PHYSHUMAN: PHYSICS-AWARE HUMAN AND OBJECT INTERACTION VIDEO AND DYNAMICS JOINT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human–object interaction (HOI) video generation promises scalable creation of interactive human motions for graphics and embodied applications. However, despite rapid progress, most existing models primarily optimize for RGB fidelity while overlooking physical plausibility and actionable physical signals that govern human motion, often producing interpenetrations, implausible impulses, and motions that fail to vary coherently with physical settings and interactions. To tackle this issue, we propose **PhysHuman**, a physics-informed approach that jointly generates HOI video and dynamics by conditioning a diffusion-based video model on text prompt and first frame. By predicting torque-derived actuation maps as an auxiliary target, the model learns structured dynamic information, thereby improving the overall quality of the generated video without providing actuator torque information. Furthermore, to facilitate learning and evaluation of physics-informed HOI generation, we introduce **PhysHuman-Dataset (PHD)**, a controllable corpus constructed in simulation by executing HOI policies while sweeping object mass and friction, providing synchronized RGB, metrically accurate depth and actuation traces, together with geometry and physical-parameter annotations. Experiments show that PhysHuman improves visual quality and perceived physical plausibility over strong video generation baselines, while simultaneously unifying the generation and perception of HOI.

1 INTRODUCTION

Video generation for human–object interaction (HOI) is a challenging yet fascinating problem in graphics and embodied AI, with applications in video synthesis (Fan et al., 2025), character animation (Li et al., 2024; Taheri et al., 2022; Yi et al., 2024; Pan et al., 2025), and simulation-driven agent training (Xu et al., 2025b; Xiang et al., 2020). Recent studies further show that video diffusion models provide strong priors for geometry-related tasks (Chen et al., 2025c; Millán et al., 2026).

However, despite significant advancements in photorealism and temporal coherence achieved by recent diffusion-based video generation methods (Lin et al., 2024; Jiang et al., 2025; Wan et al., 2025; Blattmann et al., 2023; Yang et al., 2024), these methods often fail to ensure physical validity. Generated interactions frequently violate fundamental physical dynamics and contact constraints, resulting in issues such as interpenetration, implausible impulses, and accelerations that do not respond consistently to changes in physical settings (Chefer et al., 2025; Gillman et al., 2025).

A key obstacle to physics-aware HOI video generation is two-fold. First, most video generation pipelines are trained and evaluated primarily in pixel space: they are conditioned on appearance-centric cues and optimized for perceptual similarity, without explicit access to physically meaningful state, action, or parameters. This limitation is amplified in HOI, where rigid-body dynamics are governed by physical settings, high-frequency contact events, and articulated human actuation; capturing such phenomena typically requires closed-loop control and robust contact handling, whereas open-loop Newtonian rollouts or purely visual priors tend to drift and accumulate non-physical artifacts. Second, prior physics-informed generation has largely remained object-centric, e.g., combining differentiable simulation (such as MPM (Jiang et al., 2016), Warp, Brax) with learned 3D representations to reproduce material dynamics (Xie et al., 2024; Zhang et al., 2024a). While promising,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Given a reference image and a text prompt, PhysHuman jointly generates a human–object interaction video and its physics-interpretable dynamics signals. We visualize the synthesized RGB frames together with the corresponding actuation magnitude and direction maps, enabling physics-grounded HOI generation with explicit dynamics outputs. To support training and evaluation, we also build a synthetic data pipeline that simulates human–object dynamics in a physics simulator and construct the dataset PHD.

these methods can be sensitive to object-specific parameters, cover limited categories, and are not designed for articulated human control with dense, sustained contact. Compounding these modeling gaps, progress is also bottlenecked by data: existing HOI video datasets rarely provide synchronized actuation traces or per-scene physical parameters at scale, making it difficult to learn physically consistent behavior.

To address this gap, we propose **PhysHuman**, a physics-informed HOI generation framework that *jointly* synthesizes interaction videos and dynamics. Instead of optimizing RGB fidelity alone, PhysHuman predicts per-frame *actuation maps* (magnitude and direction), a 2D proxy derived from physically simulated control torques and spatially aligned with the human body in the video. Technically, we train a latent Diffusion Transformer (Peebles & Xie, 2023) into a multi-input, multi-output model: conditioned on a reference image and a text prompt, a single denoising process produces the RGB video together with the actuation magnitude and direction map sequences. We achieve this by concatenating the latent sequences of video and actuation maps with lightweight delimiter tokens, enabling unified latent-space diffusion while preserving the base model's generative prior. At inference time, PhysHuman requires only a reference image and text instruction, and outputs both the interaction video and its dynamics descriptors, enabling downstream physics-aware analysis.

To bridge this dataset gap and to support training and evaluation of physics-grounded HOI generation models, we introduce *PhysHuman-Dataset* (PHD), a controllable large-scale corpus for physics-grounded HOI generation. We construct PHD by executing HOI control policies in a physics simulator (Makoviychuk et al., 2021) while systematically sweeping physical parameters, namely object mass and friction. We export metrically accurate trajectories and render them in a high-quality engine (Unreal Engine 5) to obtain realistic videos paired with rich annotations: depth, geometry, actuation traces, and scene-specific physics parameters. In total, PHD contains over 15K clips spanning diverse interactions and parameter settings, enabling controlled studies of how HOI dynamics should vary as physics changes.

We evaluate PhysHuman on held-out PHD clips and on real-world HOI scenes. Results show that PhysHuman achieves higher visual fidelity than prior video generators and is preferred by human raters for physical plausibility, while additionally producing body-aligned actuation maps that RGB-only models do not provide. Moreover, by coupling actuation inference with video generation, PhysHuman unifies the *perception* and *generation* of HOI in a single physics-grounded pipeline.

In summary, our main contributions are:

- We propose PhysHuman, a physics-informed diffusion-based generative model that *jointly* synthesizes interaction videos and physics-interpretable actuation from a reference image and a text prompt. Building on joint appearance–motion generation (Chefer et al., 2025)

and temporal-concatenation conditioning (Bai et al., 2025), we use co-generated actuation as an auxiliary dynamics target that improves the generated HOI video.

- We introduce PHD, a large-scale dataset for physics-aware HOI video generation, constructed via physical simulation with systematic parameter sweeps and paired with photorealistic renderings and aligned actuation annotations.
- We introduce body-aligned actuation maps, a 2D torque-derived representation of human effort. Compared with geometric auxiliary targets such as depth maps and optical flow, co-generating actuation maps yields better scores on most visual-quality metrics and fewer interpenetration and motion artifacts in qualitative comparisons.

2 Related Work

2.1 Controllable Video Generation

Conditioned Video Generation. To enhance controllability of video generation, research has focused on incorporating guidance signals. Recent works integrated granular signals, including poses (Ma et al., 2024; Chang et al., 2025), trajectories (Yin et al., 2023; Zhang et al., 2025; Geng et al., 2025), 3D points (Gu et al., 2025), and camera parameters (He et al., 2024; Bai et al., 2025). Recent approaches also integrate physical concepts. Force Prompting (Gillman et al., 2025) conditions a video model on user-specified force vectors learned from synthetic data, so physics enters the generator only as a learned prior. PhysCtrl (Wang et al., 2025a) goes further by generating physics-grounded 3D point trajectories with a diffusion model trained on large-scale simulated animations under explicit physics constraints, and then conditioning video generation on these trajectories. Both target objects and materials driven by external forces, and neither models articulated humans whose motion is produced by internal joint actuation under dense contact.

Joint Representation Synthesis. To bridge the gap between visual realism and structure, recent methods explore co-generating video alongside depth (Xi et al., 2025; Yang et al., 2025b), motion trajectories (Chefer et al., 2025), or 3D geometry (Zhen et al., 2025; Gu et al., 2025). PhysHuman adopts this joint-generation paradigm, building on the joint appearance–motion formulation of VideoJAM (Chefer et al., 2025) and the temporal-concatenation design of ReCamMaster (Bai et al., 2025). Our contribution lies in the auxiliary target: instead of kinematic or geometric signals, we co-generate body-aligned maps of human joint actuation obtained from physics simulation.

