

ProPhy: Progressive Physical Alignment for Dynamic World Simulation

Supplementary Material

In this supplementary material, we first provide the full implementation details of ProPhy (Section A). We then present a detailed analysis showing how the key modules acquire and utilize physical knowledge, together with additional qualitative ablation results and comparisons with prior methods (Section B). Finally, we discuss the limitations of our approach (Section C) and its potential social impact (Section D). Additional raw video results are included in the supplementary video file. All code will be released after the final publication.

A. Implementation Details

A.1. Model Architecture and Settings

We build our model on top of Wan2.1-T2V-1.3B and CogVideoX-5B. For the PB module without REB, its structure and initialization directly follow the corresponding Transformer layers in the backbone. Due to practical GPU memory and runtime considerations, in the 30-layer Transformer of Wan2.1-T2V-1.3B we reuse Blocks [0, 7, 14, 21, 28], and in the 42-layer Transformer of CogVideoX-5B we reuse Blocks [0, 9, 18, 27, 36]. As described in the main paper, REB is only attached to the last PB layer. To avoid unnecessary parameter inflation in the MoE implementation of REB, we set each expert’s hidden size to match the hidden states. With this configuration, the total additional parameters are approximately 31.3% for Wan2.1-T2V-1.3B and 19.4% for CogVideoX-5B.

For computational efficiency, we set the number of physical basis maps B_e in SEB to $E_s = 32$, with each map sharing the same dimensionality as the hidden states of the corresponding model. We also set the number of refinement experts in REB to $E_r = 32$, and implement the refinement router using a top-4 selection. This configuration provides a balanced trade-off between memory usage and computational cost.

Regarding the loss-weight hyperparameters in the mixed objective of Equation 6, we use $\lambda_1 = 0.1$, $\lambda_2 = 0.02$, and $\lambda_3 = 0.01$ across all experiments. These values follow standard practice in balancing multi-term objectives and work reliably to stabilize convergence in our setting.

For the Router design, since the text encoder is typically frozen during finetuning and the forward path of the Semantic Router contains no additional trainable components, we implement it using a lightweight MLP to provide adequate capacity for distinguishing physical categories. In contrast, because the Refinement Router already receives rich trainable features from earlier layers, we adopt a simple Linear layer implementation, which is sufficient and avoids unne-

cessary overhead.

A.2. Dataset and Annotation Protocol

We randomly sample 20K videos from the WISA-80K dataset as our training data. Since quantitative physical categories are generally harder to annotate reliably, we only use the qualitative physical taxonomy as our alignment target. During Semantic Alignment, we adopt the complete set of physical categories from WISA-80K, i.e., $E_{\text{wisa}} = 29$ as described in the main paper. For Fine-grained Alignment, some categories in WISA-80K describe *absence* of physical phenomena (e.g., “no obvious dynamic phenomenon”), which are not suitable for defining fine-grained physical attributes. We therefore remove such categories and use $E_{\text{attn}} = 23$ attributes for fine-grained alignment. The retained attributes include:

- **Physical phenomena:** rigid body motion, collision, liquid motion, gas motion, elastic motion, deformation, melting, solidification, vaporization, liquefaction, combustion, explosion, reflection, refraction, scattering, interference and diffraction, unnatural light source
- **Physical appearances:** liquid objects, solid objects, gas objects, object decomposition and splitting, mixing of multiple objects, object disappearance

In summary, we utilize videos, text descriptions, and qualitative physical-category annotations from WISA-80K. Compared with general video datasets, our usage mainly involves incorporating the qualitative physical taxonomy. For any standard video dataset, similar qualitative physical labels can be obtained using Language Models or Vision-Language Models. Therefore, our training pipeline does not rely heavily on WISA-80K and can be adapted to other datasets without structural changes.

A.3. Training Details

We conduct all experiments using four NVIDIA H100 GPUs with 80GB memory each, and train for a fixed 8,000 steps across all settings. The learning rate is set to $1e-4$, and we only update the SEB, PB, and the REB modules. For Wan2.1-T2V-1.3B, the input video resolution is 480×832 with 81 frames at 16 fps. For CogVideoX-5B, the input resolution is 480×720 with 49 frames at 8 fps. We adopt DDP for distributed training with a per-GPU batch size of 4 (no gradient accumulation), resulting in a total batch size of 16. We use the AdamW optimizer with a *cosine with restarts* learning-rate schedule. CFG dropout strategy is enabled during training, and the dropout probability on text conditioning is set to 0.1.

