arXiv:2603.18639v3 [cs.CV] 2 Jun 2026

OrthoPhys: Physically Plausible Video Generation with Orthogonal-View Geometry Guidance

Cong Wang[1,2,3 *], Hanxin Zhu[4 *], Xiao Tang[5], Jiayi Luo[3,6]
Xin Jin[3,7], Long Chen[1 †], Zhibo Chen[3,4 †]
[1]the State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
[2]the School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Zhongguancun Academy
[4]School of Information Science and Technology, University of Science and Technology of China
[5]College of Automotive and Energy Engineering, Tongji University
[6]SKLCCSE, School of Computer Science and Engineering, Beihang University
[7]Eastern Institute of Technology
[*]Equal contribution [†]Corresponding author

Abstract

Recent progress in video generation has led to substantial improvements in visual fidelity, yet ensuring physically consistent motion remains a fundamental challenge. Intuitively, this limitation can be attributed to the fact that real-world object motion unfolds in three-dimensional space, while video observations provide only partial, view-dependent projections of such dynamics. To address these issues, we propose **OrthoPhys**, a two-stage framework that leverages orthogonal-view geometry guidance to enforce physical plausibility. Instead of directly generating unstructured 2D videos, our first stage generates synchronized, four-view orthogonal videos of the foreground dynamics. By incorporating a geometry-enhanced attention mechanism across these orthogonal views, this stage effectively enforces 3D spatial coherence and implicitly grounds the motion in physical attributes. In the second stage, these physically consistent orthogonal foregrounds serve as rigid guidance to synthesize the final complete video, seamlessly learning the interaction between foreground dynamics and the background context. To support this orthogonal-view training paradigm, we construct **PhysMV**, a dataset containing 40K scenes, each consisting of four orthogonal viewpoints, resulting in a total of 160K video sequences. Extensive experiments demonstrate that OrthoPhys significantly improves physical realism and spatial-temporal coherence over existing video generation methods. Project page: `https://anonymous.4open.science/w/Phys4D/`.

1 Introduction

Recent advances in video generation have significantly expanded the capability of visual synthesis models, enabling the generation of high-fidelity and temporally coherent videos conditioned on multimodal inputs such as text and images [1–5]. In particular, text- and image-conditioned video generation (TI2V) has emerged as a powerful paradigm for synthesizing dynamic visual content that aligns semantic intent with visual appearance. As perceptual quality improves, these models are increasingly seen as a foundation for interactive content creation, simulation-based reasoning, and embodied AI in dynamic visual environments.

Despite this progress, current video generation models still struggle to produce physically plausible and spatially consistent motion [6, 7]. While visually convincing at the pixel level, generated videos

Preprint.

often violate basic physical principles, exhibiting implausible accelerations, inconsistent object interactions, or motion patterns that are insensitive to intrinsic object properties such as mass or elasticity, as shown in Fig. 1(a). These issues become particularly evident in background-aware synthesis, where realistic interactions between moving foreground objects and the surrounding scene are essential.

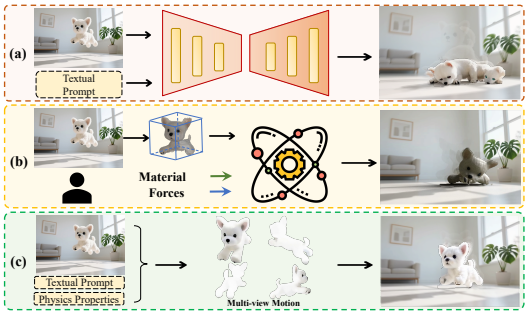


Figure 1: **Comparison with prior paradigms**. (a) Data-driven generative models. (b) Physics-engine-based methods. (c) Our physics-aware two-stage framework utilizing foreground multi-view videos.

We attribute these limitations to a mismatch between pixel-space motion modeling and physically constrained motion. Existing video generators typically learn dynamics as appearance transformations across frames, yet physically plausible motion and deformation are governed by geometry and material properties and should remain consistent across viewpoints. Explicit 3D representations and physics-based simulation [8–10] can produce valid trajectories, but they often rely on multi-stage pipelines that are difficult to scale and hard to integrate into end-to-end training. Instead, we consider a setting that directly outputs synchronized orthogonal-view foreground motion conditioned on physical attributes, and evaluates physical realism through both input-consistency with the specified physical properties and cross-view consistency, without explicitly outputting 3D structure or performing explicit 3D modeling.

In this work, we take a foreground-first view of physically grounded motion: we generate synchronized orthogonal-view foreground dynamics under explicit physical attributes, and then use them to guide final video synthesis. Building on this idea, we propose **OrthoPhys**, a two-stage framework that enables orthogonal-view geometry-guided motion modeling for physically plausible video generation without explicitly reconstructing 3D geometry. OrthoPhys adopts a two-stage design. In the first stage, the **Phys4View** module generates orthogonal-view foreground motion conditioned on a single image, a textual prompt, and explicit physical attributes. By injecting physical parameters directly into the attention mechanism and conditioning on control videos constructed from foreground masks and orthogonal-view priors, Phys4View incorporates physical and geometric cues into motion generation. We further introduce a geometry-aware cross-view attention module together with temporal attention to enhance cross-view alignment and spatiotemporal coherence, enabling simultaneous generation of synchronized four-view videos. In the second stage, the **VideoSyn** module synthesizes background-aware image-to-video results by leveraging the generated orthogonal-view motion as guidance, allowing the model to learn data-driven foreground–background interactions such as contact, occlusion, and context-consistent motion. To support training of the orthogonal-view video generation model, we employ a physics engine together with 3D-GS object representations to construct a physics-driven multi-view dataset, **PhysMV**, which comprises 10K 3D objects and 40K orthogonal video sets, totaling 160K videos.

Our contributions can be summarized as follows:

- We introduce **OrthoPhys**, a feed-forward framework for physically plausible video generation with orthogonal-view geometry guidance, conditioned on images, text, and explicit physical attributes.
- We design the **Phys4View** module, which integrates physics-aware attention with spatiotemporal and cross-view modeling to generate temporally coherent and geometrically consistent orthogonal-view motion.
- We develop the **VideoSyn** module for background-aware image-to-video synthesis, where orthogonal-view motion guidance enables data-driven learning of realistic foreground–background interactions without explicit physical simulation at inference.
- We construct a large-scale physics-oriented multi-view video dataset using a physics engine. We further show that OrthoPhys, leveraging orthogonal-view foreground motion generation followed by motion-aware video synthesis, achieves superior physical realism, temporal coherence, and cross-view consistency compared to prior video generation methods.

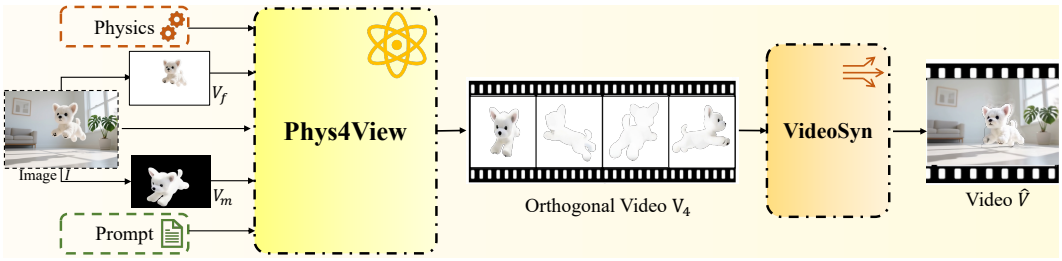


Figure 2: **Pipeline of OrthoPhys**. OrthoPhys generates physically plausible videos via a two-stage pipeline: 1) generating physics-aware orthogonal-view foreground videos, 2) generating a plausible video with the guidance of foreground motion.

2 Related Works

2.1 Controllable Video Generation

Video generation models trained on large-scale text–video paired datasets have demonstrated remarkable capabilities in synthesizing high-quality videos [11, 1, 3, 4]. Previous studies have shown that pretrained models can be additionally guided by various control signals, including camera motion [12, 13], point trajectories [14–16], and anchor-frame videos [17, 18], enabling more controllable video generation. Meanwhile, other research efforts have explored leveraging different modalities to guide motion-aware video synthesis. Some approaches [19, 20] employ optical flow videos as guidance, using flow maps to characterize motion dynamics and subsequently synthesize RGB videos based on them. Other methods [21, 22] use depth videos as conditioning inputs to guide the generation of temporally coherent video sequences. In addition, several text-driven approaches [23] attempt to infer more fine-grained descriptions of motion dynamics from textual instructions, thereby enabling more precise control over motion behaviors during video generation. However, these methods generally lack the modeling of physical laws and often produce results that violate basic principles of physical plausibility. To address this limitation, OrthoPhys generates physics-aware orthogonal-view videos as motion representations and subsequently renders them into plausible video outputs.

