PhysLayer: Language-Guided Layered Animation with Depth-Aware Physics

Tianyidan Xie[1], Zhentao Huang[2], Mingjie Wang[3], Xin Huang[3], Jun Zhou[4], Minglun Gong[2], Zili Yi[1,*]
[1]Nanjing University [2]University of Guelph [3]Zhejiang Sci-Tech University [4]Dalian Maritime University

***Abstract*—Existing image-to-video generation methods often produce physically implausible motions and lack precise control over object dynamics. While prior approaches have incorporated physics simulators, they remain confined to 2D planar motions and fail to capture depth-aware spatial interactions. We introduce PhysLayer, a novel framework enabling language-guided, depth-aware layered animation of static images. PhysLayer consists of three key components: First, a language-guided scene understanding module that utilizes vision foundation models to decompose scenes into depth-based layers by analyzing object composition, material properties, and physical parameters. Second, a depth-aware layered physics simulation that extends 2D rigid-body dynamics with depth motion and perspective-consistent scaling, enabling more realistic object interactions without requiring full 3D reconstruction. Third, a physics-guided video synthesis module that integrates simulated trajectories with scene-aware relighting for temporally coherent results. Experimental results demonstrate improvements in CLIP-Similarity (+2.2%), FID score (+9.3%), and Motion-FID (+3%), with human evaluation showing enhanced physical plausibility (+24%) and text-video alignment (+35%). Our approach provides a practical balance between physical realism and computational efficiency for controllable image animation.**

***Index Terms*—image animation, physics simulation, depth-aware layered dynamics, video generation, vision-language models**

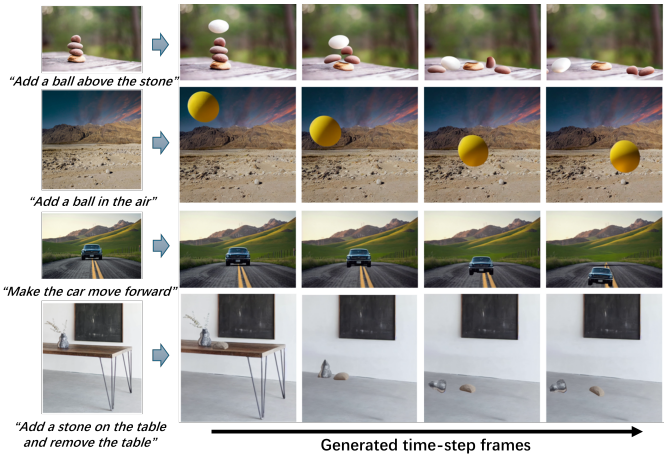

Fig. 1. Visualizing PhysLayer's capabilities across diverse animation scenarios. Our method supports (a) object insertion with depth-aware collision handling (rows 1 and 2); (b) controlled object manipulation with accurate perspective scaling (row 3); and (c) consistent, physically aware scene reconstruction (row 4). Across all examples, our approach maintains both physical plausibility and visual coherence throughout the generated sequence.

I. Introduction

The advent of advanced diffusion models has revolutionized image-to-video generation, leading to unprecedented improvements in quality and performance [1], [2]. Despite significant advancements, a fundamental challenge remains: generating videos that not only maintain high visual fidelity but also exhibit physical realism, accurately adhering to real-world physics while allowing precise control over object motion and interactions. This challenge is particularly crucial for applications in content creation, visual effects, and interactive media, where both visual quality and physical accuracy are indispensable. Current methods generally fall into two categories: those that rely solely on generative models, which often produce unnatural motions that violate physical laws [3]–[5], and those that integrate physics simulations but are constrained to 2D planar dynamics with limited controllability [6].

Specifically, we identify three key challenges in existing methods: First, they often require complex motion parameters or generate unnatural motions that violate physical laws. Second, while some approaches incorporate physics simulation, they are typically confined to 2D planar dynamics, failing to capture the rich spatial interactions inherent in real-world 3D scenes. Third, current methods lack a unified framework that enables flexible and precise control over diverse animation tasks using both language guidance and visual control.