2.2 Physics-Grounded Simulation

Physics-Integrated 3D Representations. Integrating physical laws into generative models has emerged as a critical direction to address the spectral and dynamic inconsistencies in pure pixel-based generation. A prominent line of research focuses on embedding physical properties directly into 3D scene representations to enable interactive dynamics. Representative works like PhysGaussian (Xie et al., 2024) embed continuum mechanics into 3D Gaussians for real-time generative dynamics. Similarly, PhysGen3D (Chen et al., 2025a) and PhysDreamer (Zhang et al., 2024a) leverage video generation priors to synthesize interactive miniature worlds or 3D objects that respond to physical interactions, successfully bridging the gap between static geometry and dynamic simulation.

Generative Physics and Control in Video. To move beyond static scenes to temporal video generation, recent works have explored using physical signals as conditions. PhysGen (Liu et al., 2024a) enables rigid-body physics-grounded image-to-video generation by inferring physical parameters, while WonderPlay (Li et al., 2025) facilitates dynamic 3D scene generation conditioned on actions. Closer to our work, Force Prompting (Gillman et al., 2025) and PhysCtrl (Wang et al., 2025a) demonstrate that video models can interpret physics-based control signals. PhyGenHOI (Benishu et al., 2026) targets physically aware 4D HOI: it drives a 3D Gaussian human with a motion diffusion model, simulates the object with MPM at generation time, and couples the two through contact-triggered re-simulation and video score distillation. In contrast to these methods, which take physics signals as inputs or run a simulator during generation, PhysHuman produces 2D videos from a single image and text without simulation at inference, and treats human actuation as a co-generated output modality that provides auxiliary supervision rather than a hard physical constraint.

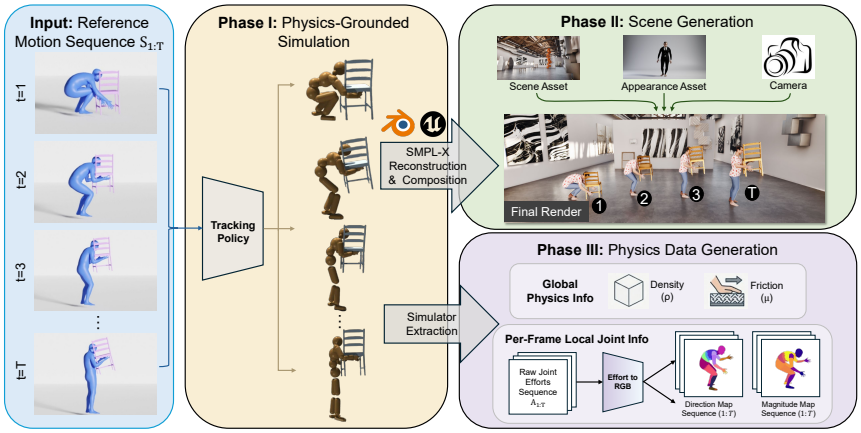


Figure 2: **Dataset construction pipeline.** We start by using a tracking-based policy InterMimic (Xu et al., 2025b) to generate both motion sequences and corresponding actuation signals. InterMimic takes reference human–object motion from InterAct (Xu et al., 2025a) as input and outputs per-joint PD target angles, which are converted by the simulator into actuator torques. The predicted motion sequences are reconstructed as SMPL animations and rendered in Unreal Engine 5 (UE5) (Epic Games, 2021) to produce the final videos.

3 METHOD

3.1 SMPL-BASED DYNAMICS FORMULATION

In HOI, we formulate human dynamics in the generalized-coordinate space of SMPL (Loper et al., 2023). For a given subject, the body shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{10}$ is treated as a time-invariant constant and therefore excluded from the generalized coordinates. The generalized coordinate stacks the floating base and the articulated pose, $\mathbf{q} = [\mathbf{t}, \mathbf{r}, \boldsymbol{\theta}] \in \mathbb{R}^{75}$, where $\mathbf{t} \in \mathbb{R}^3$ and $\mathbf{r} \in \mathbb{R}^3$ denote the global root translation and orientation, and $\boldsymbol{\theta} \in \mathbb{R}^{23\times 3}$ (flattened to $\mathbb{R}^{69}$) represents the articulated body pose in axis-angle representation for 23 joints; $\dot{\mathbf{q}}$ and $\ddot{\mathbf{q}}$ are defined accordingly. Applying the Euler–Lagrange equation to the Lagrangian $\mathcal{L}(\mathbf{q}, \dot{\mathbf{q}}) = T(\mathbf{q}, \dot{\mathbf{q}}) - V(\mathbf{q})$, with kinetic energy T and potential energy V, yields the standard floating-base rigid-body dynamics

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \mathbf{S}^\top \boldsymbol{\tau} + \mathbf{J}_c(\mathbf{q})^\top \boldsymbol{\lambda}, \tag{1}$$

where $\mathbf{M}(\mathbf{q}) \in \mathbb{R}^{75\times 75}$ is the generalized mass matrix, $\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}}$ captures Coriolis and centrifugal effects, and $\mathbf{g}(\mathbf{q})$ represents generalized gravitational forces. The six root coordinates are unactuated, so the selection matrix $\mathbf{S} = [\mathbf{0}_{69\times 6}, \mathbf{I}_{69}]$ maps the joint-level actuation $\boldsymbol{\tau} \in \mathbb{R}^{69}$ onto the pose coordinates only. Interaction with the ground and the manipulated object enters through the contact term $\mathbf{J}_c(\mathbf{q})^\top \boldsymbol{\lambda}$, where $\mathbf{J}_c$ is the contact Jacobian and $\boldsymbol{\lambda}$ stacks the contact forces. We reshape $\boldsymbol{\tau}$ into per-joint vectors $\boldsymbol{\tau}_j \in \mathbb{R}^3$, $j = 1, \ldots, 23$, aligned with $\boldsymbol{\theta}$; these are later used for actuator torque encoding.

3.2 PHYSHUMAN DATASET

We first introduce PhysHuman-Dataset (PHD), a large-scale physics-grounded dataset designed for HOI generation. Unlike traditional visual-based datasets, PhysHuman is constructed via a fully automated physics simulation pipeline. We leverage reference motions from the InterAct (Xu et al., 2025a) dataset and employ a robust tracking controller within IsaacGym simulator (Makoviychuk et al., 2021), ensuring the simulated motions are physically grounded. To capture diverse physical dynamics, we simulate each reference motion sequence with varying object masses and friction coefficients, explicitly generating distinct interaction effects for identical kinematic intentions. More details can be found in our Appendix.

3.2.1 DATA GENERATION PIPELINE

As shown in Figure 2, we propose a fully automated, physics-grounded data generation pipeline that produces pixel-aligned multimodal tuples for HOI generation. Unlike purely kinematic animation workflows, our pipeline explicitly simulates contact-rich human-object interactions in a dynamics environment, ensuring physical plausibility and enabling the extraction of synchronized visual observations (RGB), global physical parameters, and dynamics descriptors (actuation maps).

Reference Motion Acquisition. Firstly, we generate high-quality whole-body HOI reference motions. Specifically, we adopt InterAct (Xu et al., 2025a) as the primary motion source and focus on its OMOMO subset (Li et al., 2023), which contains complex full-body interactions with dense hand manipulation, foot contact, and diverse object contacts. These kinematic sequences serve as target trajectories for training a physics-based tracking controller in simulation.

