PHANTOM: Physics-Infused Video Generation via Joint Modeling of Visual and Latent Physical Dynamics

Supplementary Material

A. Implementation Details

Training Details. For our main experiments, we build upon the Wan2.2-TI2V-5B [34] due to its ability to accept both text and image inputs. We integrate our physics branch into this architecture as described in Section 4.2. The physics branch is initialized from scratch, while the visual branch is kept frozen to preserve the strong generative prior of the base model. For extracting physics-aware embeddings, we leverage V-JEPA2 [4], a pretrained video encoder shown to capture intuitive physics properties [11]. In particular, we use the VJEPA2-ViT-H-fpc64-256 variant. We have trained the model for two epochs. We train all models with a global batch size of 128 using the AdamW optimizer with a learning rate of $4e-5$ and weight decay $1e-3$. We use cosine learning rate decay with a 5% warmup ratio. All experiments are performed on 4 NVIDIA H200 GPUs.

Evaluation Details. We conduct evaluations on all benchmarks using their official protocols and codebases to ensure comparability with prior work. For VideoPhy [5], we use the official auto-rater for all evaluations. Results are reported using both the original prompts provided in the dataset and the more detailed prompts used in VideoREPA [41]. Following VideoREPA [41], we set Semantic Adherence (SA) = 1 and Physical Commonsense (PC) = 1 when their values are greater than or equal to 0.5, and values less than 0.5 are set as SA = 0 and PC = 0. The final SA and PC scores correspond to the fraction of videos assigned a score of 1 after thresholding.

In VideoPhy-2 [6], we follow the official evaluation protocol. Both SA and PC are computed as the proportion of videos that receive a rating of at least 4 out of 5 from the benchmark's auto-evaluator. We directly use the official up-sampled prompts for evaluation. For Vbench2 [42], we report the results using its original prompts.

For Physics-IQ [25], we evaluate under both single-frame and multi-frame conditioning. In the single-frame setting, the model receives only the initial frame and the caption as inputs, whereas in the multi-frame setting, the model observes a short initial clip and the corresponding caption.

B. Baselines

B.1. General-Purpose Video Models

We compare against several state-of-the-art general-purpose text-to-video (T2V) diffusion models that serve as strong baselines in open-domain video generation, including CogVideoX-5B [39], HunyuanVideo [18], Wan2.1-T2I-14B, and Wan2.2-TI2V-5B [34]. These models demonstrate strong open-domain generalization and high-fidelity video synthesis but are not designed to model or enforce physical principles.

Table 4. **Results on VideoPhy and VideoPhy2 Benchmarks.** Semantic Adherence (SA) measures video-text alignment and fidelity. Physical Commonsense (PC) measures whether generated videos follow real-world physics laws intuitively. † denotes results reported from VideoREPA [41] with the original prompt. Improvements over the base model Wan2.2-TI2V are highlighted in ↑green. Best results in **bold**, second-best underlined. Following VideoREPA [41], we also report results with detailed prompts, denoted by *.

Method	VideoPhy		VideoPhy-2	
	SA↑	PC↑	SA↑	PC↑
General-Purpose				
VideoCrafter2 [8]	50.3	29.7	25.89	55.67
LaVIE [36]	48.7	31.5	-	-
Cosmos-Diffusion-7B [1]	57.0	18.0	26.32	54.19
CogVideoX-5B [39]	63.1	31.4	**28.86**	68.42
Wan2.2-TI2V-5B [34]	41.5	25.2	24.53	69.20
Wan2.2-TI2V-5B* [34]	64.7	28.6	24.53	69.20
Physics-Focused				
PhyT2V (Round 4)† [38]	61	37	-	-
WISA† [35]	62	33	-	-
VideoREPA [41]	51.9	22.4	21.02	**72.54**
VideoREPA*† [41]	**72.1**	**40.1**	21.02	**72.54**
PHANTOM (Ours)	47.5 ↑14.5%	37.9 ↑50.4%	27.75 ↑13.1%	71.74 ↑2.6%
PHANTOM* (Ours)	70.3 ↑8.7%	39.4 ↑37.8%	27.75 ↑13.1%	71.74 ↑2.6%

B.2. Physics-Focused Video Models

In addition to general-purpose video generators, we compare against a set of recent physics-focused video generation approaches that aim to improve physical plausibility.

PhyT2V [38] uses large language models to iteratively refine prompts via chain-of-thought and step-back reasoning. By repeatedly analyzing and rewriting the prompt, it guides existing text-to-video models toward generating videos that better adhere to real-world physical laws without retraining the generation model.

WISA [35] is a physics-aware video generation approach that incorporates explicit physical categories and properties. These physical attributes are embedded into the generation process through Mixture-of-Physical-Experts Attention (MoPA) and a dedicated Physical Classifier, enabling the model to incorporate richer physical priors during synthesis.