To overcome these challenges, we propose PhysLayer, a novel framework that extends physics-simulation methods into the 3D domain and enables language-guided animation of static images. Our approach is structured into a three-phase procedure:

First, Language-Guided Scene Understanding and Layer Decomposition: We leverage large vision foundation models and advanced image parsing techniques for comprehensive scene understanding. This involves reasoning about object composition, material properties, and physical parameters. Based on depth information, we decompose the scene into multiple interactive layers, capturing spatial relationships and enabling individual object manipulation.

Second, Depth-Aware Physics Simulation: We extend traditional 2D rigid-body physics into 3D space through depth-aware multi-planar dynamics simulation. This allows us to simulate object dynamics with physical accuracy, accounting for interactions between objects and their environment across varying depth planes. By incorporating depth variations, we

*Corresponding author: yi@nju.edu.cn

enable realistic simulations of object-object and object-scene interactions that reflect real-world physics.

Third, Physics-Guided Video Synthesis Pipeline: We design a video synthesis pipeline that seamlessly integrates the physically simulated trajectories with scene-aware relighting. By utilizing diffusion-based refinement and rendering techniques, we enhance visual quality and temporal coherence. This results in visually compelling videos that maintain both physical plausibility and high aesthetic appeal.

Our contributions can be summarized as:

- **A Language-Guided Depth-Aware Animation Framework:** We present a framework that offers flexible and precise control over static image animation through natural language instructions. By decomposing scenes into depth-based layers and reasoning about physical properties, our approach enables intuitive animation control while maintaining physical plausibility.
- **A Layered Depth-Aware Physics Simulation:** We introduce a depth-aware physics simulation technique that extends traditional 2D rigid-body dynamics with depth motion and perspective-consistent scaling. By operating on discretized depth layers, our method captures more realistic spatial interactions than pure 2D methods while avoiding the complexity of full 3D reconstruction. This provides a practical trade-off between physical realism and computational efficiency.
- **Comprehensive Evaluation:** Extensive experiments demonstrate PhysLayer's effectiveness across multiple metrics. We achieve improvements in CLIP-Similarity (2.2%), FID score (9.3%), and Motion-FID (3%). Human evaluation validates substantial gains in physical plausibility (24%) and text-video alignment (35%). We provide ablation studies and physical consistency analysis to validate each component's contribution.

II. Related Works

A. Large Model Based Physical Reasoning

Visual physical reasoning—the ability to understand and predict physical phenomena from visual observations—has long been a challenging problem in computer vision and AI. Traditional approaches often depend on specialized physics simulators or hand-crafted rules [7], [8]. Early benchmarks like CLEVRER [9] and PHYRE [10] focused on reasoning about collision events and physical interactions.

The advent of large foundation models has revolutionized zero-shot capabilities in physical reasoning. GPT-4V [11] exhibits impressive abilities in inferring physical properties and predicting object interactions without explicit physics training. Recent studies have leveraged these capabilities for various physical tasks: PAC-NeRF [12] utilizes Vision-Language Models (VLMs) for physics-aware scene reconstruction, while [13] employs language-embedded feature fields for understanding physical properties.

Beyond direct inference, large models show promise in guiding physics simulations. The Dynamic Concept Learner [14] integrates visual, linguistic, and physics-based reasoning to enhance dynamic predictions. Recent work, PhysGen [6], illustrates the potential of using vision-language models for physical property estimation through multiple object-centric queries. Our method builds on this concept by performing holistic scene understanding in a single pass.

B. Image-based Physical Simulation

Physical simulation from visual inputs has been extensively studied in computer vision and graphics. Traditional approaches rely on specialized physics engines [15], [16] for rigid body dynamics, fluid simulation, and deformable objects. While full 3D simulation provides accurate physics, it requires complete 3D scene reconstruction and understanding, which remains challenging from a single image.

Recent work, PhysGen [6], demonstrates using vision-language models for physical property estimation through object-centric queries. However, PhysGen operates purely in 2D space without depth awareness, limiting its ability to model perspective changes and depth-based interactions. Our method builds on this concept by performing holistic scene understanding in a single pass while extending the simulation to handle depth-aware dynamics, providing a more realistic representation of spatial relationships.