To maintain stability, PB is initialized from the corresponding Transformer Blocks in the backbone, and the projection from PB to the input layer is initialized to zeros. This prevents undesirable interference with the pretrained backbone at the early stage of training.

A.4. Inference Details

For inference, we use DDIM with 50 sampling steps. CFG is enabled by default. We measure inference time using only the forward pass of the Transformer modules. The added components introduce an overhead of 20.3% on Wan2.1-T2V-1.3B and 11.5% on CogVideoX-5B relative to their respective baselines. ProPhy integrates physical-category reasoning directly into the routers and the model. As a result, the entire inference pipeline is fully end-to-end and does not rely on external models to provide physical priors. In practical use, the extra cost is partly offset because no external physical-prediction VLMs/LLMs are needed.

A.5. Evaluation Details

As described in the main paper, we evaluate generated video quality using VideoPhy2 and VBench. For consistency across all experiments, we use the 600 *upsampled caption* prompts provided by VideoPhy2 as the unified refined text input for video generation.

B. Analysis and Ablation

B.1. Relative Physical Semantic Analysis

To further verify that our Router learns the physical principles behind different phenomena, we collect a set of prompts from the WISA-80K dataset that never appear during training. For each physical phenomenon defined in the Semantic Router, we randomly sample 100 unseen prompts. We pass these prompts through the text encoder and the Semantic Router to obtain a series of logits. Each logit vector has a dimension of 32, corresponding to the number of physical basis maps E_s . We provide in the main paper the logits distributions for four representative phenomena. For the complete set, direct visualization with histograms is not intuitive. We therefore apply principal component analysis (PCA) to project the 32-dimensional logits into a 2D space. This projection allows for a clearer comparison of their relative relationships.

Figure 1 shows the resulting 2D PCA plot. We use dashed boxes to highlight the physical macro-categories of each phenomenon. As shown, the three macro-categories form compact clusters with limited overlap. This pattern indicates that the Semantic Router has captured meaningful relationships among the phenomena. Otherwise, the distributions would appear random or uniform. We also observe that liquid motion and scattering lie close to each other, even though they belong to different categories. This proximity

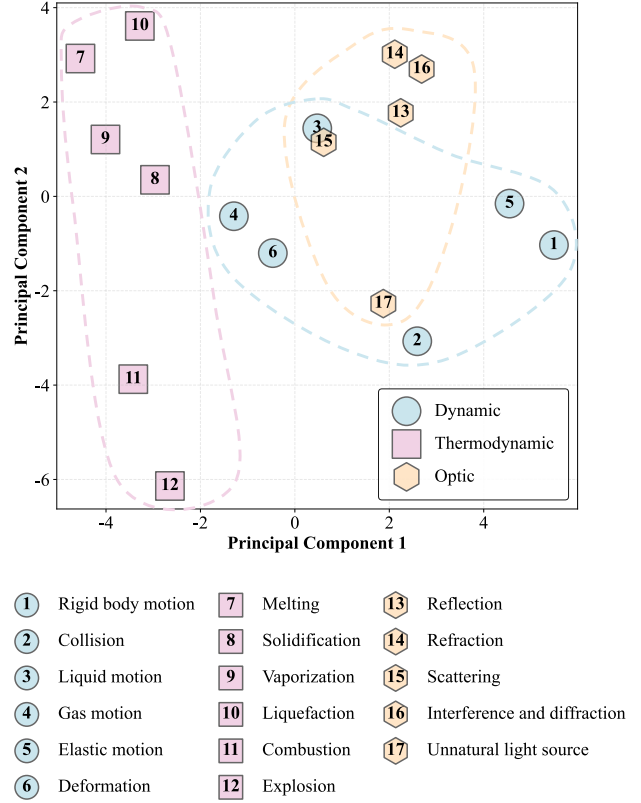


Figure 1. Principal component analysis of the activation distribution of the Semantic Router under different input prompt categories.

