

# PhysMotion: Physics-Grounded Dynamics From a Single Image

## Supplementary Material

### 1. Additional Implementation Details

In this section we provide comprehensive implementation details.

#### 1.1. Model Use

For the pre-trained text-to-image model, we apply the publicly available UNet-based checkpoints of Stable-Diffusion-2-1<sup>1</sup>. We train LoRA weights based on above models for personalization. For ControlNet models, we utilize public checkpoints for canny-edge-ControlNet<sup>2</sup> and depth-ControlNet<sup>3</sup>. We extract the canny-edge control signal using the OpenCV library [2], and we adopt the depth map  $\mathbf{D}$  from Eq. (2) in the paper or use a depth map extracted from coarse dynamics using MiDaS [7].

#### 1.2. Parameter Settings

##### 1.2.1 Geometry-Aware Reconstruction

To obtain a 3DGS representation ready to generate reasonable dynamics, our training parameters are carefully chosen: for most of our experiments, the number of training epoch is 3000, with parameters' learning rates set within the range  $[10^{-4}, 10^{-3}]$ .

We apply a learning rate decay strategy by down-scaling the learning rates by  $10^{-1}$  after 1500 epochs; we apply hard-depth supervision every 10 epochs and after epoch 500. We do not apply the soft-depth supervision as indicated in [5] since we do not observe significant change of quality in reconstruction output under our settings.

##### 1.2.2 Generative Video Enhancement

We set the deterministic DDIM+ inversion total steps as 1000, and we set the step size as 20; following [8] and [3], we set the classifier-free-guidance (CFG) [4] scale to 1. We apply deterministic DDIM+ sampling with 50 steps.

We notice that in general, enhanced video achieves better temporal consistency with higher number of key-frames sampled. For most of our experiments, we randomly choose key-frames every 5 frames, and we set the guidance scale to 7.5.

<sup>1</sup><https://huggingface.co/stabilityai/stable-diffusion-2-1>

<sup>2</sup><https://huggingface.co/thibaud/controlnet-sd21-canny-diffusers>

<sup>3</sup><https://huggingface.co/thibaud/controlnet-sd21-depth-diffusers>

### 1.3. Experiment Details

We observed for different physical scenes in our experiments, VideoPhy [1] provides scores of different scales in both physics commonsense (PC) and semantic adherence (SA). Therefore, simply calculating the average score is unfair as it does not account for the varying scales of the scores, which could disproportionately influence the results and lead to biased comparisons. To mitigate the influence of heterogeneous scales and ensure a fair comparison across different models, we perform a  $z$ -score normalization. Specifically, for each scene  $t$  (13 in total) and score  $x_{i,t}$  of model  $i$  (chose from {ours, CogVideoX-5B, Dyan-iCrafter, I2VGen-XL, MotionI2V, DragAnything}), we calculate the  $z$ -score as follows:

$$z_{i,t} = \frac{x_{i,t} - \mu_t}{\sigma_t}, \quad (1)$$

where  $\mu_t$  and  $\sigma_t$  represent the mean and standard deviation of scores across all models for scene  $t$ . This normalization allows for a fair comparison across scenes with different scoring scales by transforming scores into a common scale,. We then compute each model's overall performance by averaging its  $z$ -scores across all scenes:

$$\bar{z}_i = \frac{1}{N} \sum_{t=1}^N z_{i,t} \quad (2)$$

where  $N$  denotes the total number of scenes ( $N = 13$  in our experiments). Models with higher average  $z$ -scores demonstrate stronger overall performance across scenes.

### 1.4. Additional Preliminary Knowledge

We provide additional preliminary knowledge on the attention mechanism [9].

In self-attention blocks within transformer blocks, the features  $\mathbf{f} \in \mathbb{R}^{n \times d_f}$  ( $n$  is the sequence length and  $d_f$  is the dimension of feature) are projected into queries  $\mathbf{Q}$ , keys  $\mathbf{K}$ , and values  $\mathbf{V}$  using

$$\mathbf{Q} = \mathbf{f}\mathcal{W}_{\mathbf{Q}}, \quad \mathbf{K} = \mathbf{f}\mathcal{W}_{\mathbf{K}}, \quad \mathbf{V} = \mathbf{f}\mathcal{W}_{\mathbf{V}} \quad (3)$$

where  $\mathcal{W}_{\mathbf{Q}}, \mathcal{W}_{\mathbf{K}}, \mathcal{W}_{\mathbf{V}} \in \mathbb{R}^{d_f \times d}$  are learned weights matrices for queries, keys and values respectively.  $d$  is the dimension of the embedded vector as in Eq. (12) in the paper.