C. Video Generation and Image Animation

Video generation has witnessed remarkable progress with the emergence of diffusion models [17]. Recent large-scale text-to-video models like I2VGen-XL [4] and SEINE [2] demonstrate impressive capabilities in open-domain video synthesis through cascaded diffusion and temporal modeling. CogVideoX [5] and DynamiCrafter [3] further improve generation quality by introducing expert transformers and specialized video diffusion priors.

Image animation offers a more controlled approach to video synthesis. Traditional methods focus on bringing static images to life through motion synthesis [18]. Recent works like PIA [19] introduce plug-and-play modules to achieve personalized image animation. However, most existing methods operate primarily in 2D image space, lacking the ability to model depth-aware motion and spatial interactions accurately.

Our method addresses these limitations by combining physics-based simulation with modern diffusion models. By explicitly modeling depth-aware object dynamics and interactions, we achieve more precise control over object movements in both planar and depth directions.

III. Method

Given a single input image and natural language instructions, our goal is to generate a realistic video that features physically plausible dynamics while adhering to the instructed animation task. Our key insight is to bridge natural language understanding with depth-aware physics simulation through a three-stage framework.

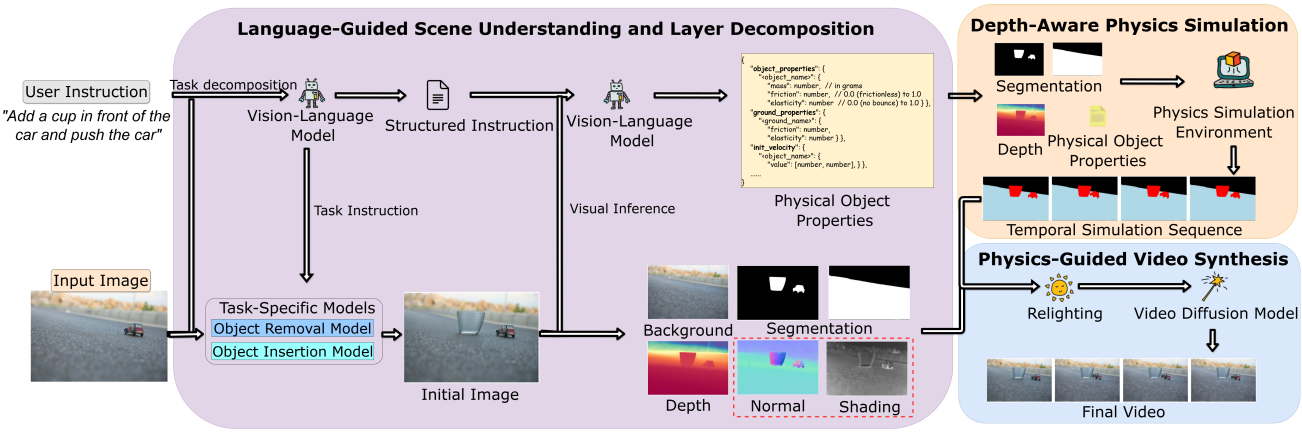

Fig. 2. **PhysLayer framework**. Our language-guided image animation framework consists of three components: (1) Language-Guided Scene Understanding and Layer Decomposition, (2) Depth-Aware Physics Simulation, and (3) Physics-Guided Video Synthesis.

A. Language-guided Scene Understanding and Layer Decomposition

The language-guided scene understanding and layer decomposition module generates a comprehensive scene representation by jointly reasoning about semantic, geometric, and physical properties through scene understanding and visual inference.

Scene Understanding: Our framework leverages large vision-language models (VLMs) [11], [20] to develop a thorough understanding of user intentions and scene physics. This process is achieved through two primary analyses: (1) *Task Decomposition*: The VLM interprets natural language instructions to identify animation objectives and associated physical interactions. Based on this analysis, animations are categorized into three core operations: object insertion, removal, and control. (2) *Physical Property Inference*: Building on recent advances in language-driven physical reasoning [13], [21], the VLM infers key physical properties such as mass, density, friction coefficients, and elasticity.

Vision Inference: Our framework integrates multiple specialized vision models to build a detailed and accurate scene representation. We employ Grounded-SAM [22] for instance-level semantic segmentation, which accurately delineates objects, boundaries, and supporting surfaces. Spatial configurations and depth relationships are captured using GeoWizard [23], a state-of-the-art depth and normal estimation network that enables depth-aware physical simulation.

