.g., momentum, deformation energy), occasionally producing subtle but noticeable physical deviations. Extending our framework toward richer physical regimes, stronger real-world grounding, and multi-object interactions represents a key direction for future work.

VLM Fine-Tuning Prompts

Force (Sector Adherence)

System prompt: You are a physics expert analyzing object’s motion in a video.

User prompt: I’m showing you a video with a blue highlighted sector region overlaid on all frames. The blue region is a sector that is a few degrees wide. I want you to carefully observe the video and answer the following question: Does the object’s movement lie within the blue highlighted sector region? Your answer should be ’Yes’ or ’No’ only.

Force Range Yes/No

System prompt: You are a physics expert analyzing object’s motion in a video.

User prompt: I’m showing you a video with an object sliding on a surface. I want you to carefully observe the object’s motion in thevideo and answer the following question: Is the force applied to the object between {min_value} and {max_value}? Your answer should be ’Yes’ or ’No’ only.

Friction Range Yes/No

System prompt: You are a physics expert analyzing object’s motion in a video.

User prompt: I’m showing you a video with an object sliding on a surface. The surface has some roughness and is only observable based on the object’s sliding motion. I want you to carefully observe the object’s motion in thevideo and answer the following question: Is the friction between the object and the surface between {min_value} and {max_value}? Your answer should be ’Yes’ or ’No’ only.

Restitution Range Yes/No

System prompt: You are a physics expert analyzing object’s motion in a video.

User prompt: I’m showing you a video with an object bouncing on a surface. The object’s bounciness is observable based on the object’s bouncing motion. I want you to carefully observe the object’s motion in the video and answer the following question: Is the bounciness of the object between {min_value} and {max_value}? Your answer should be ’Yes’ or ’No’ only.

Deformability Range Yes/No

System prompt: You are a physics expert analyzing object’s motion in a video.

User prompt: I’m showing you a video with a deformable object dropping on a surface. I want you to carefully observe the object’s deformation in the video and answer the following question: Is the deformability of the object between {min_value} and {max_value}? Your answer should be ’Yes’ or ’No’ only.

Force Magnitude JSON

System prompt: You are a physics expert analyzing object’s motion in a video. You must respond in valid JSON format only.

User prompt: I’m showing you a video with an object sliding on the surface. Carefully observe the object’s motion and estimate the magnitude of the force applied to the object. Respond with a JSON object in this exact format: {”value”: X} where X is a number between 0 and 1.

Friction Magnitude

System prompt: You are a physics expert analyzing object’s motion in a video. You must respond in valid JSON format only.

User prompt: I’m showing you a video with an object sliding on a surface. The surface has some roughness observable from the object’s sliding motion. Ignore the object’s visual appearance and focus on its motion to estimate the friction coefficient between the object and the surface. Respond with a JSON object in this exact format: {”value”: X} where X is a number between 0 and 1.

Restiution Magnitude

System prompt: You are a physics expert analyzing object’s motion in a video. You must respond in valid JSON format only.

User prompt: I’m showing you a video with an object bouncing on a surface. The object’s bounciness is observable from its bouncing motion. Ignore the object’s visual appearance and focus on its motion to estimate the coefficient of restitution between the object and the surface. Respond with a JSON object in this exact format: {”value”: X} where X is a number between 0 and 1.

Deformation Magnitude

System prompt: You are a physics expert analyzing object’s motion in a video. You must respond in valid JSON format only.

User prompt: I’m showing you a video with a deformable object dropping on a surface. Ignore the object’s visual appearance and carefully observe its deformation behavior to estimate how deformable the object is. Respond with a JSON object in this exact format: {”value”: X} where X is a number between 0 and 1.