Physics-Grounded Simulation and Parameterized Augmentation. Given the reference motions, we leverage InterMimic (Xu et al., 2025b) to train a robust tracking policy in the IsaacGym to drive the character through physically valid joint torques, interacting with objects under contact constraints. Our HOI includes complex actions like pushing, pulling, lifting, rotating, and repositioning objects. This pipeline yields diverse and high-quality physics-based HOI data with frame-wise joint-level actuation signals directly from the simulator. Besides, simulation provides controllable *parameter sweeps*. For each interaction intent, we procedurally vary global physical parameters, namely object mass (density) and surface friction, to generate motion variants under different force regimes. This augmentation expands the dataset from "lightweight grasping" to "effortful lifting" while keeping the underlying task semantics consistent, and provides ground-truth global parameters for each clip. The resulting human–object motion sequences are subsequently used for SMPL mesh reconstruction, and the per-joint actuator torque signals are used for actuation map construction.

Visual Synthesis and Domain Randomization. To achieve photorealistic rendering while preserving physically simulated trajectories, we export simulated motions to Blender for geometric animation reconstruction and then render them in Unreal Engine 5 (UE5) (Epic Games, 2021). We apply multi-factor domain randomization to improve generalization: (i) *Appearance diversity:* we randomize human textures, clothing, and materials using assets from BEDLAM (Black et al., 2023); (ii) *Environment diversity:* we render in a set of high-fidelity indoor and outdoor scenes as well as randomly sample backgrounds per clip; (iii) *Viewpoint diversity:* conditioned on the character's facing direction, we randomly sample camera positions and orientations within a bounded range to vary viewpoint while maintaining interaction visibility.

Physics-to-Visual Representation of Actuation. To expose latent physical driving forces to video generators, we first reconstruct the SMPL mesh from the recovered human motion sequences. Subsequently, we transform the per-joint actuator torque matrix into two spatially aligned 2D representations. Following the encoding spirit of PhysPT (Zhang et al., 2024b), at each frame we produce two RGB maps: (i) a *Direction Map*, whose three channels encode the 3D actuation direction (X/Y/Z components); (ii) a *Magnitude Map*, whose intensity encodes the actuation magnitude, reflecting effort and impulsiveness across joints. To align these maps with the rendered video, we project simulated body joints to the image plane using the same camera parameters as rendering, and rasterize per-joint actuation values into dense images. The resulting actuation-map sequences are temporally synchronized and pixel-aligned with RGB frames.

Output. The pipeline outputs a complete pixel-aligned tuple that shares the same set of camera parameters for each clip: *(1) photorealistic RGB video, (2) actuation-map sequences (direction & magnitude), and (3) ground-truth global physical parameters*. This dataset captures not only visual interaction details but also the underlying physical causal structure, enabling training and evaluation of physics-grounded HOI video generation models. In total, our dataset contains approximately **18 hours** of high-fidelity video data interacting with **15** everyday object categories. It features **100** diverse characters endowed with **1,700** distinct clothing assets and is rendered across **20** complex indoor and outdoor scenes using UE5.

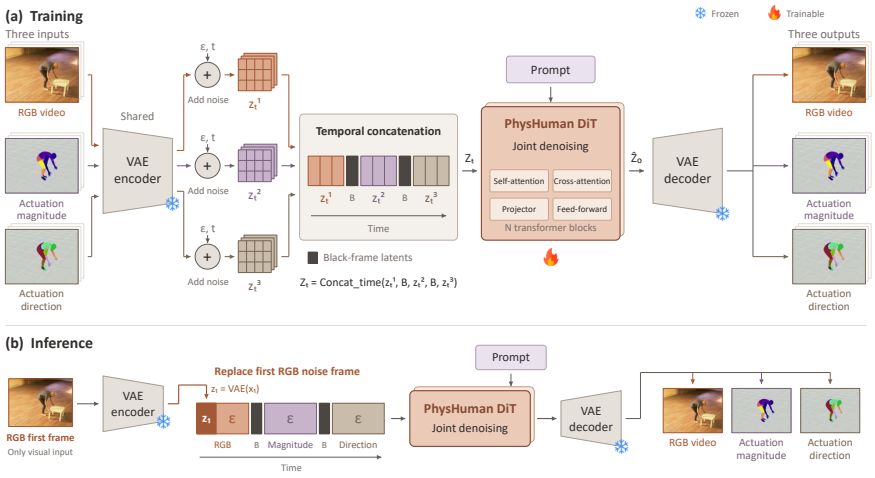


Figure 3: **(a) Training.** During training, the video frames, actuation magnitude, and actuation direction are separately encoded by the VAE and corrupted with noise, resulting in latent representations z_v, z_m, and z_n. The first frame of z_v is replaced with the image latent. The three latents are concatenated along the temporal dimension, with black-frame latents inserted to distinguish different modalities, and are then denoised by multiple DiT blocks. **(b) Inference.** During inference, we construct the noisy latent sequence and replace the first frame with the image latent. The sequence is then processed by multiple DiT blocks to generate the final multi-output results.

3.3 PhysHuman Model

3.3.1 Actuator Torque Encoding

Inspired by VideoJAM (Chefer et al., 2025), we represent our auxiliary signals, namely torque-derived actuation maps, as RGB videos to avoid the need for training an additional VAE. We convert the per-frame per-joint actuation magnitude $m_{t,j} = \|\boldsymbol{\tau}_{t,j}\|_2$ and the per-frame per-joint actuation direction $\mathbf{d}_{t,j} = \frac{\boldsymbol{\tau}_{t,j}}{\|\boldsymbol{\tau}_{t,j}\|_2+\epsilon}$ into video-aligned representations $\mathbf{V}_m \in \mathbb{R}^{H\times W\times 3}$ for magnitude and $\mathbf{V}_n \in \mathbb{R}^{H\times W\times 3}$ for direction. The actuation direction $\mathbf{d}_{t,j}$, expressed in the local frame of joint j, is mapped to RGB values after linear rescaling to the range $[0, 1]$, while scalar magnitudes are mapped to RGB using color mapping. Each SMPL vertex is colored according to the joint with the highest skinning weight, and the colored mesh is rendered into the camera frame with the same camera as the RGB video, so that $\mathbf{V}_m$ and $\mathbf{V}_n$ are pixel-aligned with the video.

PhysHuman is based on Wan2.2-TI2V-5B (Wan et al., 2025), a latent-space Diffusion Transformer (DiT) model for image-to-video generation. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and a text prompt $\mathbf{y}$, the model aims to generate video $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times 3}$ and the corresponding actuation magnitude map V_m, actuation direction map V_n with the same shape as video.

3.3.2 Model Architecture

Prior work on multi-modality joint generation has explored a range of architectures. VideoJAM (Chefer et al., 2025) concatenates multiple modality sequences along the channel dimension. In ablation studies, we find that this design tends to disrupt the distribution of the pretrained backbone, which in turn makes optimization unstable and training more difficult. Following the design of RecamMaster (Bai et al., 2025), we reformulate the model as a multi-input, multi-output framework by temporally concatenating the latent representations of the actuation magnitude map $\mathbf{z}_m$, the actuation direction map $\mathbf{z}_n$, and the target video $\mathbf{z}_v$.

Temporal Concatenation We project the actuation magnitude map $\mathbf{V}_m$, the actuation direction map $\mathbf{V}_n$, and the video $\mathbf{V}$ into the latent space using a shared VAE encoder $\mathcal{E}(\cdot)$, resulting in $\mathbf{z}_m = \mathcal{E}(\mathbf{V}_m)$, $\mathbf{z}_n = \mathcal{E}(\mathbf{V}_n)$, and $\mathbf{z}_v = \mathcal{E}(\mathbf{V})$, where all latent sequences share the same shape $(T', 48, H', W')$, with $T' = (T-1)/4 + 1$, $H' = H/16$, and $W' = W/16$ (e.g., $T = 81$ frames yields $T' = 21$ latent frames). Since all modalities are concatenated along the temporal dimension and fed into the model as a single sequence, the model may easily confuse adjacent modalities. To address

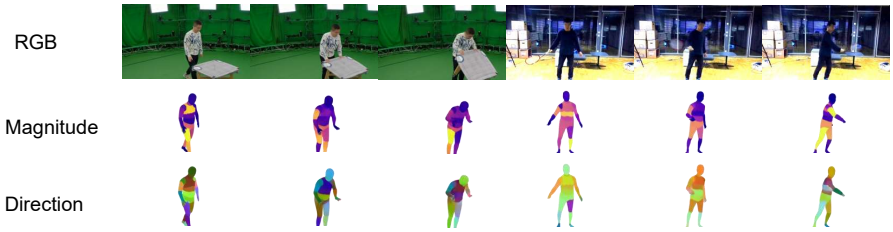


Figure 4: **Qualitative Results on Real World Data.** We visualize representative generated samples on real-world cases, showing generated RGB video frames alongside the corresponding predicted actuation maps (magnitude and direction). Our method produces visually realistic HOI interactions while generating temporally coherent, spatially aligned dynamics cues that reflect the underlying physical effort and contact events.