2.2 Physics-grounded Video Generation

Physics engines typically use the 3D representation of foreground objects as input to simulate and generate the 3D state at each time step, based on defined material and motion properties [8, 24]. Leveraging the renderable 3D-GS representation [25], some methods [8] propose using the Material Point Method (MPM) solver to update the 3D Gaussian representation at each time step after force-induced motion. However, due to the limitations of MPM, it is only suitable for simulating elastic objects. To address the shortcomings of MPM solvers, some approaches [26, 27] incorporate Position-Based Dynamics (PBD) solvers to simulate fluid motion. The use of physics engines requires additional definition of physical properties, and some methods [28, 29] propose using MLLMs to estimate properties such as Young's modulus, Poisson's ratio, and density. Furthermore, other methods [30–32] suggest utilizing the prior knowledge of pre-trained video generation models to iteratively optimize the material properties of objects through SDS loss [33] after rendering the video. Although physics-engine-based methods ensure physical plausibility, they require manual definition of simulation conditions such as material properties and external forces and depend on high-quality explicit 3D representations, limiting their ability to automatically generate high-quality videos from a single image.

3 Methodology

We propose **OrthoPhys**, a two-stage generation framework that first produces orthogonal-view videos with physically grounded motion and subsequently synthesizes visually realistic videos with foreground–background interactions, as shown in Fig. 2. Its orthogonal-view generation module, **Phys4View**, is presented in Sec. 3.1 and Sec. 3.2, followed by its video synthesis module, **VideoSyn**, in Sec. 3.3.

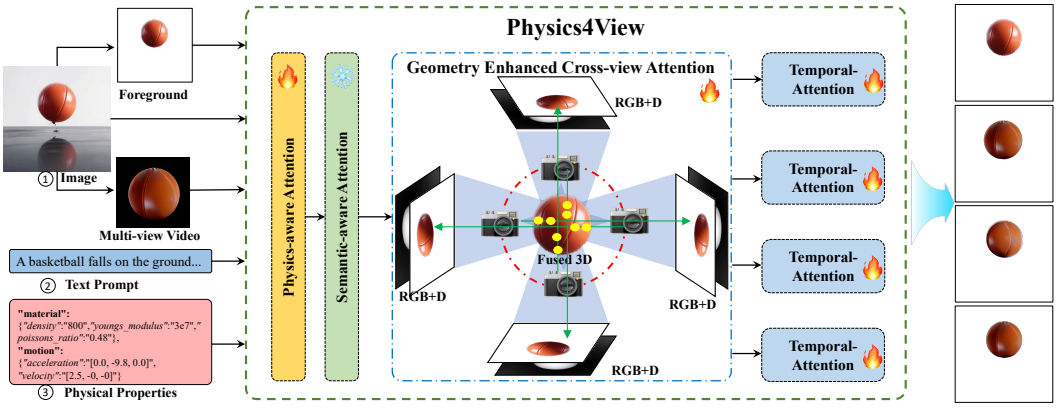


Figure 3: **Overview of Phys4View**. This stage of OrthoPhys incorporates physics-aware attention to model physical motion and introduces a geometry-enhanced cross-view attention module to generate spatially and temporally consistent orthogonal-view videos.

3.1 Physics Conditioning and Video Prior Fusion

Phys4View aims to generate temporally coherent and geometrically consistent orthogonal-view videos conditioned on a single image I, a textual prompt T, and physical attributes P. Given an input image containing both foreground and background regions, we first extract a foreground object mask M_f using a semantic segmentation model [34]. Based on M_f, we isolate the foreground image I_f and construct a foreground-initialized video V_f by replicating I_f across the temporal dimension. To provide 3D structural priors for subsequent video generation, we further employ a pretrained image-to-3D generation model [35] to obtain a static multi-view video V_m, which serves as geometric guidance.

Video prior conditioning. We condition video generation on two complementary video priors: a foreground-initialized video V_f and an orthogonal-view structural prior V_m. Both priors are encoded into latent features H_f and H_m, and are adaptively fused via a lightweight gating mechanism. Specifically, we predict a gating tensor from their concatenation and compute the fused representation as

$$H_v = G \odot H_f + (1 - G) \odot H_m, \tag{1}$$

where $G = \mathrm{Proj}(H_f \oplus H_m)$ and $\mathrm{Proj}(\cdot)$ is a lightweight 3D convolutional projector. We then inject H_v into the intermediate hidden representation H_i by replacing a designated subset of channels:

$$H_c = \mathrm{Inject}(H_i, H_v). \tag{2}$$

Physical and semantic conditioning. We condition generation on object-level physical attributes, including motion cues (e.g., acceleration $\mathbf{a}$ and velocity $\mathbf{v}$) and material properties (e.g., density ρ, Young's modulus E, and Poisson's ratio ν). To avoid semantic distortion from an expressive attribute encoder, we adopt a parameter-free tokenization scheme and obtain a unified physical conditioning token sequence C_{phys}.

We incorporate C_{phys} into the generator via a physics-aware attention module that modulates the intermediate representation H_c:

$$H_{\text{phys}} = H_c + \alpha_{\text{phys}} \cdot \mathrm{Attn}_{\text{phys}}([H_c; C_{\text{phys}}]), \tag{3}$$

where α_{phys} is a learnable scaling factor. For textual conditioning, we follow the standard text cross-attention used in the pretrained TI2V model [4]:

$$H = H_{\text{phys}} + \alpha_t \cdot \mathrm{Attn}_t(H_{\text{phys}}; C_p), \tag{4}$$

where C_p denotes the encoded prompt.

3.2 Geometry-enhanced Cross-view and Temporal Attention

To prove spatial and temporal consistency, we introduce a geometry-enhanced cross-view and temporal attention, as illustrated in Fig. 3.

Camera pose fusion. Before applying spatiotemporal attention, we inject camera pose information into the four-view latent features in a lightweight manner. Specifically, we encode the per-view camera pose into a learnable embedding, concatenate it with the corresponding latent along the channel dimension, and use a small convolutional block to fuse them, yielding pose-aware features for subsequent cross-view interaction.

Geometry-enhanced cross-view attention. To encourage spatially coherent cross-view interactions while avoiding explicit 3D representation, we introduce a geometry-guided bias into the cross-view attention mechanism based on predicted depth.

Given the latent feature map H, we predict a dense depth map at the same spatial resolution:

$$D = \mathcal{P}_\theta(H), \tag{5}$$

where $\mathcal{P}_\theta$ denotes a learnable depth prediction head. Each spatial location i in the latent grid, associated with coordinates (u_i, v_i), is lifted into the camera coordinate system via

$$\mathbf{x}_i^{\text{cam}} = D_i \, \mathbf{K}^{-1} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}, \tag{6}$$

where $\mathbf{K}$ is the camera intrinsic matrix.

Using the reconstructed 3D points in camera coordinates, we compute the pairwise geometric distance between locations i and j as

$$d_{ij} = \left\| \mathbf{x}_i^{\text{cam}} - \mathbf{x}_j^{\text{cam}} \right\|_2. \tag{7}$$

This distance is converted into a geometry-based affinity weight:

$$w_{ij,d} = \exp\left(-\frac{d_{ij}^2}{\tau} \right), \tag{8}$$

where τ controls the spatial sensitivity of the geometric prior.

Since depth prediction may be unreliable near occlusion boundaries or textureless regions, we further introduce a depth-confidence term to downweight ambiguous correspondences. Specifically, we define:

$$w_{ij,\text{conf}} = \alpha + (1 - \alpha) \left(1 - f_\theta \left(\frac{\|\nabla D_i\|_2 + \|\nabla D_j\|_2}{|D_i| + |D_j| + \epsilon} \right) \right), \tag{9}$$

where ∇D_i denotes the local spatial depth gradient at location i, $f_\theta(\cdot)$ is a learnable confidence mapping function, ϵ is a small constant for numerical stability, and α provides a lower bound on the confidence weight.

The final geometry-aware affinity between locations i and j is given by:

$$w_{ij} = w_{ij,d} \cdot w_{ij,\text{conf}}. \tag{10}$$

We incorporate this geometric prior into cross-view attention by augmenting the attention logits as:

$$s_{ij}^{\text{geo}} = \frac{Q_i \cdot K_j}{\sqrt{d}} + \log(w_{ij}), \tag{11}$$

where Q_i and K_j denote the query and key features at locations i and j, respectively. This formulation softly biases the attention mechanism toward spatially compatible and geometrically plausible regions under the predicted depth, while suppressing unreliable interactions.

Temporal attention. To further enhance temporal consistency in the generated videos, we introduce a lightweight temporal attention mechanism. Given the feature representation H, we first partition it into four disjoint components along the channel dimension as $H = [H_1;\, H_2;\, H_3;\, H_4]$, where $[\cdot;\cdot]$ denotes channel-wise concatenation.

For each component H_i, we apply an independent temporal self-attention operation to model temporal dependencies:

$$H_i' = H_i + \alpha_t \cdot \text{Attn}(H_i), \tag{12}$$

where $\text{Attn}(\cdot)$ denotes the temporal attention module and α_t is a learnable scaling factor that controls the strength of temporal aggregation.

Finally, the temporally enhanced features from all groups are concatenated to form the final representation $H_{\text{final}} = [H_1';\, H_2';\, H_3';\, H_4']$.

3.3 Motion-guided Plausible Video Generation

After obtaining a physics-aware four-view video V_4, our goal is to synthesize a visually coherent RGB video that captures realistic foreground–background interactions. To this end, we first extract the original-view video V_1 from V_4. We then apply a frame interpolation model [36] to densify the temporal resolution by interpolating four intermediate frames between consecutive frames, resulting in an extended video sequence V_{1i}.