The attention mechanism computes the weighted sum of the values, with the weights determined by the similarity between queries and keys. Specifically, the attention scores are calculated as the scaled dot product between the queries

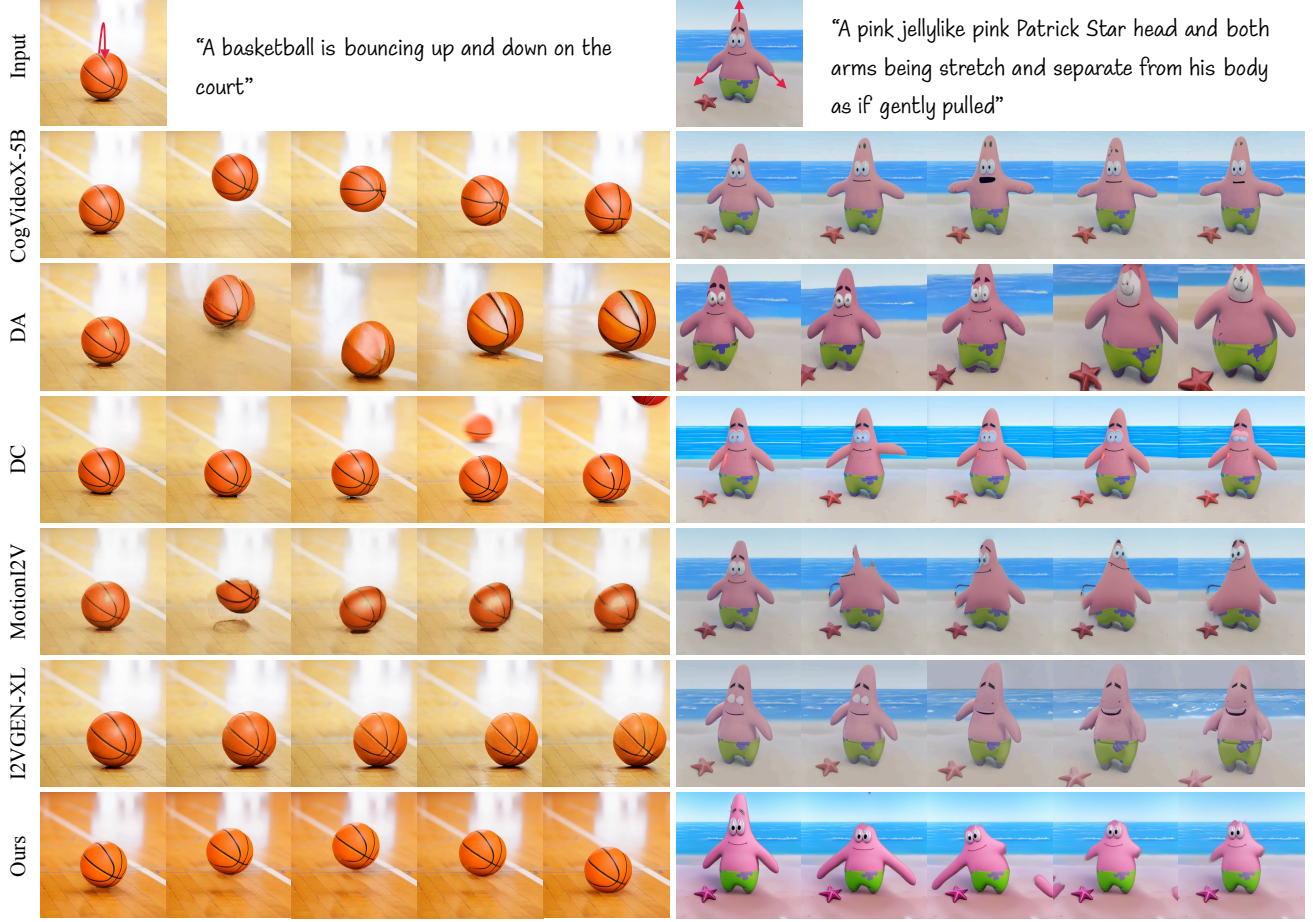


Figure 1. **Additional Qualitative Comparison.** We provide additional comparison results against MotionI2V [6], DragAnything (DA) [10], CogVideoX-5B [12], DynamiCrafter (DC) [11] and I2VGen-XL [6]. Text prompts for CogVideoX-5B, I2VGen-XL and DynamiCrafter are generated using ChatGPT-4o, while trajectories are used for DragAnything and Motion-I2V.