To achieve photorealistic rendering, we utilize an intrinsic decomposition model [24] to analyze scene illumination conditions. For static scene structure reconstruction, we apply advanced generative inpainting techniques [25], [26], which recover complete background content in regions initially occupied by dynamic objects.

B. Depth-aware Physics Simulation

Building on the physical properties and depth information extracted during scene understanding, we introduce a depth-aware layered physics simulation framework. Unlike previous methods [6] that operate exclusively in 2D space, our approach extends traditional rigid-body dynamics with depth motion constraints and perspective-consistent scaling. Rather than requiring full 3D reconstruction, we achieve depth awareness through a layered representation that discretizes the depth dimension while maintaining computational efficiency.

a) Depth Layer Representation: We discretize the scene depth into L layers based on the estimated depth map from GeoWizard [23]. Each layer l corresponds to a depth range $[d_l, d_{l+1}]$, where objects within the same range interact physically. The number of layers is determined by analyzing the depth histogram and clustering objects at similar depths: $L = \min(N_{obj}/2, 8)$, where N_{obj} is the number of dynamic objects. This layered representation enables efficient collision detection while approximating depth-aware interactions.

We formulate our depth-aware dynamics in an augmented 2.5D representation. At time t, each rigid object i is represented by a state vector including translation $\mathbf{t}_i(t) \in \mathbb{R}^3$ (where the third dimension represents depth displacement) and rotation $\theta_i(t) \in SO(2)$ constrained to the image plane. The velocity state consists of linear velocity $\mathbf{v}_i(t) \in \mathbb{R}^3$ and angular velocity $\omega_i(t) \in \mathbb{R}$.

The complete state of object i at time t is formally expressed as $\mathbf{q}_i(t) = [\mathbf{t}_i(t), \theta_i(t), \mathbf{v}_i(t), \omega_i(t)]$, with dynamics governed by:

$$\frac{d}{dt}\mathbf{q}(t) = \frac{d}{dt}\begin{bmatrix}\mathbf{t}(t)\\ \theta(t)\\ \mathbf{v}(t)\\ \omega(t)\end{bmatrix} = \begin{bmatrix}\mathbf{v}(t)\\ \omega(t)\\ \frac{\mathbf{F}_{3D}(t)}{M}\\ \frac{\tau(t)}{I}\end{bmatrix} \tag{1}$$

Here, $\mathbf{F}_{3D} = [\mathbf{F}_{xy}, F_z]^T$ represents the composite force vector decomposed into planar and depth components, τ denotes planar torque, M is object mass, and I is the moment of inertia. The planar dynamics $\mathbf{F}_{xy}$ are computed using Pymunk's rigid-body engine, while F_z is regulated to maintain smooth depth motion.

b) Perspective-Consistent Scaling: A key advantage of our approach over 2D methods is realistic perspective handling. As objects move along the depth axis, their apparent size changes according to perspective projection:

$$S(d) = S_0 \cdot \frac{f}{f+d} \tag{2}$$

where S_0 is the original size at reference depth, d is the depth displacement, and f is the effective focal length estimated from the input image using camera calibration heuristics [23]. This linear scaling approximates the perspective foreshortening effect, enabling visually coherent animations where objects naturally appear larger when approaching the camera and smaller when receding.

c) Language-guided Initialization: Our framework executes three key tasks: (1) language-guided object insertion, which introduces objects at specified locations with physically accurate initialization; (2) language-guided object removal, which eliminates target objects while maintaining physical consistency; (3) language-guided object control, which applies user-specified forces and torques in the augmented space.

d) Collision Handling: Our system employs a layer-wise collision processing pipeline. Objects are assigned to depth layers based on their current depth $d_i(t)$. Collision detection and response are computed within each layer using Pymunk's 2D engine. This design choice trades off cross-layer interaction accuracy for computational efficiency and implementation simplicity. Objects moving across layer boundaries are reassigned dynamically. While this approximation cannot handle collisions between objects at significantly different depths, it effectively captures interactions among objects in similar depth ranges, which constitute the majority of realistic scenarios. The depth-dependent scaling ensures that collision responses appear visually consistent with the objects' apparent sizes.

