PhyCo: Learning Controllable Physical Priors for Generative Motion

Supplementary Material

A. Video Results on Webpage

All video results are available at phyco-video.github.io. 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.

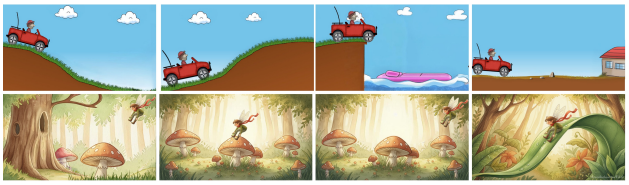

Figure 8. Shows the first frames that were used in the generated video storyline for two different examples

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.

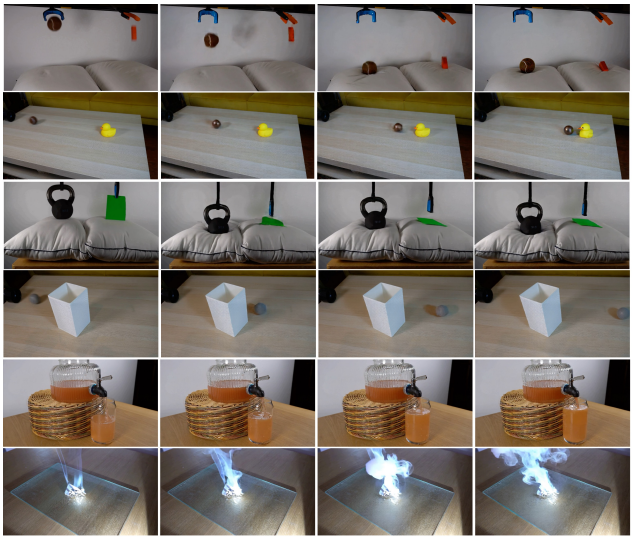

Figure 9. Qualitative results from Physics-IQ [26] benchmark.

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

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

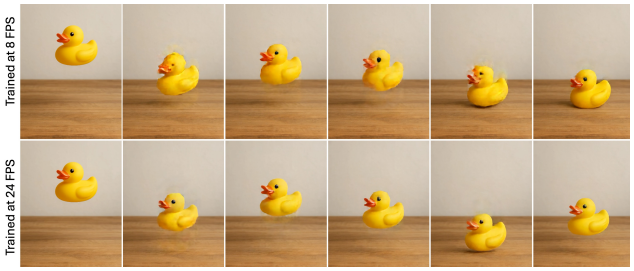

Figure 11. Illustration highlighting the impact of training FPS towards flickering artifacts on Cosmos-Predict base model.

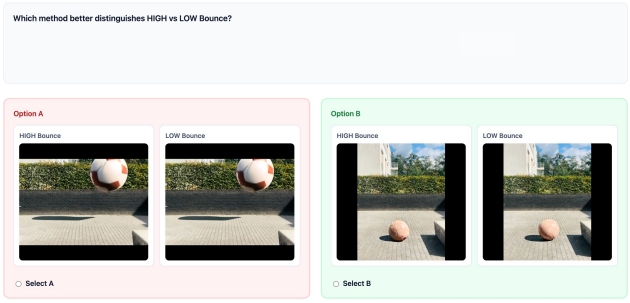

Figure 12. User study interface. Participants were shown two videos for each question and asked to choose which method better expresses the specified physical property.

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.