Figure 5: **Qualitative Results on Synthetic Data.** We visualize representative generated samples on synthetic data scenarios, showing generated RGB video frames alongside the corresponding predicted actuation maps (magnitude and direction).

this issue, we insert black-frame latents z_b between neighboring modalities, which serve as lightweight delimiters. The black-frame latent is obtained by encoding an all-black frame using the VAE. We then concatenate the three latents along the temporal dimension $z = [\mathbf{z}_v, \mathbf{z}_b, \mathbf{z}_m, \mathbf{z}_b, \mathbf{z}_n]$. The input image $\mathbf{I}$ is also encoded by the same VAE $\mathcal{E}(\cdot)$, and its latent feature replaces the first temporal slice of $\mathbf{z}$ to provide visual conditioning. During training, Gaussian noise is added only to the latent tokens, excluding the first frame and the black-frame delimiters, and the denoising loss is computed only over these noised latent positions. The resulting latent sequence $\mathbf{z} \in \mathbb{R}^{(3T'+2)\times 48\times H'\times W'}$ is then fed into the DiT backbone, which is trained with the flow-matching objective of the base model to predict the velocity field and iteratively denoises the sequence in the latent space. After denoising, the unified latent sequence is split into three segments according to the black-frame delimiters, corresponding to $\mathbf{z}_v$, $\mathbf{z}_m$ and $\mathbf{z}_n$, respectively, and each segment is decoded by the VAE to recover the final video and actuation outputs.

4 Experiments

4.1 Experimental Setup

Implementation details. All videos are resized to a spatial resolution of 480×832 for training efficiency. The input of the model is Prompt $\mathbf{y}$ and first frame image $\mathbf{I}$. All video captions are automatically annotated using Qwen3-8B-VL-Instruct (Yang et al., 2025a). The model generates videos with 81 frames at 15 FPS for each modality, equal to 243 frames in total. We train the model for 6.5K steps using a batch size of 32 and an AdamW optimizer with a learning rate of 1×10^{-5}.

Evaluation metrics. We report Fréchet Image Distance (FID) and Fréchet Video Distance (FVD) to measure distribution-level realism in image and video feature spaces, LPIPS to measure perceptual similarity to the ground-truth videos, and Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) to measure pixel-level fidelity.

Baselines. Since actuation represents a novel modality for which no existing generative models are available, this comparison focuses on the visual quality of the generated RGB videos. As baselines, we first compare our method with the pretrained Wan2.2-TI2V-5B model (Wan et al., 2025) without fine-tuning. We further compare against state-of-the-art open-source video generation

Table 1: **Quantitative Comparison**. Best in **bold**, second-best underlined.

Method	Video Generation Quality				
	FID↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓
CogVideoX1.5-5B-I2V	89.47	466.78	18.35	0.761	0.283
Wan2.2-TI2V-5B	86.05	496.22	16.56	0.593	0.422
HunyuanVideo-I2V	74.80	359.16	19.12	0.720	0.289
Ours	**53.76**	**184.31**	**22.39**	**0.835**	**0.163**

Table 2: **User study (favor %).** Higher is better.

Criterion	Ours	CogVideoX	Wan	HunyuanVideo
Video quality	**71.3**	10.7	9.0	9.0
Physical plausibility	**74.0**	9.3	7.7	9.0
Text alignment	**69.7**	11.0	9.7	9.7
Motion quality	**72.3**	9.0	9.3	9.3
Average	**71.8**	10.0	8.9	9.3

models, including CogVideoX1.5-5B-I2V (Yang et al., 2024) and HunyuanVideo-I2V (Kong et al., 2024).

4.2 Comparison

Qualitative Result on synthetic and real-world data. As shown in Figures 4 and 5, the actuation maps generated by PhysHuman are temporally coherent and spatially aligned with the body. The model also generalizes to a certain degree to real-world scenes, producing consistent body structures, faithful prompt following, and anatomically plausible motions.

Quantitative Evaluation. In the quantitative comparison (Table 1), we evaluate our method against several state-of-the-art zero-shot baselines. Our model consistently outperforms all competing methods across all evaluation metrics. This clearly demonstrates that training on the target dataset plays a crucial role in improving the overall visual quality, and enables the model to generate more realistic and coherent video results.

User Study. We conducted a user study with 30 participants on 10 video samples, comparing our method against CogVideoX, Wan, and HunyuanVideo under identical reference images and text prompts. For each set, participants performed a 4-way forced choice on four criteria: overall video quality, physical plausibility, text alignment, and interaction motion quality. As shown in Table 2, our method is preferred across all criteria. The largest gains appear in physical plausibility and interaction motion quality, suggesting that co-generating dynamics signals (actuation maps) helps reduce perceived physically implausible interaction artifacts while maintaining visual quality.

Table 3: **Ablation Study.** Best in **bold**, second-best underlined.

Method	FID↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓
single-output	70.14	287.37	21.90	0.827	0.189
channel concat	82.29	305.33	18.67	0.805	0.220
optical flow	59.70	219.48	22.34	0.825	0.155
metric depth	62.99	227.12	22.28	0.828	**0.152**
relative depth	63.35	226.44	21.97	0.822	0.162
Ours	**53.76**	**184.31**	**22.39**	**0.835**	0.163

4.3 Ablation study

Ablation on model design. We compare our temporal concatenation design with two alternatives. The first is a *single-output* baseline, which fine-tunes Wan2.2-TI2V-5B (Wan et al., 2025) on PHD with the same image and text inputs but generates only the RGB video, without actuation maps. The second is *channel concatenation*, in which we concatenate the three latent representations $\mathbf{z}_v, \mathbf{z}_m, \mathbf{z}_n$ along the channel dimension, i.e., $\mathbf{z} = [\mathbf{z}_v, \mathbf{z}_m, \mathbf{z}_n]_{\text{channel}}$. To distinguish different modalities, we further introduce a learnable modality embedding $\text{me} \in \mathbb{R}^{f\times h\times w}$ for each modality, and the fused latent becomes $\mathbf{z} = [\mathbf{z}_v + \text{me}_v,\ \mathbf{z}_m + \text{me}_m,\ \mathbf{z}_n + \text{me}_n]$. Meanwhile, we expand the channel dimension of the patch embedding W_{in} by a factor of three and the output head W_{out} accordingly, in order to accommodate the increased channel size of the latent representation. As shown in the first two rows of Table 3, our design outperforms both alternatives on all metrics; the improvement over the single-output baseline indicates that co-generating actuation maps benefits RGB generation itself. Compared with channel concatenation, temporal concatenation achieves better results and converges noticeably faster, as it keeps the input and output layers of the pretrained backbone unchanged.