C. Physics-guided Video Synthesis

Our video synthesis pipeline integrates physics-based motion, depth-aware appearance modeling, and neural rendering to produce temporally coherent and physically plausible animations. The synthesis process consists of three key stages: motion-guided composition, depth-aware relighting, and physics-constrained refinement.

Motion-guided Composition: We create an initial video sequence by alpha-blending the warped foreground objects onto the inpainted background according to simulated transformations, with scale adjustments based on relative depth:

$$\hat{X}(t) = composite(B, warp(X^i, T_i(t)) \cdot S_i(d(t))_i) \tag{3}$$

where X^i represents the segmented image of the i-th object appearance, B denotes the clean background, $T_i(t)$ represents the affine transformation matrix, and $S_i(d(t))$ is a depth-dependent scaling factor.

Depth-aware Relighting: To model appearance changes due to varying depths, we employ a relighting module that computes shading based on transformed albedo $\hat{A}(t)$, surface normals $\hat{N}(t)$, and estimated directional light L:

$$\tilde{X}(t) = f(\hat{X}(t), \hat{A}(t), \hat{N}(t), L, d(t)) \tag{4}$$

Physics-constrained Refinement: The relit sequence is refined through a latent diffusion model [25] that ensures physical consistency while improving visual fidelity. We apply an adaptive fusion strategy in latent space:

$$z_{t-1} = (1-m)z_{t-1} + m[w(t)z_{t-1} + (1-w(t))\tilde{z}_{t-1}] \tag{5}$$

where m is the foreground mask and $w(t)$ regulates fusion strength.

IV. Experiments

A. Implementation Details

Scene Understanding: Our implementation leverages several state-of-the-art models. We employ Gemini-pro-vision [20] for language-vision understanding and physical property inference. The VLM prompt includes: (1) task description (insert/remove/control), (2) object identification requirements, (3) physical property queries (mass range: 0.1-10kg, friction: 0.1-1.0, elasticity: 0.1-0.9). We use Grounded-SAM [22] for instance detection and segmentation, Geowizard [23] for depth and normal estimation, Clipaway [26] for background inpainting and object removal, and Diffree [27] to generate objects and masks from text descriptions.

Animation Pipeline: We implement physics simulation using Pymunk [15] with our depth-aware extensions. The depth dimension is discretized into $L = 4$ to 8 layers depending on scene complexity. We run 160 simulation steps at 30Hz and sample 16 frames uniformly to create 512×512 resolution videos. The perspective scaling factor $S(d)$ is computed using Eq. 2 with focal length f estimated from the depth map. For video refinement, we employ SEINE with DDIM sampling (25 steps, noise strength $s = 0.5$).

B. Experimental Setup

Dataset: To evaluate our approach across diverse scenarios, we curate indoor and in-the-wild images with variations in illumination, viewpoints, and scene complexity. Our evaluation dataset comprises three language-guided tasks (object insertion, removal, and control), each containing 20 unique initial images. For each image, we generated 5 videos with different physical parameters and initial conditions, yielding a total of 60 source images and 300 videos.

Baselines: We compare against five state-of-the-art image-to-video generation methods: DynamicCrafter [3], I2VGen-XL [4], PIA [19], CogVideoX [5], and PhysGen [6] as a physics-aware baseline.

C. Quantitative Results

Following prior works, we evaluate using four metrics: CLIP-Similarity [28] to measure alignment between instructions and video frames; FID [29] to assess visual quality; Motion-FID to evaluate motion coherence; Motion-S to assess smoothness of generated motions.