VideoREPA [41] injects physics understanding into diffusion-based video generators by aligning their hidden states with the representation from video foundation models via distillation.

C. Additional Results

Quantitative Results. Table 4 presents extended results on VideoPhy [5] and VideoPhy2 [6], including both the evaluation on original prompt and detailed prompts (denoted by *) following VideoREPA [41]. Across both settings, PHANTOM yields substantial performance gains over the base

Table 5. **Text-to-video evaluation on VBench-2.** Best results in **bold**. Improvements over base model Wan2.2-TI2V highlighted in ↑green.

Model	Total	Creativity	Commonsense	Controllability	Human Fidelity	Physics
Wan2.2-TI2V-5B [34]	51.57	**52.50**	60.57	18.50	86.10	40.19
PHANTOM (Ours)	**51.84**$_{\uparrow 0.5\%}$	45.51	**61.43**$_{\uparrow 1.4\%}$	**20.23**$_{\uparrow 9.4\%}$	**88.39**$_{\uparrow 2.7\%}$	**43.61**$_{\uparrow 6.0\%}$
Model	**Human Anatomy**	**Human Clothes**	**Human Identity**	**Composition**	**Diversity**	**Mechanics**
Wan2.2-TI2V-5B [34]	87.32	92.31	**78.70**	40.35	**64.67**	59.13
PHANTOM (Ours)	**90.19**$_{\uparrow 3.3\%}$	**96.85**$_{\uparrow 4.9\%}$	78.12	**45.07** $_{\uparrow 11.7\%}$	45.95	**60.48**$_{\uparrow 2.3\%}$
Model	**Material**	**Thermotics**	**Multi-view**	**Dynamic Spatial Rel.**	**Dynamic Attribute**	**Motion Order**
Wan2.2-TI2V-5B [34]	36.49	54.11	11.05	24.64	**9.52**	10.77
PHANTOM (Ours)	**37.33**$_{\uparrow 2.3\%}$	**54.61**$_{\uparrow 0.9\%}$	**22.01**$_{\uparrow 99.2\%}$	**32.37**$_{\uparrow 31.4\%}$	6.23	**12.46**$_{\uparrow 15.7\%}$
Model	**Human Interact.**	**Complex Landscape**	**Complex Plot**	**Camera Motion**	**Motion Rationality**	**Instance Preservation**
Wan2.2-TI2V-5B [34]	37.33	**18.89**	9.52	**18.83**	27.59	**93.57**
PHANTOM (Ours)	**47.00**$_{\uparrow 25.9\%}$	18.22	**10.23**$_{\uparrow 7.5\%}$	15.12	**29.89**$_{\uparrow 8.3\%}$	92.98

Wan2.2-TI2V-5B model, demonstrating improved physical commonsense and semantic fidelity. The improvements are especially pronounced under the original-prompt setting, where no dense textual description is provided, indicating that PHANTOM has learned strong intrinsic physics-awareness without relying on enriched prompts. Despite the fact that VideoREPA [41] is built upon CogVideoX-5B, a considerably stronger backbone than Wan2.2, PHANTOM still delivers large improvements over its base model and achieves competitive performance, underscoring the effectiveness of our approach.

Table 5 reports fine-grained performance across all 18 VBench-2 metrics. PHANTOM outperforms the base Wan2.2-TI2V-5B on the majority of these dimensions, demonstrating that joint physics-aware modeling not only boosts physics-related metrics but also helps improve overall perceptual realism, semantic consistency, and temporal coherence.

Ablation Studies. In addition, we replace the VJEPA2 encoder with VideoMAEv2, an alternative video encoder, while keeping the same training setup on Wan2.2-TI2V. Table 6 shows PHANTOM w/ VJEPA2 achieves better performance across all metrics, supporting the choice of VJEPA2 for physics-aware latent representation.

Qualitative Results. We provide additional qualitative comparisons against both state-of-the-art T2V models and recent physics-focused approaches, as shown in Figure 4 and 5. Since most physics-focused baselines operate solely in the text-to-video setting, Figure 5 compares PHANTOM only with general-purpose T2V models.

D. Physics-based Video Control

To further evaluate the ability of PHANTOM to model and respond to explicit physical control signals, we apply our framework to the Force-Prompting dataset [1]. Force-Prompting provides paired video sequences and temporally aligned force annotations describing external physical interactions applied to static images. Specifically, we focus on the local point force setting, in which a localized force is applied to an object at a specific image coordinate.