Ablation on modality. To demonstrate that actuation map serves as a more effective latent representation for human–object interaction, we conduct an ablation study by comparing different latent

Table 4: **Physical plausibility evaluation.** Metrics are computed from human and object motion recovered from the generated videos. *Foot sliding*: horizontal displacement per frame of feet in contact with the ground. *Joint jitter*: mean magnitude of joint jerk. *Foot–ground* $|h|$: mean absolute height of the lowest foot vertex relative to the true floor, whether below or above it. *Ground pen.*: ground penetration, the depth of the lowest foot vertex below the true floor. *H–O non-pen.*: human–object non-penetration rate, the percentage of frames in which the human penetrates the object by less than 2 cm, evaluated on the clips whose ground-truth video can be reliably tracked (Appendix C). Grey rows are references: the simulator ground truth, and the ground-truth videos processed by the same estimators. Best in **bold**, second-best underlined, among PhysHuman and the auxiliary-modality ablations.

Method	Foot sliding↓ (mm/frame)	Joint jitter↓ (m/s^3)	Foot–ground $\lvert h\rvert$↓ (cm)	Ground pen.↓ (cm)	H–O non-pen.↑ (%)
GT (simulator)	3.77	31.5	–	0.04	–
GT video	5.00	23.1	4.17	1.69	86.2
optical flow	11.03	46.8	6.55	4.64	**79.0**
metric depth	11.23	46.1	6.68	5.19	74.9
relative depth	17.64	82.4	11.27	9.33	68.0
Ours	**9.76**	**37.7**	**6.41**	**3.96**	76.4

motion representations, including optical flow, metric depth, relative depth and actuation map. We use RAFT (Teed & Deng, 2020) to extract optical flow, Marigold (Ke et al., 2025) to estimate relative depth, and MoGe2 (Wang et al., 2025b) to obtain metric depth. As shown in Table 3, the actuation map achieves the best FID, FVD, PSNR, and SSIM among all auxiliary modalities, with the largest margin on FVD, while its LPIPS is slightly higher than that of the geometric variants. These results suggest that actuation provides richer and more informative motion cues, enabling the model to better capture human–object interactions; we evaluate the physical plausibility of the resulting motion directly in Sec. 4.4.

As highlighted in Figure 9 (Appendix G), optical flow suffers from motion blur, while metric depth suffers from human–object interpenetration, and relative depth produces physically implausible motion (e.g., the person bouncing upward). In contrast, the actuation map leads to more stable interactions, indicating that it captures physics-related cues that improve physical plausibility.

4.4 Physical Plausibility Evaluation

Visual-quality metrics do not directly measure whether the generated motion is physically plausible. We therefore recover human and object motion from the generated videos and compute physics-related statistics on the same test clips (Table 4). All quantities are expressed in metric units rather than pixels. Foot sliding, joint jitter, and human–object penetration are measured on reconstructed meshes of real-world size: a world-grounded SMPL-X body estimated by GVHMR (Shen et al., 2024), and the known object mesh tracked by FoundationPose (Wen et al., 2024). Foot–ground deviation and ground penetration are measured against the true floor, which we recover from the ground-truth camera height and pitch shared by all methods. Detailed metric definitions are provided in Appendix C.

Among the auxiliary-modality variants, PhysHuman achieves the lowest foot sliding, joint jitter, foot–ground deviation, and ground penetration, and the second-highest human–object non-penetration rate after optical flow. The advantage of optical flow on this metric is expected, as it directly supervises the dense motion of both the human and the object and thus their relative displacement at contact, whereas the actuation map describes only the human. Relative depth degrades all metrics substantially, and a gap to the GT-video reference remains on every metric.

5 Conclusion

We presented PhysHuman, which jointly synthesizes human–object interaction videos and physics-interpretable actuation maps with a single multi-output latent diffusion model. To support learning and evaluation, we introduced PHD, a large-scale synthetic dataset built via physics-based simulation and photorealistic rendering, with aligned actuation annotations and ground-truth physical parameters.

AI USE STATEMENT

In this work, we used generative AI tools for the following purposes. (1) *Data annotation:* the text captions of all videos in the synthetic dataset PHD were automatically generated with the vision-language model Qwen3-8B-VL-Instruct (Yang et al., 2025a), as described in Sec. B; the captions only serve as text prompts, while all motion, actuation, and physical-parameter annotations come directly from the physics simulator. (2) *Metric computation:* AI coding assistants were used to help implement the scripts that compute the evaluation metrics reported in the paper (FID, FVD, LPIPS, SSIM, and PSNR); the authors checked these scripts against standard reference implementations. (3) *Writing:* large language models were used to polish the language of the manuscript (grammar, wording, and readability). Generative AI tools were not used to generate research ideas, design the method, or produce experimental results. We have reviewed all AI-assisted work, and we take full responsibility for the final content of this paper, including its text, claims, and artifacts.

ETHICS STATEMENT

The training data of PhysHuman consist of synthetic human–object interaction videos rendered from physics simulation with parametric SMPL body models and virtual assets, so they contain no images of real people; real-world evaluation uses the publicly available InterCap dataset (Huang et al., 2022) under its original license. The user study involved volunteer participants who rated anonymized generated videos. As with other human video generation models, PhysHuman could potentially be misused to create misleading content depicting people; we will release the model and data for research purposes with usage terms that prohibit such applications.

REPRODUCIBILITY STATEMENT

The construction of the PHD dataset, including the simulation policy, physical parameter sweeps, actuation extraction, and rendering, is described in Sec. 3.2 and Fig. 2. The model architecture and training/inference procedures are described in Sec. 3.3 and Fig. 3. Implementation details, including data preprocessing, video resolution and frame rate, captioning, and torque acquisition in the simulator, are provided in Appendix B; the evaluation metrics are described in Sec. 4, and the physical plausibility metrics are defined in Appendix C. Additional qualitative results are available on the anonymous project website submitted with the paper.

REFERENCES

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. ReCamMaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025.

Omer Benishu, Gal Fiebelman, and Sagie Benaim. Phygenhoi: Physically-aware 4d generation of dynamic human-object interactions. *arXiv preprint arXiv:2605.30268*, 2026.

Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8726–8737, 2023.

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.

Di Chang, Hongyi Xu, You Xie, Yipeng Gao, Zhengfei Kuang, Shengqu Cai, Chenxu Zhang, Guoxian Song, Chao Wang, Yichun Shi, et al. X-Dyna: Expressive dynamic human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5499–5509, 2025.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

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*, pp. 6178–6189, 2025a.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22831–22840. IEEE, 2025b.

Wenyue Chen, Peng Li, Wangguandong Zheng, Chengfeng Zhao, Mengfei Li, Yaolong Zhu, Zhiyang Dou, Ronggang Wang, and Yuan Liu. SyncHuman: Synchronizing 2D and 3D generative models for single-view human reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025c.

Epic Games. Unreal Engine 5. `https://www.unrealengine.com/`, 2021.

Yingying Fan, Quanwei Yang, Kaisiyuan Wang, Hang Zhou, Yingying Li, Haocheng Feng, Errui Ding, Yu Wu, and Jingdong Wang. Re-HOLD: Video hand object interaction reenactment via adaptive layout-instructed diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17550–17560, 2025.

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*, pp. 1–12, 2025.

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.

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*, pp. 1–12, 2025.

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.

Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pp. 281–299. Springer, 2022.

Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. The material point method for simulating continuum materials. In *ACM SIGGRAPH 2016 Courses*, SIGGRAPH '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342896. doi: 10.1145/2897826.2927348. URL `https://doi.org/10.1145/2897826.2927348`.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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.

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pp. 54–72. Springer, 2024.

Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. WonderPlay: Dynamic 3D scene generation from a single image and actions. *arXiv preprint arXiv:2505.18151*, 2025.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-Sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4117–4125, 2024.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac Gym: High performance GPU-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Marc Benedí San Millán, Angela Dai, and Matthias Nießner. Animating the uncaptured: Humanoid mesh animation with video diffusion models. *International Conference on Learning Representations*, 2026.

Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. TokenHSI: Unified synthesis of physical human-scene interactions through task tokenization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5379–5391, June 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13263–13273, 2022.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

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.

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. PhysCtrl: Generative physics for controllable and physics-grounded video generation. *arXiv preprint arXiv:2509.20358*, 2025a.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025b.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879. IEEE, 2024.