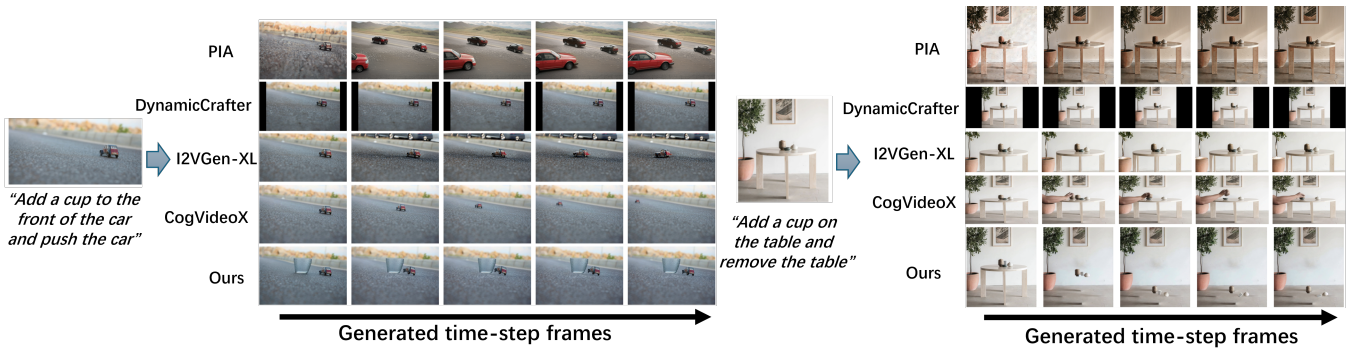

Fig. 3. Qualitative comparison of different methods on two challenging scenarios. Our method generates more realistic and coherent results with improved physics plausibility and temporal consistency compared to baseline approaches.

Table I shows our method achieves superior performance across all metrics. The improved CLIP-Sim score (16.64 vs $\leq$ 16.28) indicates better instruction-video alignment, while the lower FID score (102.32 vs $\leq$ 112.83) indicates enhanced visual quality. The improvement in Motion-FID (389.64 vs $\leq$ 401.61) demonstrates that our depth-aware physical simulation generates more coherent and realistic object motions.

TABLE I
QUANTITATIVE COMPARISON WITH BASELINE METHODS.

Methods	Clip-Sim	FID	M-FID	M-S
PIA	16.17	215.06	418.51	0.981
DynamiCrafter	16.16	185.86	429.22	0.954
CogVideoX	16.18	120.91	406.32	0.978
I2VGen-XL	16.28	187.35	442.91	0.956
PhysGen	-	112.83	401.61	0.982
Ours	**16.64**	**102.32**	**389.64**	**0.986**

D. Ablation Study

To validate the contribution of each component, we conduct ablation experiments by removing key elements from our framework. We evaluate on 30 videos across all three tasks using the same metrics as the main comparison.

TABLE II
ABLATION STUDY RESULTS. EACH ROW REMOVES ONE KEY COMPONENT.

Variant	Clip-Sim	FID	M-FID	M-S
Full Method	**16.64**	**102.32**	**389.64**	**0.986**
w/o Depth scaling ($S(d) = 1$)	16.52	118.45	405.28	0.979
w/o Depth motion ($v_z = 0$)	16.48	115.67	398.42	0.981
w/o Relighting	16.61	108.94	392.11	0.985
w/o Language guidance	15.93	105.28	391.87	0.984

Depth scaling is the most critical component: removing it ($S(d)$=1) causes the largest FID degradation (102.32→118.45). Depth motion removal (v_z=0) similarly hurts visual quality (FID→115.67). Relighting removal moderately affects FID (→108.94), while removing language guidance most impacts text-video alignment (CLIP-Sim: 16.64→15.93). All components contribute meaningfully to the full model.

E. User Study

We conducted a user study with 10 participants who evaluated 30 randomly selected videos based on three criteria—text-video alignment, visual quality, and physical plausibility—using a 5-point Likert scale.

As shown in Table III, our method outperforms all baselines significantly. It achieves the highest score for text alignment (4.23 vs $\leq$ 3.12), visual quality (3.83 vs $\leq$ 3.56), and most notably, physical plausibility (4.10 vs $\leq$ 3.31).

TABLE III
HUMAN EVALUATION RESULTS (1-5 SCALE).

Methods	Text-Align	Visual	Physics
PIA	1.86	1.68	1.55
DynamiCrafter	1.97	2.02	1.82
I2VGen-XL	2.21	2.54	2.16
CogVideoX	3.12	3.35	2.63
PhysGen	-	3.56	3.31
Ours	**4.23**	**3.83**	**4.10**

V. LIMITATIONS AND FUTURE WORK

While PhysLayer advances depth-aware image animation, several limitations remain. Our method is restricted to rigid body dynamics and cannot handle deformable objects, articulated structures, or complex physical phenomena such as fluid dynamics. The depth-layered collision detection prioritizes computational efficiency over complete accuracy, preventing realistic interactions between objects at significantly different depths. Our approach assumes a fixed camera viewpoint and cannot support significant view changes without full 3D scene reconstruction. From a practical perspective, the pipeline requires approximately 100–105 seconds per 16-frame video on an A6000 GPU after model initialization (with the diffusion renderer accounting for roughly 58% of the total time), limiting real-time applications, and the reliance on VLMs for

physical property estimation can introduce inaccuracies for uncommon materials. Future work will focus on extending to non-rigid motion, improving cross-layer interaction modeling, incorporating learned physical property predictors, accelerating inference for interactive applications, and supporting camera motion with multi-view consistency.