[1] https://force-prompting.github.io

Table 6. **Alternative Video Encoders.**

Visual Encoder	VideoPhy		VideoPhy-2	
	SA↑	PC↑	SA↑	PC↑
Wan2.2-TI2V-5B	41.5	25.2	24.53	69.20
PHANTOM w/ VJEPA-2	47.5	37.9	27.75	71.74
PHANTOM w/VideoMAEv2	45.8	37.6	26.90	70.56

We convert each point-force annotation into a *force tensor* that encodes both the spatial distribution and temporal evolution of the applied forces. Each tensor is then rendered as a short video sequence at a resolution of 256×256, providing a consistent spatiotemporal representation of the external force. These force videos are processed by the V-JEPA2 encoder in the same way as ordinary video inputs, producing physics-aware embeddings that are fed through the physics branch.

Since the original video captions in the Force-Prompting dataset do not contain force-related information, we additionally construct a textual *force prompt* that describes the applied force in natural language. This prompt encodes all relevant physical parameters and is fed into the physics branch during training and inference:

```
Simulate the scene under an external point
force applied at (x, y) = ({coordx}, {coordy}),
with magnitude = {force} and direction = {angle}
degrees, and generate the resulting video dynamics.
```

In this application, the two branches of PHANTOM receive different inputs and textual conditions. The *video branch* models the visual evolution of the scene and is conditioned on the original video caption. In contrast, the *physics branch* processes the force-tensor video and is guided by the constructed force prompt. During inference, PHANTOM is conditioned on a single static image along with the first frame of the force-tensor sequence, and it synthesizes the resulting physically driven dynamics.

We follow the same experimental hyperparameters as in our main setup and fine-tune from the PHANTOM for 1.1K steps. Figure 6 shows that PHANTOM can synthesize dynamic and physically plausible motion that evolves consistently with the applied force, demonstrating its ability to generalize force-based control signals.