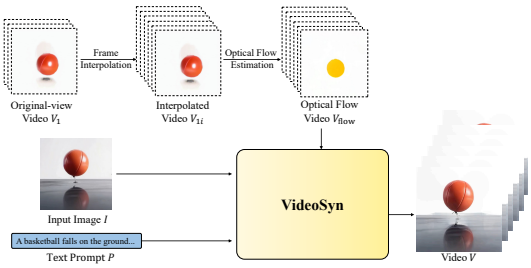


Figure 4: **Overview of VideoSyn**. The framework extracts foreground motion cues from a pre-generated physics-aware video and leverages them to guide the full video generation process.

To explicitly leverage the motion information encoded in V_{1i} as guidance for video generation, we adopt optical flow as a motion control signal. Specifically, we employ a pretrained optical flow estimation model [37] to compute the corresponding flow video V_{flow} from V_{1i}. The resulting optical flow captures fine-grained motion dynamics while remaining agnostic to appearance variations, making it a suitable representation for motion conditioning.

We then incorporate V_{flow} as an explicit motion control condition in the video generation process:

$$\tilde{V} = \text{Attn}([H;\, H_{\text{flow}};\, C_p]), \tag{13}$$

where H_{flow} is the encoded latent of the optical flow video V_{flow}. The attention module integrates appearance, motion, and semantic guidance for prediction.

The final video latent $\hat{V}$ is obtained through the diffusion denoising process:

$$\hat{V} = \mathcal{D}_\theta(\tilde{V}_t, t), \tag{14}$$

where $\tilde{V}_t$ denotes the noisy latent sampled at diffusion timestep t.

3.4 Training

PhysMV dataset. To facilitate the training of our physics-aware, orthogonal-view video generation model, we construct PhysMV, a large-scale dataset comprising 40K distinct motion sequences, each captured from four orthogonal viewpoints, resulting in a total of 160K single-view videos. To build up the dataset, we generate 10K foreground objects represented as 3D Gaussian Splatting (3D-GS) models using Trellis [35]. Additional dataset details are provided in Appendix C.

Phys4View training. We train the proposed Phys4View model using a standard diffusion objective. Specifically, the base training loss is defined as

$$\mathcal{L}_{\text{base}} = \mathbb{E}_{\tilde{F},t}\Big[w_1(t)\big\|\mathcal{D}_\theta(\tilde{F}_t, t) - \bar{F}\big\|_2^2\Big], \tag{15}$$

where $\tilde{F}_t$ denotes the noisy latent at diffusion timestep t, $\bar{F}$ is the target latent, and $w_1(t)$ is a timestep-dependent weighting function.

To provide geometric supervision, we additionally impose a depth consistency loss using relative depth estimates. Given the predicted depth map D and the estimated relative depth D_r, we define the depth supervision loss as

$$\mathcal{L}_{\text{depth}} = \left\| \frac{\text{Cov}(D, D_r)}{\sqrt{\text{Var}(D)\,\text{Var}(D_r)}} \right\|_1, \tag{16}$$

which measures the scale-invariant correlation between the predicted and reference depth maps.

VideoSyn training. To train VideoSyn, we adopt a diffusion-based image-to-video learning paradigm in which the model learns to generate a video conditioned on the motion guidance from the original view. We preprocess the OpenVid [38] and WISA [39] datasets and train the VideoSyn



Figure 5: **Qualitative comparisons.** The top row shows the given image, text prompt and corresponding physics settings. For the definition of velocity and acceleration directions, the positive x-axis is defined as horizontal left, the positive y-axis as vertical upward, and the positive z-axis as perpendicular to the paper plane inward.

Table 1: **Quantitative comparisons on WorldScore and VBench. Bold**: Best. Underline: Second Best.

Method	WorldScore			VBench			
	Photo↑	Motion↑	3D Consist↑	Motion↑	Subject↑	Flickering↑	Image↑
CogVideoX-I2V-5B [4]	64.72	67.39	74.43	0.993	0.886	0.989	0.636
Wan2.2-TI2V-5B [5]	65.97	64.90	63.86	0.994	0.926	0.991	0.653
Force Prompting [40]	74.38	72.53	50.97	0.994	0.938	0.992	0.641
OmniPhysGS [31]	58.67	90.65	37.24	0.995	0.929	0.995	0.378
PhysGen [41]	90.80	82.45	84.27	0.995	0.956	0.994	0.648
PhysGen3D [24]	86.68	86.34	86.89	0.996	0.910	**0.998**	0.605
Ours	**94.09**	**93.62**	**87.19**	**0.996**	**0.970**	0.996	**0.655**

model on both datasets. During training, we minimize the mean squared error between the denoised prediction and the ground-truth frame under random diffusion steps:

$$\mathcal{L}_2 = \mathbb{E}_{\tilde{R},\,t}\Big[w_2(t)\big\|\mathcal{D}_\theta(\tilde{R}_t, t) - \bar{R}\big\|_2^2\Big], \tag{17}$$

where $\bar{R}$ is the ground-truth RGB video. This training objective guides VideoSyn to synthesize temporally coherent and visually realistic video sequences that faithfully follow the motion guidance.

4 Experiments

4.1 Experimental Setups

Implementation details. Our model is trained on 8 NVIDIA A100 GPUs, each with 80GB of memory, while all inference experiments are performed on a single A100 GPU. To assess the effectiveness of the proposed approach, we construct a benchmark dataset consisting of 50 test samples. Each sample comprises a single RGB image I with a resolution of 720×480, an associated textual prompt P, and object-level physical attributes generated by GPT-4o. The benchmark set includes both rigid and deformable objects to evaluate the model's generalization across diverse physical regimes. All input images are synthesized in a photorealistic style using a text-to-image generation model [42]. For each input image, we obtain the foreground object mask M_f using Grounded-SAM [34]. Based on the extracted foreground, we further generate multi-view videos and corresponding 3D Gaussian Splatting (3D-GS) representations with Trellis [35], which are used for downstream evaluation. The comparison baselines are detailed in Appendix Sec. A.1.

Table 2: **GPT-4o and human evaluation results**. The left block reports GPT-4o scores, and the right block reports results from a fair blind user study with 59 participants. Both use three metrics: physical realism (Physical), photorealism (Photo), and semantic consistency (Semantic).

Method	GPT-4o			Human Study		
	Physical↑	Photo↑	Semantic↑	Physical↑	Photo↑	Semantic↑
CogVideoX-5B [4]	0.612	0.858	0.617	2.80	3.14	3.14
Wan2.2-5B [5]	0.609	0.877	<u>0.630</u>	<u>3.24</u>	<u>3.41</u>	<u>3.46</u>
ForcePrompt [40]	0.615	0.875	0.611	2.25	2.41	2.74
OmniPhysGS [31]	<u>0.622</u>	0.723	0.402	2.19	2.55	1.92
PhysGen [41]	0.547	0.808	0.550	2.17	2.07	2.76
PhysGen3D [24]	0.621	<u>0.879</u>	0.589	2.07	2.27	2.06
Ours	**0.627**	**0.885**	**0.636**	**3.47**	**3.52**	**3.53**

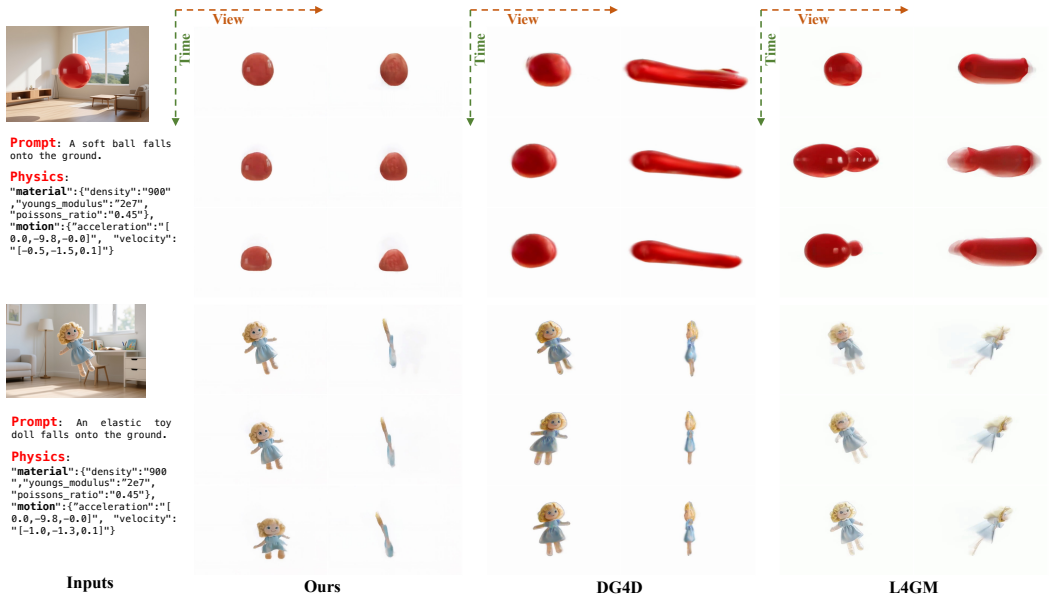


Figure 6: **Qualitative comparison of orthogonal-view video synthesis results**.