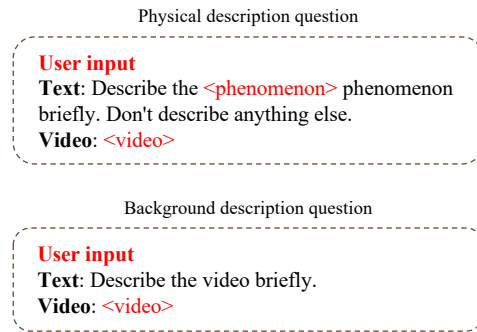


Figure 2. Details of the two types of user inputs used to obtain token-level physical properties annotations. The angle brackets '<>' are replaced with the specified physical phenomenon or the given video.

appears mainly in flowing-water videos, where splashing droplets often scatter light. This observation further suggests that the Semantic Router has learned a reasonable and physically meaningful correlation between these two phenomena.

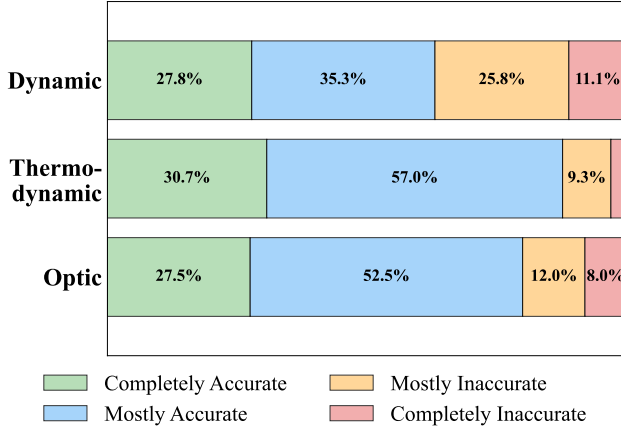


Figure 3. Human analysis of fine-grained physical annotation accuracy.

B.2. Annotation Details and Analysis

We propose a method for obtaining token-level physical attributes by computing attention maps over the answers to physical-description questions and background-description questions, as detailed in Section 3.3. Figure 2 illustrates how we obtain these answers through simple user instructions. We also experimented with using more elaborate prompts to induce longer responses from the VLM, but found that the resulting attention maps were essentially indistinguishable from those derived from shorter answers. Therefore, for efficiency, we request relatively concise responses, typically around 30–50 words.

To extract token-level physical attributes, we first use the prompts in Figure 2 to generate an answer. We then locate the answer tokens and video tokens in the token sequence, where the answer length is S_a and the video-token length is S_v . After passing them through the corresponding projection layers and computing scaled dot-product attention, the resulting attention map has a shape of $[N_{\text{vlm}}, S_a, S_v]$, where N_{vlm} denotes the number of decoder layers in the VLM. We average over the query dimension to treat the entire answer as a single unit, and further average across all layers. This produces an attention map of shape $[S_v]$. Since S_v corresponds to the number of video tokens, it can be reshaped into the three-dimensional form $[F/r'_t, H/r'_s, W/r'_s]$, representing the compressed number of frames, height, and width. Here (F, H, W) denote the original video length, height, and width, whereas (r'_t, r'_s) represent the temporal and spatial downsampling ratios used by the VLM. For example, in the case of Qwen2.5-VL-32B, the spatial downsampling ratio is $r'_s = 14$, and the temporal downsampling ratio is computed based on video duration rather than frame count. To standardize processing, we rescale all videos to a duration of 6 seconds to match the VLM’s sampling rate of two frames per second. The video generation models

Wan2.1-1.3B and CogVideoX-5B both use video encoders with temporal downsampling $r_t = 4$ and spatial downsampling $r_s = 8$, which introduces a size mismatch between the alignment target and the video model features. We subtract the attention map obtained from physical prompts from that obtained from background prompts to produce the diff attention map, which is then upsampled via tricubic interpolation to match the larger hidden-state resolution, followed by mild smoothing to fill minor gaps. Because sign-based filtering of the diff attention map still leaves some noise, such as regions that do not correspond to the actual physical phenomenon, we limit the alignment supervision during training to at most 10% of the tokens in the diff map. These selected tokens are used to compute $\mathcal{L}_{\text{fine-align}}$, which encourages the model to focus on the correctly annotated regions.