Dianbing Xi, Jiepeng Wang, Yuanzhi Liang, Xi Qiu, Yuchi Huo, Rui Wang, Chi Zhang, and Xuelong Li. OmniVDiff: Omni controllable video diffusion for generation and understanding. *arXiv preprint arXiv:2504.10825*, 2025.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. SAPIEN: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11097–11107, 2020.

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*, pp. 4389–4398, 2024.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, et al. InterAct: Advancing large-scale versatile 3D human-object interaction generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7048–7060, 2025a.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12266–12277, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Lehan Yang, Lu Qi, Xiangtai Li, Sheng Li, Varun Jampani, and Ming-Hsuan Yang. Unified dense prediction of video diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28963–28973, 2025b.

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.

Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*, pp. 246–263. Springer, 2024.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. DragNUWA: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

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*, pp. 388–406. Springer, 2024a.

Yufei Zhang, Jeffrey O Kephart, Zijun Cui, and Qiang Ji. PhysPT: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2305–2317, 2024b.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2063–2073, 2025.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. TesserAct: learning 4D embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

A Appendix Overview

This appendix provides additional details on experiments in Section B, defines the physical plausibility metrics in Section C, discusses the limitations of the current framework and possible future extensions in Section D, and presents the dataset overview together with more qualitative results on synthetic and real-world data in Sections E and F, respectively, and qualitative comparisons of auxiliary modalities in Section G. Please refer to the anonymous website submitted with the paper for additional results.

B Experiment Details

All videos are sampled at 15 FPS. The original rendered sequences are produced at a resolution of 1920×1080 and are resized to a spatial resolution of 480×832 for training efficiency. The input to the model consists of a text prompt $\mathbf{y}$ and the first-frame image $\mathbf{I}$. All video captions are automatically annotated using Qwen3-8B-VL-Instruct (Yang et al., 2025a).

To preserve the generative capability of the base model, the trainable parameters are restricted to the linear layers (query, key, value, and output projections) of the self-attention and cross-attention modules, while all other parameters are kept frozen. In addition, we apply a random dropout rate of 5% to the text prompt during training to improve robustness and prevent overfitting to specific linguistic patterns. Training takes about one day on 16 NVIDIA H100 GPUs.

To increase visual diversity, we apply extensive scene-level augmentations. Specifically, we vary clothing assets using BEDLAM (Black et al., 2023), employ multiple indoor and outdoor environments rendered in Unreal Engine 5 (UE5) (Epic Games, 2021), and randomize lighting directions, camera poses, object masses, and friction coefficients. For camera placement, we fix the distance between the camera and the human subject and randomly sample camera positions within the frontal viewing solid angle of the subject, ensuring diverse viewpoints while maintaining clear visibility of the human–object interaction.

For per-joint actuator torque acquisition, we enable PhysX DOF force sensors in Isaac Gym (Makoviychuk et al., 2021) and retrieve per-step joint effort (generalized force) measurements via `acquire_dof_force_tensor` followed by `refresh_dof_force_tensor`. The resulting 1D float tensor provides one scalar value per DOF, corresponding to torque (in N·m) for revolute joints and force (in N) for prismatic joints.

Actuation maps. For each joint j, the torques of its three revolute DOFs form the vector $\boldsymbol{\tau}_j$. The magnitude map encodes $\|\boldsymbol{\tau}_j\|_2$ after min–max normalization, mapped to RGB with the plasma colormap. The direction map encodes the unit vector $\boldsymbol{\tau}_j/\|\boldsymbol{\tau}_j\|_2$, expressed in the local frame of joint j, whose components are linearly rescaled from $[-1, 1]$ to $[0, 1]$ as RGB values. The torque measured at an unloaded joint is never exactly zero but a small residual value; we treat such values as zero, so these joints take the lowest color of the colormap in the magnitude map and a neutral gray in the direction map. Each SMPL vertex takes the color of the joint with the highest skinning weight, and the colored mesh is rendered with the same camera as the RGB video.

Test split. We hold out 50 clips of PHD as the test set. The test clips are randomly sampled while keeping their scenes, characters, and actions as different from each other as possible.

C Physics Metric Details

We detail the physical plausibility metrics in Table 4. All metrics are computed on the 81 generated frames of each test clip, and near-blank degraded frames are excluded for every method.

Human reconstruction. We run GVHMR (Shen et al., 2024) on each generated video to obtain a world-grounded SMPL-X body sequence in a gravity-aligned frame. Because the SMPL-X body model has real-world dimensions, its mesh and joints provide metric scale without access to depth.

Foot sliding and joint jitter. A foot joint (left/right ankle and foot) is considered in contact when its height is below its 5th-percentile height over the clip plus 2 cm. *Foot sliding* is the mean horizontal displacement of a foot joint between consecutive contact frames (mm/frame). *Joint jitter* is the mean norm of the third finite difference of the 22 body joint positions, scaled by the cube of the frame rate

(m/s^3). For the simulator reference, the same rules are applied to the recorded rigid-body positions of the corresponding bodies.

Foot–ground deviation and ground penetration. Since every method is conditioned on the same ground-truth first frame, all generated videos share the ground-truth camera, whose roll, pitch, and height above the floor are known from rendering. This defines the true floor plane in camera coordinates: the height of a camera-space point $\mathbf{p}$ above the floor is $\mathbf{u}^\top \mathbf{p} + H$, where $\mathbf{u}$ is the world-up direction given by the camera pitch and H is the camera height. As the camera-space body estimated by GVHMR carries a per-clip scale error, each clip is rescaled once at the first frame, which is identical for all methods, so that its lowest foot vertex matches the simulator's lowest foot height at that frame. For every subsequent frame, let h be the height of the lowest foot vertex above the true floor. *Foot–ground deviation* $|h|$ is the mean of $|h|$ over frames, and *ground penetration* is the mean of $\max(0, -h)$.

Human–object non-penetration rate. We obtain the object mask in the first frame from a Grounding DINO (Liu et al., 2024b) detection of the object category prompted into SAM2 (Ravi et al., 2025), and propagate it through the clip with SAM2. Metric depth is estimated with Video Depth Anything (Chen et al., 2025b) and rescaled per clip to agree with the rendered depth of the GVHMR body. FoundationPose (Wen et al., 2024) registers the known object mesh in the first frame and tracks it through the clip, re-registering whenever the IoU between the rendered object silhouette and the SAM2 mask drops below 0.5. Human vertices are then mapped into the object frame, an inside test is performed with the generalized winding number, and the penetration depth of a frame is the maximum distance of the inside vertices to the object surface. To avoid penalizing methods for failures of the estimators themselves, we evaluate this metric only on the 36 test clips in which the object is successfully tracked in at least 80% of the evaluable frames of the ground-truth video; in the remaining clips, tracking fails even on the ground truth. A frame is evaluable if at most 30% of the object is occluded or out of view. For each clip and method, the non-penetration rate is the percentage of evaluable frames whose penetration depth is below 2 cm, and frames in which the object cannot be tracked count as failures; the rates are then averaged over clips. We use a 2 cm threshold rather than a stricter one because the reconstructed human and object meshes carry estimation errors of this order, as reflected by the ground-truth videos, which do not reach 100% even though they contain no penetration.

D Limitations and Future Works

Currently, the model is only capable of predicting a 2D representation of the actuation map, meaning that we cannot directly obtain the underlying 3D per-joint actuator torques. Consequently, it is difficult to convert the predicted 2D representation into explicit actuator torque values for real-world deployment.

In addition, our current framework does not enforce constraints from SMPL-based human body models, which makes the reconstructed human meshes less stable in certain cases.

Finally, PhysHuman is trained only on simulated and rendered data. Real-world videos differ in appearance, body shape, contact behavior, camera motion, and object categories, and no ground-truth actuation is available for them. Our real-world results are therefore qualitative, and the predicted actuation maps should not be interpreted as accurate measurements of real human effort.