Metrics. We evaluate the quality of the generated videos using a combination of automatic metrics and VLM–based assessments. Specifically, WorldScore [43] is used to quantify photometric consistency (Photo), multi-view geometric consistency (3D Consist), and motion smoothness (Motion). Complementarily, we employ VBench [44] to assess motion quality (Motion), subject consistency (Subject), temporal flickering (Flickering), and overall visual fidelity (Image). To further measure adherence to the input prompts, following the evaluation protocol of PhysGen3D [24], we utilize GPT-4o to score the generated videos along three aspects: physical realism, photorealism, and semantic alignment with the textual prompt. We additionally conduct a fair blind user study with 59 participants under the same three criteria to complement the VLM-based assessment.

4.2 Comparisons with State-Of-The-Art Methods

Physics plausibility comparison. Quantitatively, as shown in Table 1, our method achieves superior motion coherence, consistently outperforming prior video generative models and showing comparable results with physical-engine-based methods across key dynamic metrics. Moreover, the generated videos exhibit higher visual quality than physical-engine-based methods. Regarding physical plausibility, as demonstrated in Table 2, our method surpasses all competing approaches under both GPT-4o evaluation and the blind user study, while simultaneously maintaining strong alignment with the input text and high semantic consistency. Fig. 5 shows two qualitative comparisons. Existing methods suffer from texture degradation, incorrect motion dynamics, or inconsistent object–background interactions, often leading to object replacement. In contrast, our method consistently generates physically plausible and coherent object motion. More comparisons can be found in Appendix Sec. D.

Table 3: **Evaluation of foreground orthogonal-view video synthesis.**

Method	View1			View2		
	Motion↑	Subject↑	Image↑	Motion↑	Subject↑	Image↑
DG4D [45]	<u>0.986</u>	0.892	0.268	0.987	0.883	0.260
Diffusion^2 [46]	0.970	0.637	<u>0.371</u>	0.979	0.666	<u>0.362</u>
L4GM [47]	0.994	<u>0.898</u>	0.368	<u>0.992</u>	<u>0.899</u>	0.305
Ours	**0.995**	**0.912**	**0.389**	**0.994**	**0.901**	**0.368**

Table 4: **Ablation study of attention modules in the orthogonal-view generation stage**.

Method	Motion↑	Image↑	Physical↑
Baseline*	0.994	0.594	0.636
+ Physics-Attn	0.994	0.613	0.682
+ Cross-View Attn	0.995	0.648	0.689
+ Temporal-Attn	**0.996**	**0.655**	**0.691**



Figure 7: **Results of object editing**.

Orthogonal-view video quality comparison. To evaluate the quality of the generated orthogonal-view videos, we compare our method with existing 4D generation approaches, including DG4D [45], Diffusion^2 [46], and L4GM [47]. For quantitative evaluation, we adopt VBench metrics to assess video quality from different views. As shown in Table 3, our method outperforms the competing approaches in terms of both motion consistency and visual quality. For qualitative comparison, representative results are shown in Fig. 6. Our method generates motion that follows physical principles, whereas existing 4D generation methods often fail to do so. Moreover, our approach effectively captures orthogonal-view structure, producing consistent results from novel viewpoints.

4.3 Ablation Study

We conduct ablation studies on the orthogonal-view generation stage of OrthoPhys to verify the effectiveness of its key components. Specifically, we incrementally incorporate physics-aware attention (Phys-Attn), geometry-enhanced cross-view attention (Cross-view Attn), and temporal attention (Temporal-Attn). We evaluate results using two VBench metrics, namely Motion and Image, together with a physical realism score by GPT-4o. As shown in Table 4, introducing physics-aware attention leads to a significant improvement in physical realism. Adding geometry-enhanced cross-view attention further enhances motion smoothness and image quality, while temporal attention provides additional gains in motion consistency. The qualitative ablation comparison is shown in Appendix Fig. 8. Additional ablation study on condition videos is detailed in Appendix Fig. 9.

5 Downstream Applications

While our primary focus is video generation rather than full 4D modeling, the orthogonal four-view foreground videos produced by our method naturally enable several downstream applications. As a demonstration, we show that the generated four-view foreground videos are well-suited for video editing. As illustrated in Fig. 7, we modify the rotation of the foreground object and leverage the foreground video segment from one of the generated views to guide the synthesis of an edited video. Additionally, although not a central goal of our method, for scenes with simple backgrounds, our approach can generate four-view videos with background, producing spatially consistent foreground–background renderings across views, as shown in Appendix Fig. 11.

6 Conclusion

In this work, we propose OrthoPhys, a two-stage framework for physically plausible video generation with orthogonal-view geometry guidance. OrthoPhys introduces physics-aware and geometry-enhanced inductive biases to improve motion consistency and physical plausibility without relying on explicit 3D representation. The framework combines physics-conditioned orthogonal-view motion

modeling with background-aware video synthesis, enabling controllable video generation. Extensive experiments demonstrate that OrthoPhys achieves superior motion coherence, visual quality, and physical realism compared to existing methods. In addition, we introduce a large-scale physics-oriented multi-view video dataset to facilitate future research.

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] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

[3] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[4] 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. In *The Thirteenth International Conference on Learning Representations*, 2025.

[5] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[6] 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 preprint arXiv:2406.03520*, 2024.

[7] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

[8] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4389–4398, 2024.

[9] Jiajing Lin, Zhenzhong Wang, Dejun Xu, Shu Jiang, Yunpeng Gong, and Min Jiang. Phys4dgen: Physics-compliant 4d generation with multi-material composition perception. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10398–10407, 2025.

[10] Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3d: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024.

[11] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

[12] Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. *arXiv preprint arXiv:2412.07759*, 2024.

[13] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

[14] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1–12, 2025.

[15] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025.

[16] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13–23, 2025.

[17] Zhifei Chen, Tianshuo Xu, Leyi Wu, Luozhou Wang, Dongyu Yan, Zihan You, Wenting Luo, Guo Zhang, and Yingcong Chen. Stance: Motion coherent video generation via sparse-to-dense anchored encoding. *arXiv preprint arXiv:2510.14588*, 2025.

[18] David Romero, Ariana Bermudez, Hao Li, Fabio Pizzati, and Ivan Laptev. Learning to generate object interactions with physics-guided video diffusion. *arXiv preprint arXiv:2510.02284*, 2025.

[19] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24142–24153, 2024.

[20] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2040–2049, 2025.

[21] Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8207–8216, 2024.

[22] Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1430–1440, 2024.

[23] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18826–18836, 2025.

[24] Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. Physgen3d: Crafting a miniature interactive world from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6178–6189, 2025.

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[26] Yutao Feng, Xiang Feng, Yintong Shang, Ying Jiang, Chang Yu, Zeshun Zong, Tianjia Shao, Hongzhi Wu, Kun Zhou, Chenfanfu Jiang, et al. Gaussian splashing: Dynamic fluid synthesis with gaussian splatting. *CoRR*, 2024.

[27] Yue Gao, Hong-Xing Yu, Bo Zhu, and Jiajun Wu. Fluidnexus: 3d fluid reconstruction and prediction from a single video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26091–26101, 2025.

[28] Haoyu Zhao, Hao Wang, Xingyue Zhao, Hao Fei, Hongqiu Wang, Chengjiang Long, and Hua Zou. Physsplat: Efficient physics simulation for 3d scenes via mllm-guided gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5242–5252, 2025.

[29] Haotian Mao, Zhuoxiong Xu, Siyue Wei, Yule Quan, Nianchen Deng, and Xubo Yang. Live-gs: Llm powers interactive vr by enhancing gaussian splatting. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 1234–1235. IEEE, 2025.

[30] Tianyu Huang, Haoze Zhang, Yihan Zeng, Zhilu Zhang, Hui Li, Wangmeng Zuo, and Rynson WH Lau. Dreamphysics: Learning physics-based 3d dynamics with video diffusion priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3733–3741, 2025.

[31] Yuchen Lin, Chenguo Lin, Jianjin Xu, and Yadong Mu. Omniphysgs: 3d constitutive gaussians for general physics-based dynamics generation. *arXiv preprint arXiv:2501.18982*, 2025.

[32] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pages 388–406, 2024.

[33] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[34] 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.

[35] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025.

[36] Guangyang Wu, Xin Tao, Changlin Li, Wenyi Wang, Xiaohong Liu, and Qingqing Zheng. Perception-oriented video frame interpolation via asymmetric blending. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2753–2762, 2024.

[37] Qiaole Dong and Yanwei Fu. Memflow: Optical flow estimation and prediction with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19068–19078, 2024.

[38] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

[39] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025.

[40] Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. *arXiv preprint arXiv:2505.19386*, 2025.

[41] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378, 2024.

[42] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[43] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.

[44] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[45] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

[46] Zeyu Yang, Zijie Pan, Chun Gu, and Li Zhang. Diffusion^2: Dynamic 3d content generation via score composition of video and multi-view diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

[47] Jiawei Ren, Cheng Xie, Ashkan Mirzaei, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, Huan Ling, et al. L4gm: Large 4d gaussian reconstruction model. *Advances in Neural Information Processing Systems*, 37:56828–56858, 2024.