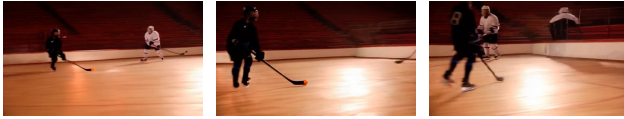
To assess the accuracy of the fine-grained physical annotations, we sample 100 videos for each phenomenon together with their per-frame diff attention maps. We then evaluate their correctness through human qualitative inspection, as illustrated in Figure 3. We define four accuracy levels. *Completely Accurate* indicates that the VLM-annotated regions almost perfectly match the true physical-phenomenon areas. *Mostly Accurate* means that the annotated regions largely overlap with the true areas but contain a small amount of activation outside them. *Mostly Inaccurate* and *Completely Inaccurate* refer to cases where the overlap is small or nonexistent. Overall, combining the completely and mostly accurate cases, our annotation method achieves an accuracy of 76.9%. Thermodynamics and optics reach 87.7% and 80.0%, respectively. The accuracy for dynamics is lower, at 63.1%. We attribute this gap to the subtle nature of many dynamic phenomena, which often occupy small regions in the video and are therefore more difficult to capture.

B.3. Additional Ablation Study

To assess the practical impact of the added SEB and REB modules, we conduct a qualitative ablation study using Wan2.1-1.3B as the baseline, as shown in Figure 4. In the first “ball-passing” scenario, the baseline model fails to correctly understand the interaction between the ball and the stick: the orange ball appears embedded inside the stick, accompanied by noticeable artifacts. With SEB added, these artifacts are alleviated, and the model becomes able to distinguish the orange ball from the stick; however, the small ball is not consistently maintained and disappears in an unnatural manner shortly afterward. When REB is added on top of SEB, ProPhy successfully completes the passing motion while preserving the ball’s shape and its physical interactions throughout the sequence.

In the second “pouring syrup” scenario, the baseline model does not generate the downward flow of syrup af-

A player uses their stick to push the ball around another player...



Baseline



Only SEB



SEB + REB

Syrup is poured onto pancakes from a bottle...



Baseline



Only SEB



SEB + REB

Figure 4. Qualitative ablation analysis on the functional roles of each module.

ter excessive accumulation on the pancakes. Adding SEB enables the correct downward-flow behavior, but without fine-grained alignment, unnatural liquid accumulation appears in regions the syrup has not touched. With REB further incorporated, the model produces a more dynamic syrup-pouring motion without any violations of gravity or fluid-flow regularities. These results clearly demonstrate that SEB enhances the generation of global physical dynamics, while REB further refines the model’s ability to capture fine-grained physical behaviors.

B.4. More Qualitative Results

To further demonstrate the ability of our model to generate videos that follow fine-grained physical dynamics, we com-

pare ProPhy with several prior state-of-the-art methods. As shown in Figure 5, under the same backbone, ProPhy produces videos that better align with physical commonsense, whereas other models often exhibit issues such as inconsistent object shapes, incorrect collision behaviors, or unnatural particle effects. To showcase the generalizability of our architecture and its capacity to model classical physical phenomena, we also present several visual results based on Wan2.1, as shown in Figure 6. These examples demonstrate that our model can reliably handle scenarios involving multiple interacting physical processes, while maintaining strong visual quality and physical plausibility.