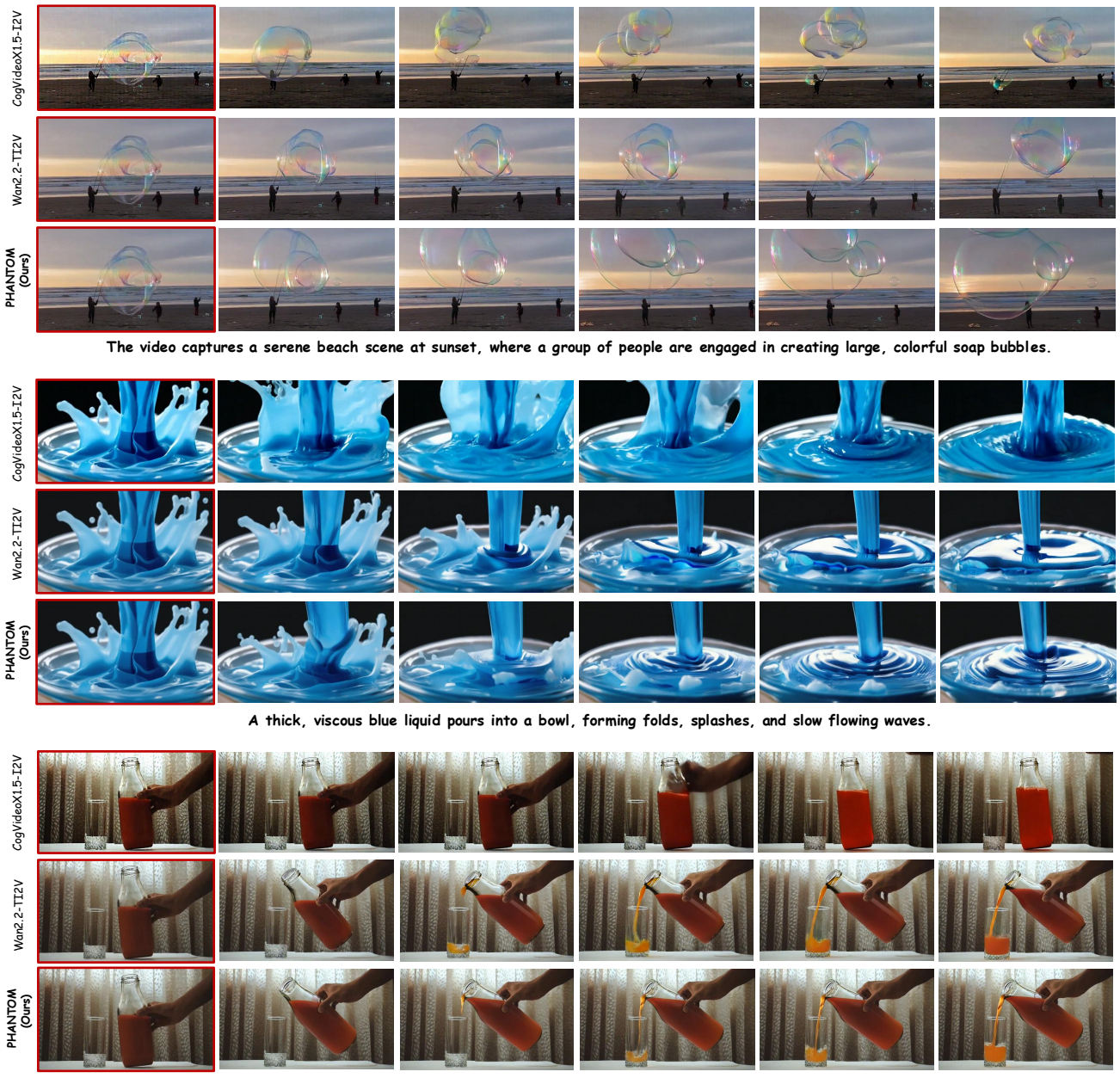

Figure 4. **Qualitative Comparison on Text-/Image-to-Video Generation.** The conditional frame is marked in red box.

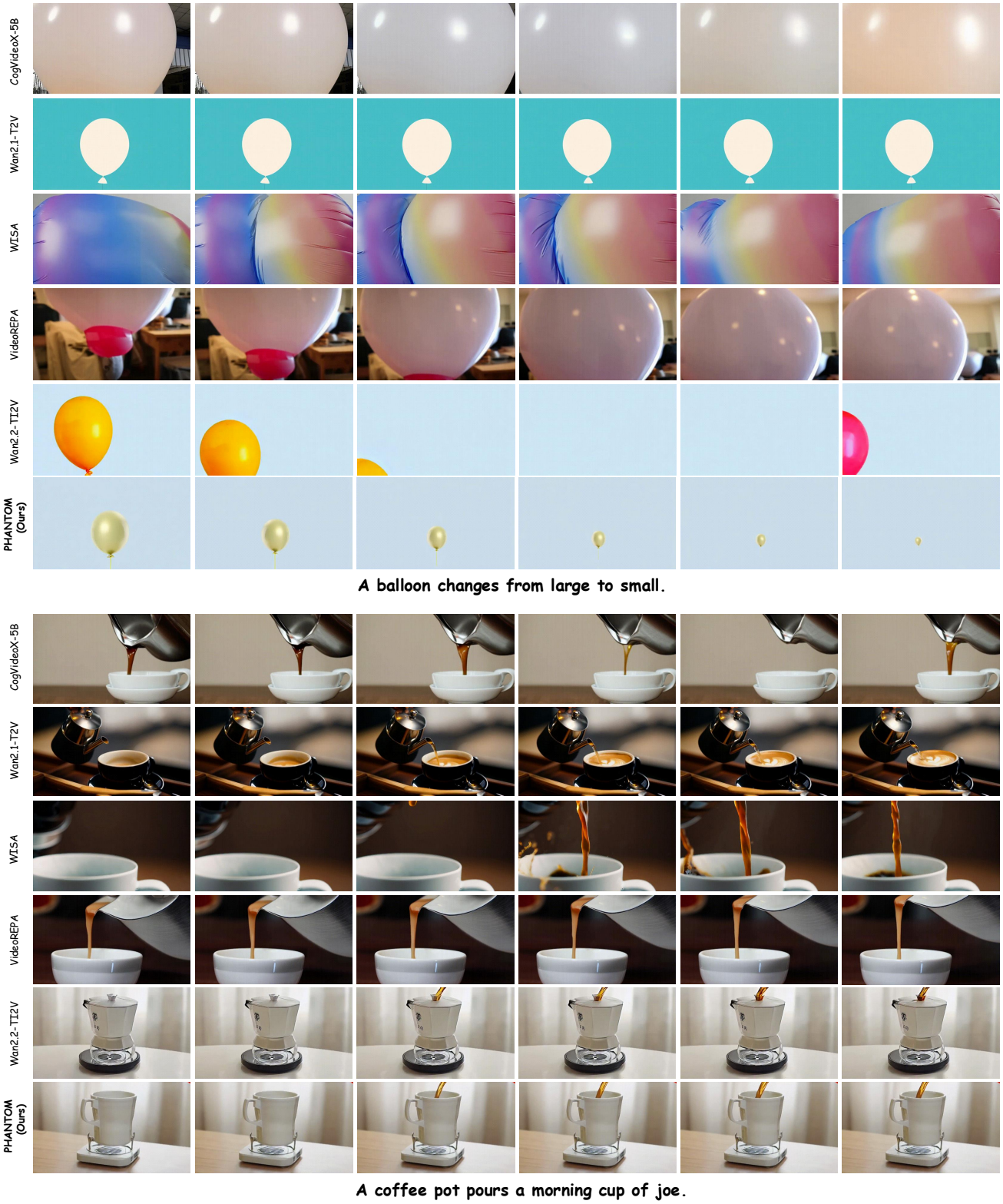

Figure 5. **Qualitative Comparison on Text-to-Video Generation.**

Figure 6. **Examples of Force-conditioned Video Generation using PHANTOM.** The conditional frame is marked in red box.