[48] Yuanming Hu, Tzu-Mao Li, Luke Anderson, Jonathan Ragan-Kelley, and Frédo Durand. Taichi: a language for high-performance computation on spatially sparse data structures. *ACM Transactions on Graphics (TOG)*, pages 1–16, 2019.

[49] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

A Additional Experiments

A.1 Baseline Details

We compare OrthoPhys with the baselines reported in Table 1 and Table 2. The comparison is organized into two categories: physics-engine-based methods and learning-based video generation models. All methods are evaluated on the same benchmark samples and with the same metric pipelines described in Sec. 4.1; no baseline is given access to the orthogonal-view geometry guidance or foreground motion priors produced by OrthoPhys.

Physics-engine-based baselines.

- **PhysGen** [41]: a physics-grounded generation method that uses physical simulation to model object dynamics. We include it as a representative baseline for simulation-based physically plausible video generation.
- **PhysGen3D** [24]: a 3D-aware physics-based video generation baseline that combines physical simulation with explicit 3D representations. It serves as a strong reference for methods that improve physical plausibility through scene-level physical modeling.
- **OmniPhysGS** [31]: a Gaussian-splatting-based physical simulation method for dynamic scene generation. We include it to compare against approaches that explicitly simulate object dynamics in a 3D scene representation.

Learning-based video generation baselines.

- **CogVideoX** [4]: a strong open video generation backbone with an expert-transformer design. It serves as a general-purpose video generation baseline.
- **Wan** [5]: a recent high-quality video generation model used to assess whether strong visual generation capacity alone can produce physically plausible motion and object-background interactions.
- **Force Prompting** [40]: a physics-aware prompting method that augments video generation with explicit physical priors. We include it to compare against learning-based generation enhanced through textual or prompt-level physical guidance.

A.2 Additional Quantitative Physical Evaluations.

VideoPhy-2 evaluation. We further evaluate all comparison methods on VideoPhy-2 [6], a complementary benchmark for physical commonsense in video generation. As shown in Table 5, OrthoPhys achieves the best physical plausibility and semantic consistency scores among all evaluated methods, indicating that the gains observed on our benchmark also transfer to an external physical evaluation protocol.

Table 5: **Quantitative comparisons on VideoPhy-2**.

Method	Physical↑	Semantic↑
CogVideoX-5B	<u>3.06</u>	3.33
Wan2.2-5B	2.91	3.45
ForcePrompt	2.72	3.27
OmniPhysGS	2.21	3.53
PhysGen	2.69	<u>3.63</u>
PhysGen3D	2.73	3.57
Ours	**3.12**	**3.83**

Controlled physical controllability. To provide a more direct quantitative validation of physical controllability, we construct a controlled synthetic paired evaluation set generated by the physics simulator. Since the available supervision consists of predicted videos paired with reference motion videos, we report three proxy physical metrics that are closely tied to physical behavior: trajectory RMSE, contact timing error, and deformation-trend error. These metrics quantify motion alignment,

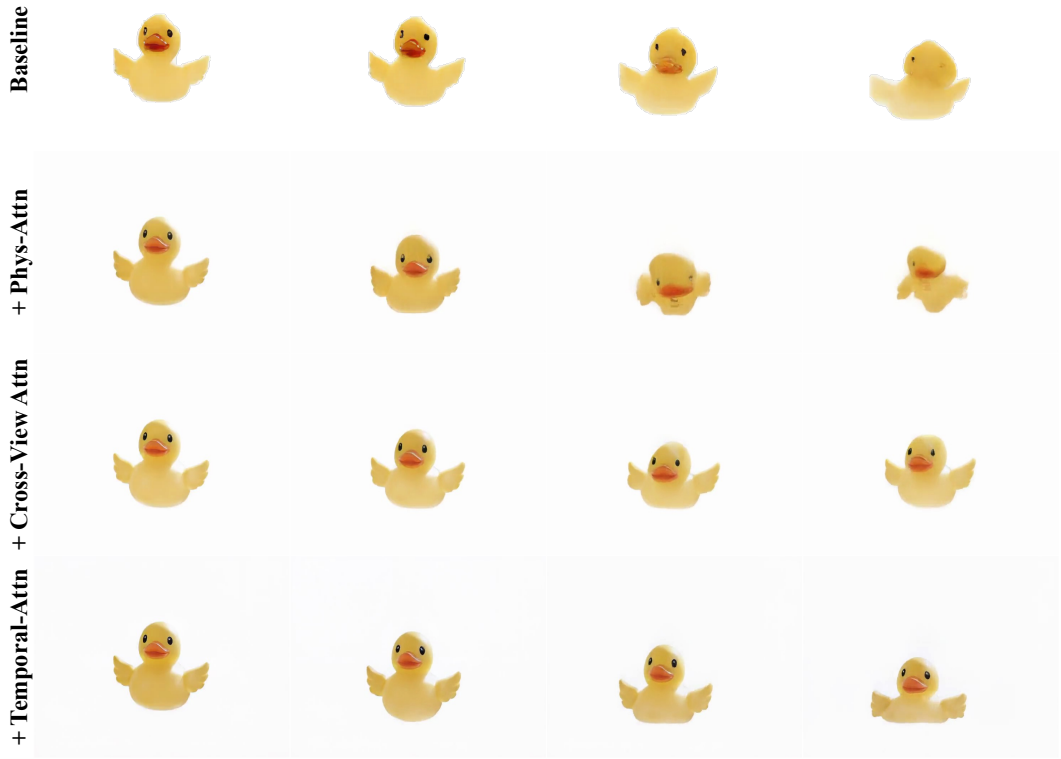


Figure 8: **Qualitative ablation results**. We compare the baseline with variants that incorporate individual components, including Phys-Attn, Cross-View Attn, and Temporal-Attn. Each row corresponds to one setting.

contact-event consistency, and deformation consistency, respectively. Contact timing error measures the deviation in the first-contact frame, while deformation-trend error measures the discrepancy between predicted and reference deformation evolution over time.

On our constructed 100-sample evaluation set, OrthoPhys consistently outperforms DG4D across all three physical motion metrics, as shown in Table 6. Although these are proxy metrics rather than full physical-state errors, they provide a direct quantitative link between physical inputs and generated motion.

Table 6: **Quantitative evaluation on the controlled synthetic evaluation set**. PSNR is reported as a reference reconstruction metric, while Traj. RMSE, Contact Timing Err., and Def. Trend Err. are proxy physical metrics.

Method	PSNR↑	Traj. RMSE↓	Contact Timing Err.↓	Def. Trend Err.↓
DG4D	18.32	0.1203	19.59	36.51
Ours	**22.24**	**0.0148**	**18.03**	**11.11**

A.3 Ablation Study of Attention Modules.

In the main manuscript, we report quantitative ablation results in Table 4 to evaluate the contribution of different attention modules. Here, we further provide qualitative ablation results in Fig. 8 to offer visual insights into their effects on orthogonal foreground video generation. As shown in Fig. 8, progressively introducing physics-aware attention, geometry-enhanced cross-view attention, and temporal attention leads to more coherent motion patterns and improved visual quality. These

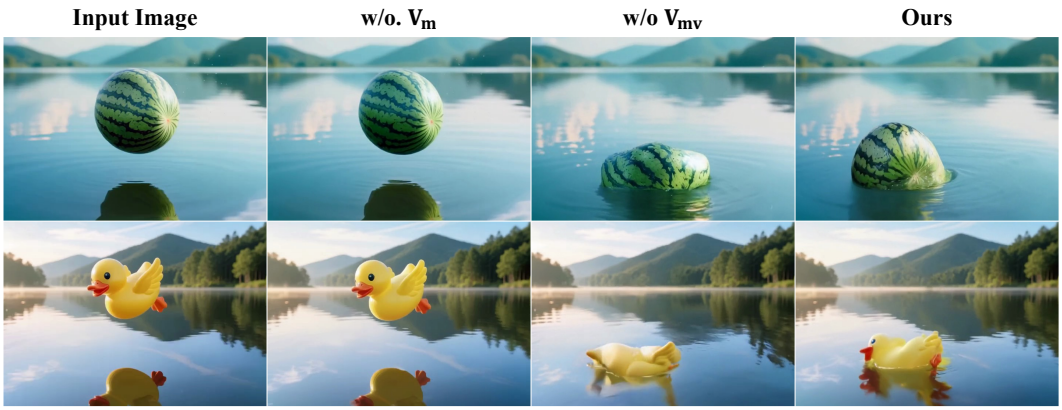


Figure 9: **Ablation study on conditional videos**. The left column shows the input single image, and the three right columns each display one representative frame from the generated RGB video.

qualitative results complement the quantitative analysis and further validate the effectiveness of each proposed component.

A.4 Ablation Study of Condition Videos.

We employ two types of video conditions to guide the orthogonal-view generation process. To evaluate the effectiveness of these conditions, we conduct an ablation study with qualitative comparisons shown in Fig. 9. The foreground mask video condition provides explicit guidance for foreground object motion, while the orthogonal-view video condition implicitly encodes 3D structural information, leading to more spatially consistent four-view video generation.