VI. Conclusion

We introduce PhysLayer, a framework for language-guided, depth-aware layered animation of static images. Our method decomposes scenes into depth-based layers using vision models, extends 2D rigid-body physics with depth motion and perspective-consistent scaling, and integrates simulated trajectories with scene-aware relighting to produce temporally coherent and physically plausible videos. This approach provides a practical balance between physical realism and computational efficiency without requiring full 3D reconstruction. Extensive experiments demonstrate that PhysLayer outperforms existing methods across multiple metrics, achieving substantial improvements in FID (9.3%), physical plausibility (24%), text-video alignment (35%), CLIP-Similarity (2.2%), and Motion-FID (3%), while offering flexible control over diverse animation effects through natural language guidance.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al., "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[2] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu, "Seine: Short-to-long video diffusion model for generative transition and prediction," in *The Twelfth International Conference on Learning Representations*, 2023.

[3] 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*. Springer, 2025, pp. 399–417.

[4] 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 preprint arXiv:2311.04145*, 2023.

[5] 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] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang, "Physgen: Rigid-body physics-grounded image-to-video generation," in *European Conference on Computer Vision*. Springer, 2025, pp. 360–378.

[7] Jiajun Wu, Joseph J Lim, Hongyi Zhang, Joshua B Tenenbaum, and William T Freeman, "Physics 101: Learning physical object properties from unlabeled videos.," in *BMVC*, 2016, vol. 2, p. 7.

[8] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum, "Learning to see physics via visual de-animation," *Advances in neural information processing systems*, vol. 30, 2017.

[9] 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.

[10] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick, "Phyre: A new benchmark for physical reasoning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[11] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[12] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan, "Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification," *arXiv preprint arXiv:2303.05512*, 2023.

[13] Albert J Zhai, Yuan Shen, Emily Y Chen, Gloria X Wang, Xinlei Wang, Sheng Wang, Kaiyu Guan, and Shenlong Wang, "Physical property understanding from language-embedded feature fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28296–28305.

[14] Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B Tenenbaum, and Chuang Gan, "Grounding physical concepts of objects and events through dynamic visual reasoning," *arXiv preprint arXiv:2103.16564*, 2021.

[15] Victor Blomqvist, "pymunk (2023)," *URL https://pymunk. org*.

[16] Erwin Coumans and Yunfei Bai, "Pybullet quickstart guide," 2021.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[18] Aleksander Holynski, Brian L Curless, Steven M Seitz, and Richard Szeliski, "Animating pictures with eulerian motion fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5810–5819.

[19] Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen, "Pia: Your personalized image animator via plug-and-play modules in text-to-image models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7747–7756.

[20] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al., "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[21] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh, "Physically grounded vision-language models for robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12462–12469.

[22] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al., "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[23] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long, "Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image," in *European Conference on Computer Vision*. Springer, 2025, pp. 241–258.

[24] Chris Careaga and Yağız Aksoy, "Intrinsic image decomposition via ordinal shading," *ACM Transactions on Graphics*, vol. 43, no. 1, pp. 1–24, 2023.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[26] Yigit Ekin, Ahmet Burak Yildirim, Erdem Eren Caglar, Aykut Erdem, Erkut Erdem, and Aysegul Dundar, "Clipaway: Harmonizing focused embeddings for removing objects via diffusion models," *arXiv preprint arXiv:2406.09368*, 2024.

[27] Lirui Zhao, Tianshuo Yang, Wenqi Shao, Yuxin Zhang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Rongrong Ji, "Diffree: Text-guided shape free object inpainting with diffusion model," *arXiv preprint arXiv:2407.16982*, 2024.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[29] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.