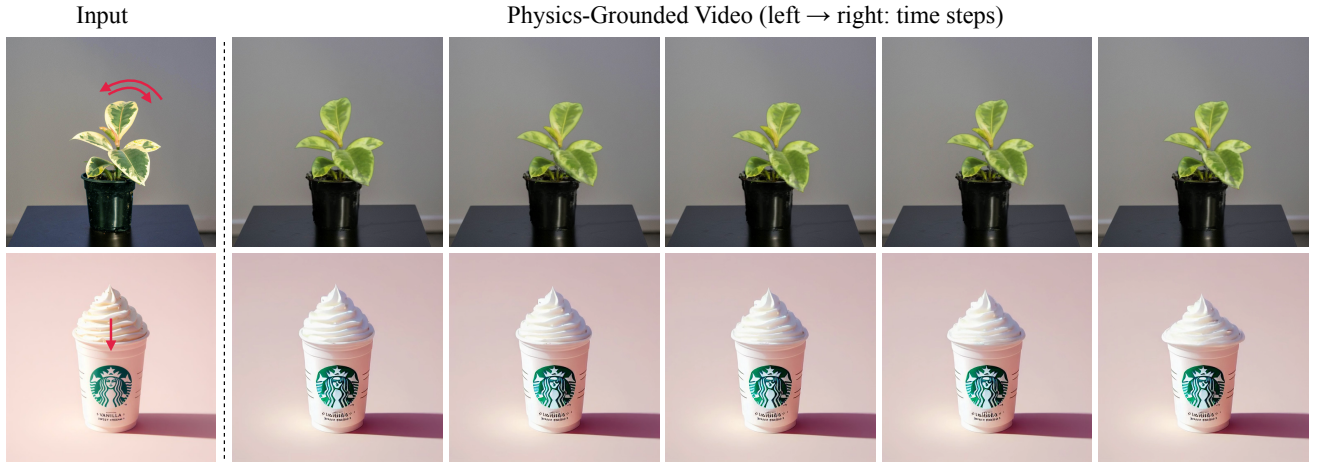


Figure 2. **Additional Showcases.** We demonstrate additional showcases created by PhysMotion.

and keys, as

$$\mathcal{A} = \text{Softmax} \left( \frac{\mathbf{QK}^T}{\sqrt{d}} \right), \quad (4)$$

where  $\mathcal{A} \in \mathbb{R}^{n \times n}$  contains the attention scores for all query-key pairs, with the weighted sum to 1 for each query. The

final output of the attention mechanism is given by

$$\phi = \mathcal{A} \cdot \mathbf{V}. \quad (5)$$

Note that in Eq. (13) in paper, the  $\mathbf{V}$ 's are concatenated to form a shared value matrix.

## 2. More Results

### 2.1. Qualitative Comparison

In Fig. 1, we provide additional qualitative comparison results with baseline methods, including CogVideoX-5B [12], Drag Anything [10], DynamiCrafter [11], Motion-I2V [6] and I2VGen-XL [13].

### 2.2. Showcases

As indicated in Fig. 2, we present additional showcases created using the proposed method. Our method enables users to generate high-fidelity, physics-grounded dynamics.

## References

- [1] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv:2406.03520*, 2024. 1
- [2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 1
- [3] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv:2307.10373*, 2023. 1
- [4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 1
- [5] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. *arXiv:2403.06912*, 2024. 1
- [6] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3
- [7] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. Ldm3d: Latent diffusion model for 3d. *arXiv:2305.10853*, 2023. 1
- [8] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 1
- [9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 1
- [10] 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, 2025. 2, 3
- [11] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 2, 3
- [12] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenqi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv:2408.06072*, 2024. 2, 3
- [13] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv:2311.04145*, 2023. 3