A.5 Supplementary Ablation Studies.

We provide additional ablation studies to isolate finer-grained design choices in OrthoPhys. Unless otherwise specified, we report Motion and Image scores from VBench and the Physical score from GPT-4o evaluation, following the metrics used in the main ablation study.

Finer-grained pipeline components. Table 7 evaluates several intermediate design choices beyond the major attention modules, including the mask prior, orthogonal-view prior, gated fusion, and depth-confidence weighting. Removing any of these components degrades the final performance, with a particularly clear drop in physical plausibility. This indicates that the gains come from a compact set of necessary design choices rather than arbitrary engineering details. Frame interpolation, optical-flow extraction, and flow-conditioned synthesis are supporting implementation steps: they align the guidance sequence, extract foreground motion, and transfer motion cues from the first-stage orthogonal-view results to full-scene synthesis without directly propagating appearance artifacts.

Table 7: **Ablation on finer-grained design choices in OrthoPhys**.

Method	Motion↑	Image↑	Physical↑
Ours (w/o mask prior)	0.991	0.646	0.609
Ours (w/o orthogonal-view prior)	0.992	0.637	0.625
Ours (w/o gated fusion)	0.995	0.651	0.648
Ours (w/o depth-confidence weighting)	0.994	0.642	0.643
Ours	**0.996**	**0.655**	**0.691**

Two-stage design. We compare the full OrthoPhys pipeline with a single-stage variant that uses the same physics conditioning but does not generate orthogonal-view foreground videos. As shown in Table 8, removing the orthogonal-view foreground stage consistently reduces all metrics. This

supports the importance of decomposing physically grounded foreground motion generation from the final background-aware synthesis stage.

Table 8: **Ablation study on the two-stage design**.

Method	Motion↑	Image↑	Physical↑
Ours (w/o orthogonal-view constraints)	0.983	0.636	0.628
Ours	**0.996**	**0.655**	**0.691**

Geometry-enhanced cross-view attention. Table 9 further analyzes the geometry-enhanced cross-view attention module. Cross-view attention alone already improves over removing cross-view interaction, while the full geometry-enhanced variant brings additional gains by better capturing spatial dependencies across synchronized orthogonal views.

Table 9: **Ablation study on geometry-enhanced cross-view attention**.

Method	Motion↑	Image↑	Physical↑
Ours (w/o cross-view)	0.994	0.594	0.636
Ours (only Cross-Attn)	0.994	0.640	0.678
Ours (w/ Geometry-enhanced Attn)	**0.996**	**0.655**	**0.691**

Physical parameter sensitivity. We conduct a controlled quantitative sensitivity study by fixing all non-physical factors and varying one physical parameter at a time. We report a motion deformation response metric, *Peak Deformation Proxy*, defined as the increase in post-contact spatial spread relative to the pre-contact baseline. For each parameter, we evaluate 30 controlled samples, sweep over 5 parameter values, and repeat each setting 3 times. Table 10 and Table 11 report the aggregated normalized deformation response. For Young's modulus, the response decreases as stiffness increases, consistent with the expected trend that stiffer materials deform less under comparable impacts. For density, the response increases monotonically with the density factor, indicating that the generated motion changes consistently with the specified physical parameter.

Table 10: **Sensitivity analysis for Young's modulus**.

Young's Modulus	1e5	3e5	1e6	3e6	1e7
Peak Deformation (norm.)	1.000	0.992	0.985	0.981	0.976

Table 11: **Sensitivity analysis for density**.

Density Factor	0.25x	0.5x	1.0x	2.0x	4.0x
Peak Deformation (norm.)	1.000	1.006	1.011	1.019	1.026

Dependency on LLM-estimated attributes. GPT-4o serves as a practical initializer for physical attributes from the input image and prompt, rather than an infallible source of supervision. To assess the dependency on LLM-estimated attributes, we compare physical attributes estimated by GPT-4o with those provided by human experts. As shown in Table 12, expert-provided attributes yield only marginal improvement over GPT-4o estimates, suggesting that GPT-4o provides sufficiently reliable physical initialization for OrthoPhys in most cases.

Table 12: **Ablation on GPT-4o-estimated versus expert-provided physical attributes**.

Method	Motion↑	Image↑	Physical↑
Ours (w/ GPT-4o)	0.996	0.655	0.691
Ours (w/ expert)	**0.996**	**0.658**	**0.694**

System Prompt

```
You are a helpful assistant that pays attention to context and estimates the perceptual quality of a
video based on physical realism, photorealism, and semantic consistency.
```

User Prompt

```
I would like you to evaluate the quality of generated videos based on the following criteria: physical
realism, photorealism, and semantic consistency. The evaluation will be based on 15 evenly sampled
frames from each video. Given the following instructions: '{instructions}', please evaluate the quality
of each video on the three criteria mentioned above. And for the definition of velocity and acceleration
directions, the positive x-axis is defined as horizontal left, the positive y-axis as vertical upward,
and the positive z-axis as perpendicular to the paper plane inward.

Note that: Physical realism focuses on input-conditioned physical fidelity: the object's motion and
deformation must be consistent with the physical properties specified in the input (e.g., motion
direction and deformation magnitude). Any noticeable mismatch with the provided physical attributes
should be considered a degradation in physical realism. Photorealism assesses the overall visual quality
of the video, including the presence of visual artifacts, discontinuities, and how accurately the video
replicates details of light, texture, and materials. Semantic Consistency evaluates how well the content
of the generated video aligns with the intended meaning of the text prompt.

Please provide scores from 0-1, with 1 being the best. The output must follow this exact format:
Physical Realism Score: [a score]; Photorealism Score: [a score]; Semantic Consistency Score: [a score]

Note that your output should strictly follow the above format, with a ';' after each score. Do not give
any other explanations.
```

Figure 10: **Prompt used for GPT-4o evaluation.**

A.6 Subjective Evaluation with GPT-4o and Human Study.

Since no widely accepted, physics-focused evaluation metrics for image-to-video generation currently exist, we follow the approach of PhysGen3D [24] and employ GPT-4o to conduct subjective assessments. These evaluations score multiple dimensions relevant to physical plausibility, including physical realism, photorealism, and semantic alignment. We adopt the evaluation prompt design originally proposed in PhysGen3D [24]; as their work has empirically demonstrated strong alignment with human judgment, we apply only minor adaptations to tailor the prompt to our dataset, without introducing substantial modifications. The complete evaluation prompt is illustrated in Fig. 10. In addition, we conduct a fair blind user study with 59 participants under the same criteria, whose aggregate results are reported in Table 2. Participants are shown anonymized generated videos in randomized order and asked to score physical realism, photorealism, and semantic consistency on a five-point scale; no personal or sensitive information is collected.

B Application of 4D Synthesis

Our method not only supports object editing while preserving a stable background, but also enables four-view video generation with simple backgrounds. As shown in Fig. 11, we render four-view videos from the front, left, back, and right viewpoints. The generated results exhibit strong cross-view consistency, and the interactions between foreground objects and background remain visually plausible. More 4D synthesis results can be viewed on our project page.

C Details of PhysMV Dataset

In the main manuscript, we introduce the PhysMV dataset. Here we provide further detailed information. We generate 10K 3D objects and reconstruct 10 diverse background scenes, each represented by a 3D-GS model that supports viewpoint exploration. Following the design of OmniPhysGS [31], we employ TaiChi [48] as the physics engine and adopt the Material Point Method (MPM) as the physics solver to simulate object dynamics conditioned on the specified physical properties. Leveraging the rendering capability of 3D-GS, we render four-view foreground videos after the physical simulation.

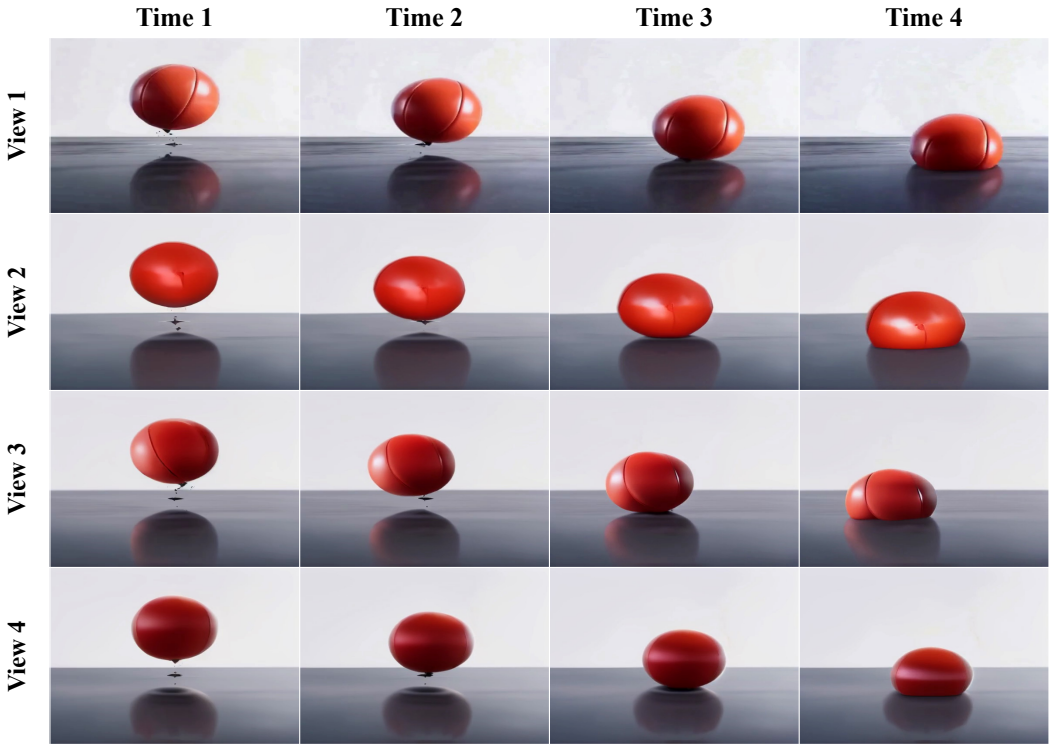


Figure 11: **Four-view video generation results**. Views 1 through 4 correspond to the front, left, rear, and right viewpoints, respectively.