In future work, we aim to extend the generated 2D representations to 3D representations, for example by representing actuation signals as a weighted point cloud, where the weights correspond to actuator torque magnitudes. Such a representation would also facilitate a more direct evaluation of actuator torque generation quality. Another direction is to use PhysHuman to generate data for robot learning, for example humanoid–object interaction videos paired with actuation maps that could supervise whole-body manipulation policies.

E Dataset Overview

Fig. 6 presents an overview of the dataset. Our dataset covers both indoor and outdoor environments and contains diverse human–object interactions, including pushing, lifting, kicking, and object repositioning.

F More Qualitative Results

We present additional qualitative results to further illustrate the capabilities of our method. Fig. 7 shows examples generated on the synthetic PhysHuman dataset, covering diverse human–object interactions such as pushing, pulling, kicking, and repositioning objects. Fig. 8 shows examples generated on InterCap (Huang et al., 2022). These examples demonstrate that the model can generate temporally coherent human motions and realistic object responses in various environments.

G Qualitative Modality Ablation

Fig. 9 qualitatively compares the auxiliary modalities studied in Sec. 4.3.

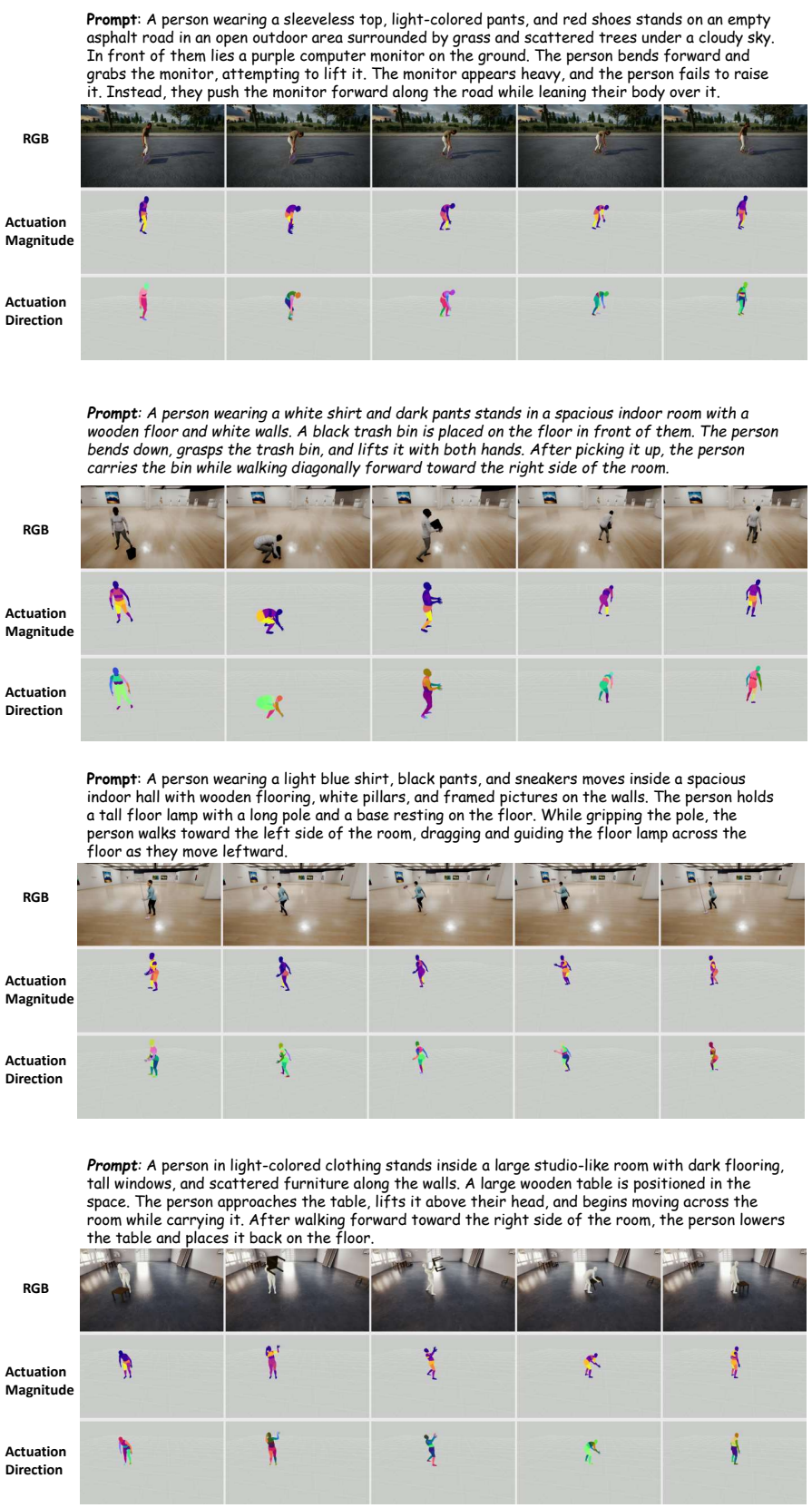


Figure 6: **Dataset Overview.** Each row shows the RGB video, actuation magnitude map, and actuation direction map, respectively. The dataset covers diverse human-object interaction actions across multiple scenes, demonstrating substantial diversity in both motion patterns and visual appearances.