C. Limitations

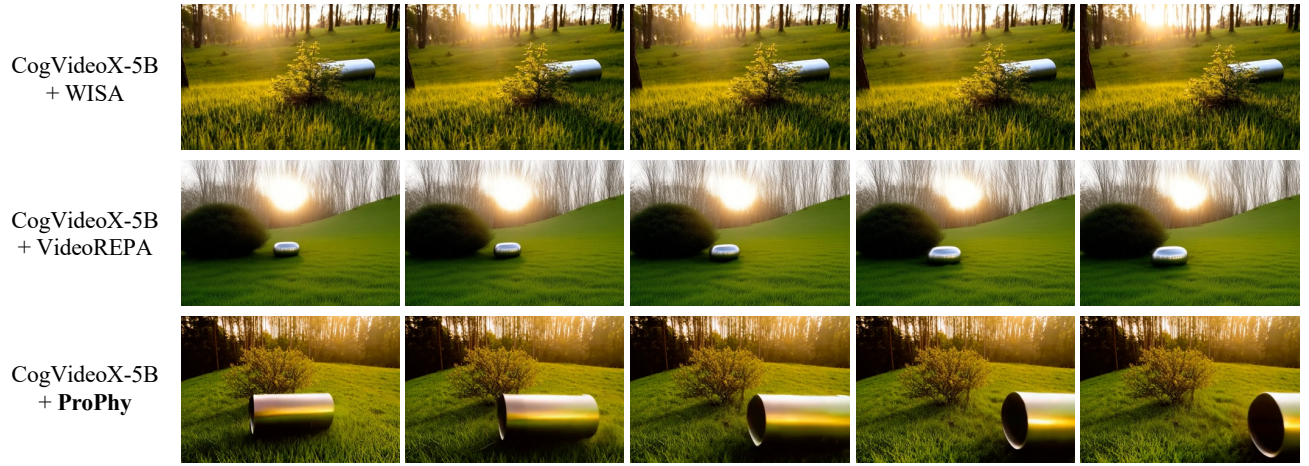
We fully recognize that our work still has several important limitations. First, our training process relies heavily on VLM-generated annotations of the physical attributes in videos. The alignment-based training does embed this physical awareness into the model and produces an end-to-end inference pipeline. However, the physical knowledge learned by the model still depends strongly on the accuracy of these annotations. In addition, our approach uses physical categories and fine-grained supervision to help the model simulate physical phenomena. This allows ProPhy to behave like an initial version of a world simulator. Even so, physical categorization only restricts the parameter space of each expert to a certain subset of general physical behaviors. It does not enforce object dynamics through explicit physical equations. The model ultimately generates plausible videos by fitting patterns in real data rather than by following precise physical laws. Future research can extend our idea by incorporating the corresponding physical differential equations. Such guidance could allow different physical phenomena to be generated in a more accurate and principled manner.

D. Social Impact

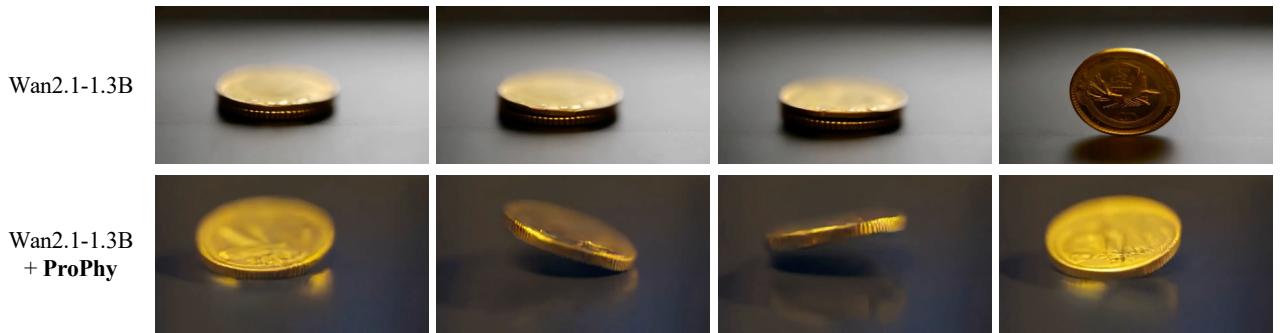
ProPhy enables text-to-video generation that aligns more closely with physical dynamics. Although it is still far from reproducing real-world physical scenes perfectly, it has the potential to support physics education in scenarios where real experiments are difficult to conduct. Regarding potential risks of identity leakage, the dataset we use contains almost no human faces. Most identity-related content, if any, comes from the generation capability of the underlying base models. Our method is designed purely for research on generative modeling and is not intended for use in safety-critical or deceptive applications. To support further research in the community, we will release all code after the final publication.



The high-speed baseball hit the brick wall...



A cylindrical metal container rolls down a grassy hill...



The coin makes its final spin on the table...

Figure 5. Comparison between ProPhy with different backbones and previous methods, including the baseline. More generated examples are provided in the supplementary video.



A car speeding through the rainy night... (reflection and liquid motion)



Cloth banner hanging from wooden twig... (deformation)



A glass fell from the cabinet... (rigid body motion and scattering)



A worker mixes concrete with a paddle... (rigid body motion and liquid motion)



A brush spreading glue on a piece of paper... (elastic motion and liquid motion)



Professional surfers make sharp turns on the waves... (human motion and liquid motion)



A spray can sprays out a layer of red paint onto a wall... (deposited and gas motion)

Figure 6. Examples of videos generated by ProPhy in response to text prompts involving complex physical phenomena.