Note that the generated four-view videos contain only foreground objects, as the background scenes do not support full 360-degree rendering across all views. The dataset examples are shown in Fig. 12.

For each simulated sample, the textual prompt, motion attributes, and material parameters used to control the physics simulation are automatically recorded as metadata. During post-processing, we apply Grounded-SAM [34] to obtain semantic foreground masks and DepthAnything V2 [49] to get relative depth maps. Finally, Trellis [35] is again utilized to generate multi-view videos of the segmented foreground objects, which are used to provide multi-view video priors.

D More Visual Comparisons with Baselines

As shown in Fig. 13, Fig. 14, Fig. 15, and Fig. 16, we provide additional visual comparisons with all baseline methods. These results highlight that our method achieves state-of-the-art quality, characterized by high visual fidelity, strong physical plausibility, and fine-grained controllability.

E Limitations and Future Work

While OrthoPhys introduces a novel framework for generating physics-aware orthogonal-view foreground videos with background interaction from a given viewpoint, it still exhibits several limitations. In particular, our approach is currently constrained in its ability to model and synthesize complex and highly dynamic backgrounds. Although the proposed method can generate diverse and physically plausible orthogonal-view foreground motions, synthesizing complete videos from novel viewpoints remains challenging when the background contains rich geometric structures or intricate appearance variations. This limitation primarily stems from the inherent difficulty of holistic scene modeling, which has not yet been fully addressed by existing 3D generation and reconstruction methods, and becomes even more challenging in the 4D setting where temporal dynamics must be consistently preserved. As a result, background reconstruction quality may degrade

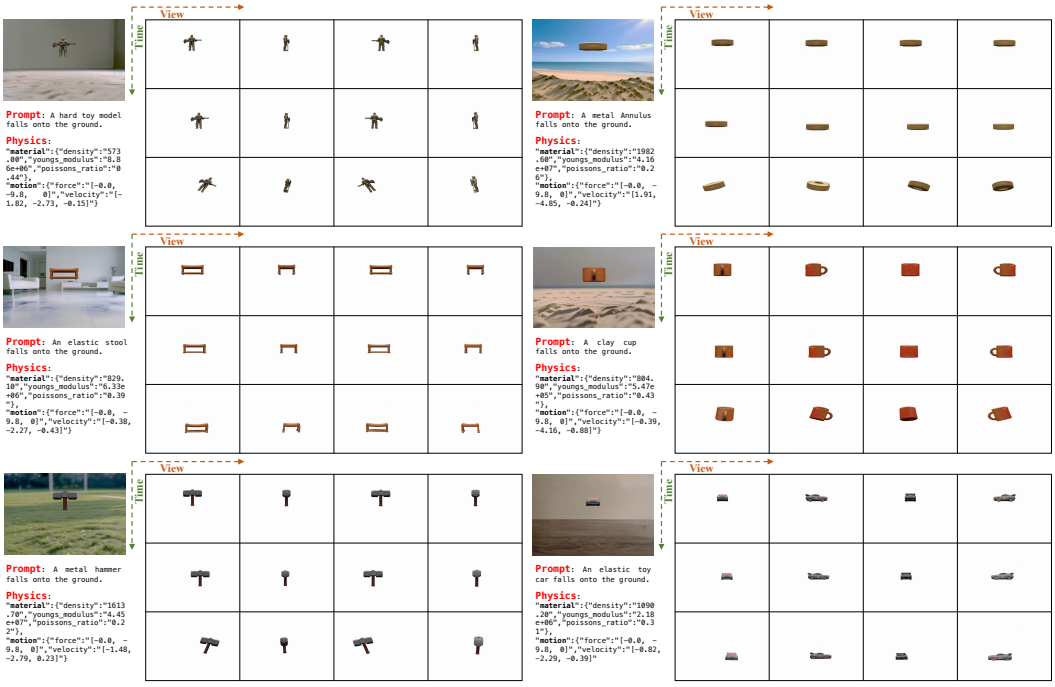
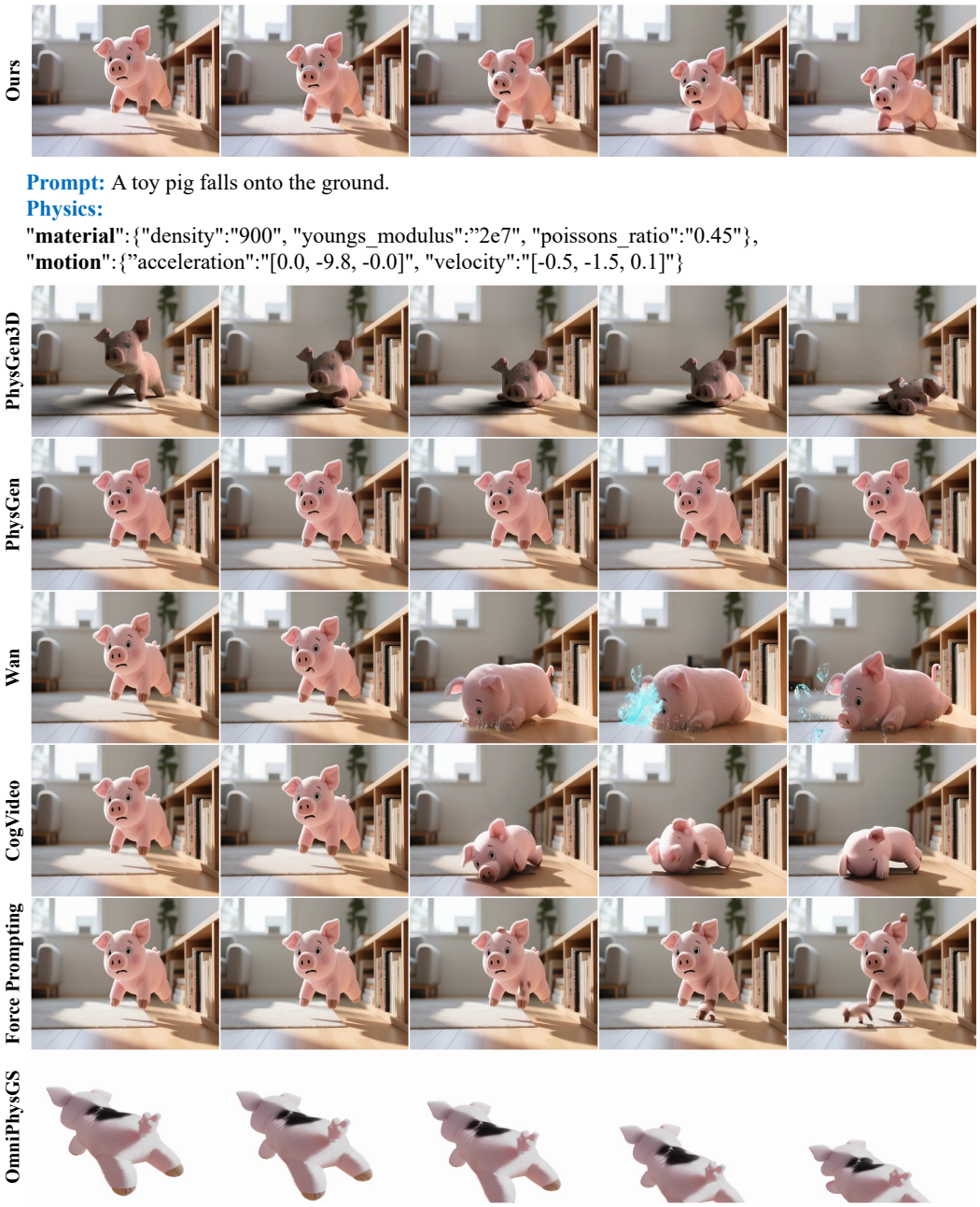


Figure 12: **Dataset demonstration**. Our dataset contains various objects and motion patterns.