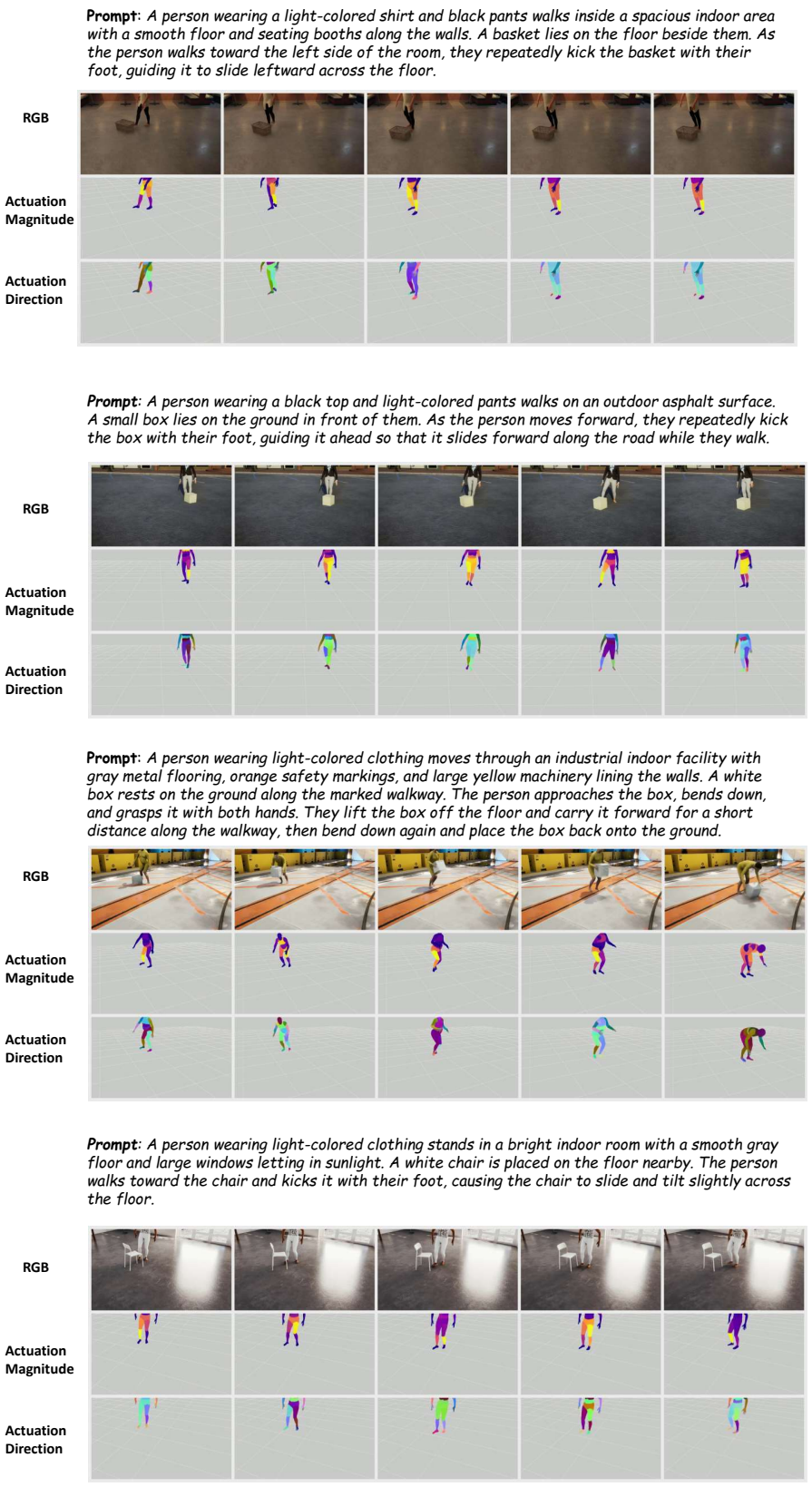


Figure 7: **More Qualitative Results on Synthetic Data.** We present additional results on the PHD test set. The generated outputs demonstrate that our model can effectively generalize to synthetic data while producing visually coherent videos and corresponding actuation maps.

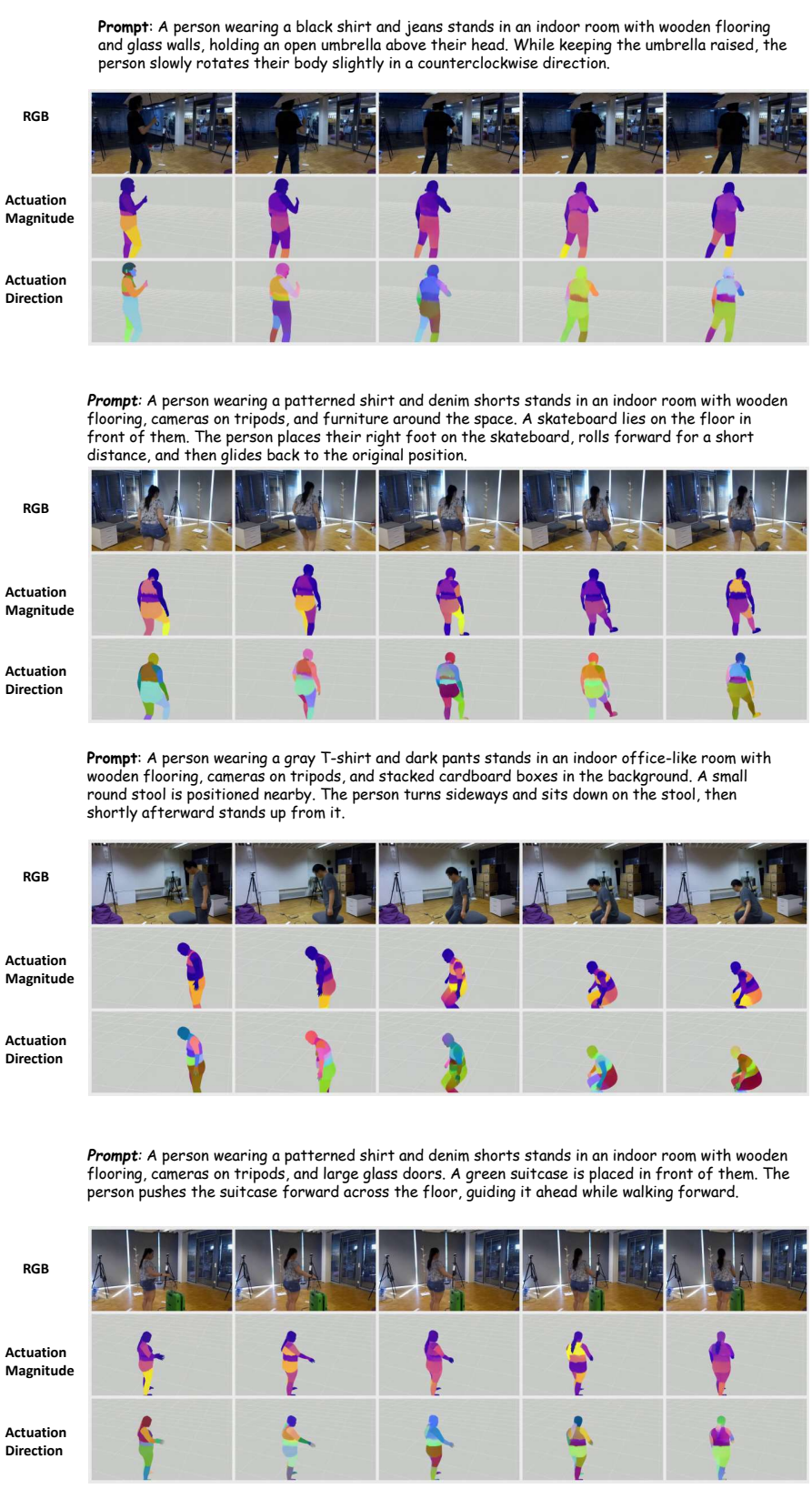


Figure 8: **More Qualitative Results on Real-World Data.** We present additional results on the InterCap dataset (Huang et al., 2022). The predicted actuation maps are well aligned with the input videos, demonstrating that our model can generalize to real-world scenarios.

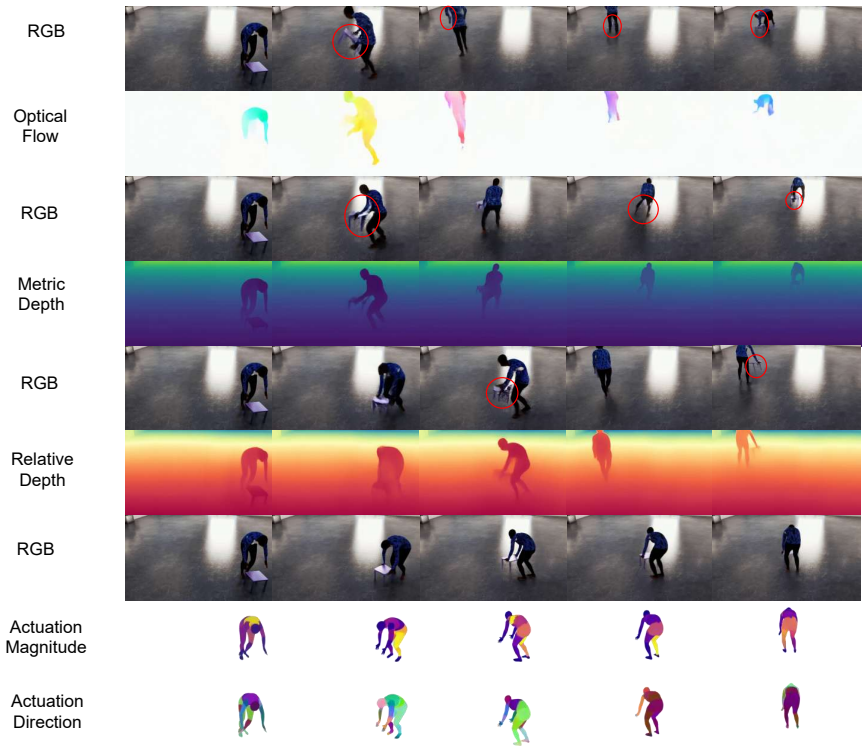


Figure 9: **Modality Ablation.** We compare different auxiliary representations used to guide video generation. The first three rows use geometric signals (optical flow, metric depth, and relative depth), while the last row uses the proposed physics-informed actuation map. Artifacts observed in geometric-signal results are highlighted with red circles, whereas the actuation representation produces more stable and coherent interactions.