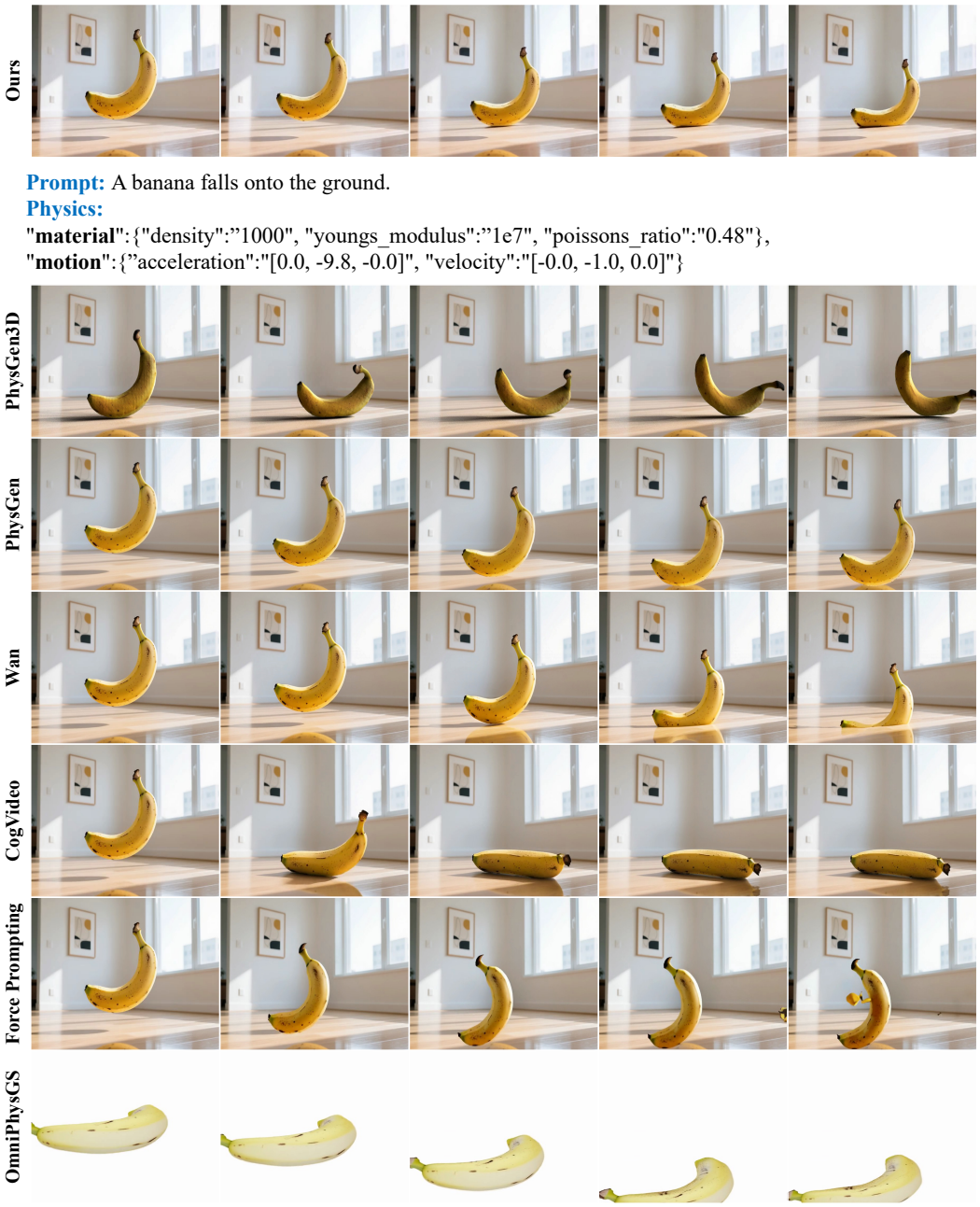
when extrapolating to unseen viewpoints, which in turn affects the overall visual coherence of the generated video.

In future work, we plan to explore more advanced scene representations and background modeling strategies to better capture complex geometry and long-range temporal consistency. Integrating explicit scene decomposition, multi-view supervision, or hybrid 3D-4D representations may help alleviate these limitations and enable more robust whole-scene generation from novel viewpoints.



Prompt: A toy pig falls onto the ground.
Physics:
"**material**":{"density":"900", "youngs_modulus":"2e7", "poissons_ratio":"0.45"},
"**motion**":{"acceleration":"[0.0, -9.8, -0.0]", "velocity":"[-0.5, -1.5, 0.1]"}

Figure 13: **Qualitative visualization results of our method and the baselines**.



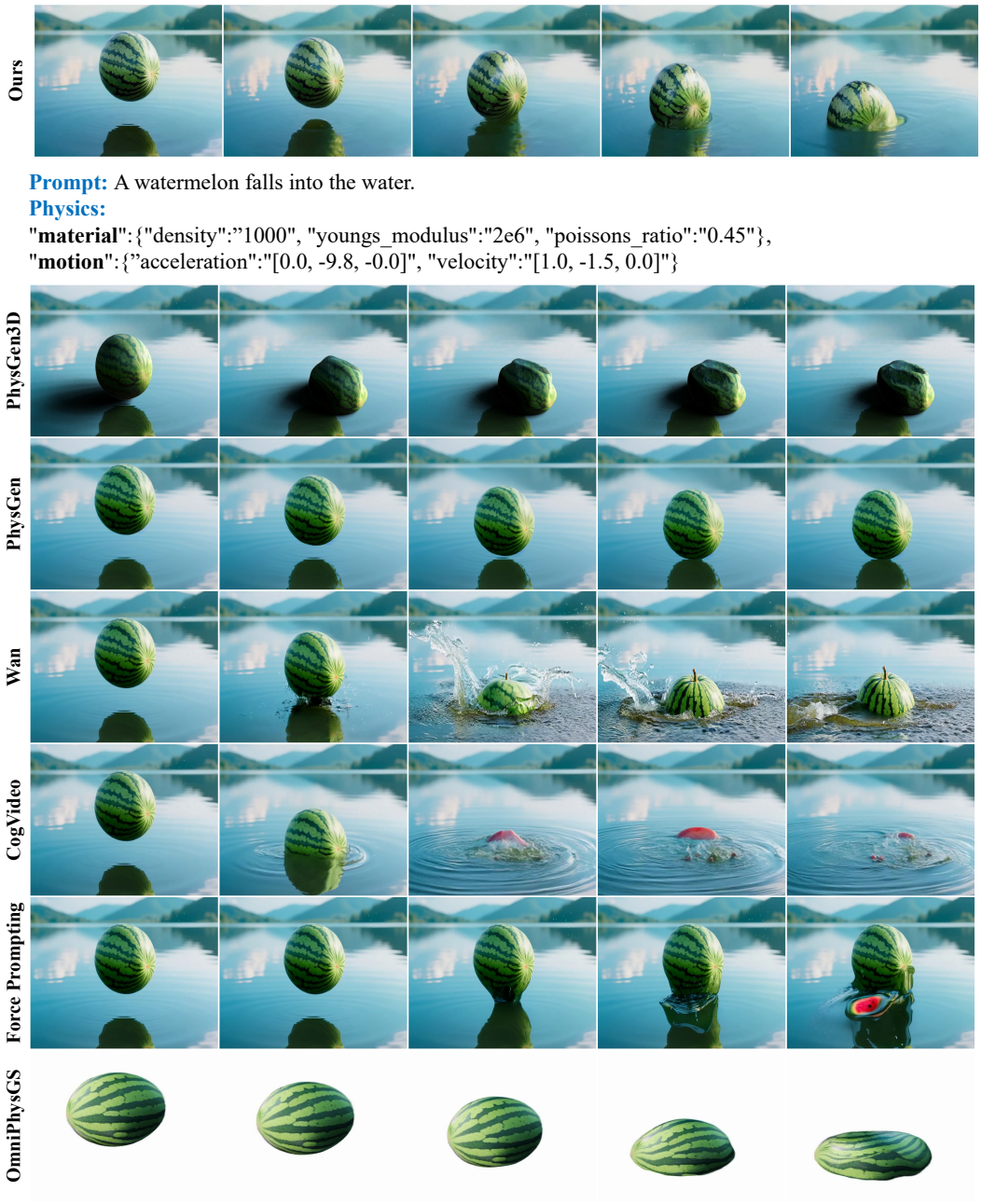
Prompt: A banana falls onto the ground.
Physics:
"**material**":{"density":"1000", "youngs_modulus":"1e7", "poissons_ratio":"0.48"},
"**motion**":{"acceleration":"[0.0, -9.8, -0.0]", "velocity":"[-0.0, -1.0, 0.0]"}

Figure 14: **Qualitative visualization results of our method and the baselines**.



Prompt: A cap falls onto the ground.
Physics:
"**material**":{"density":"1000", "youngs_modulus":"7e6", "poissons_ratio":"0.45"},
"**motion**":{"acceleration":"[0.0, -9.8, -0.0]", "velocity":"[-0.0, -0.5, 0.0]"}

Figure 15: **Qualitative visualization results of our method and the baselines**.



Prompt: A watermelon falls into the water.
Physics:
"**material**":{"density":"1000", "youngs_modulus":"2e6", "poissons_ratio":"0.45"},
"**motion**":{"acceleration":"[0.0, -9.8, -0.0]", "velocity":"[1.0, -1.5, 0.0]"}

Figure 16: **Qualitative visualization results of our method and the baselines**.