COMPADAPT: ADAPTABLE COMPOSITE MOTION MODELING FOR PHYSICS-CONSISTENT TEXT-TO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While diffusion-based text-to-video (T2V) models have demonstrated impressive capability in generating realistic and temporally coherent videos, they often fail to respect fundamental physical dynamics. Although recent physics-constrained methods incorporate explicit dynamics priors to improve physical plausibility, they remain limited to simple single-type motions, depend on manually specified parameters, and struggle to generalize to unseen physical laws. In this work, we propose CompAdapt, a physics-consistent T2V framework for adaptable generation across complex real-world scenarios. It extends neural dynamics modeling beyond single-type motions to encompass composite physical behaviors, including coupled motions, multi-stage transitions, and multi-object collisions. Furthermore, CompAdapt translates natural language prompts into structured physical semantics, enabling end-to-end specification of motion types, temporal relations, and initial physical parameters. To generalize to novel physical environments, CompAdapt introduces dynamics-aware prior matching, achieving one-shot adaptation without retraining the core dynamics module. In addition, a physics-aware latent feature fusion module improves visual fidelity under fast and complex motion. Experiments on physics-focused T2V benchmarks demonstrate that CompAdapt improves physical consistency over both general T2V models and physics-constrained baselines, while preserving high visual quality and adaptability to unseen dynamics.

1 INTRODUCTION

Diffusion-based text-to-video (T2V) models (Ho et al., 2022b;a; Singer et al., 2023; Yang et al., 2025; Bar-Tal et al., 2024; Chen et al., 2023; 2024a; Guo et al., 2023) have recently achieved impressive progress in generating realistic and temporally coherent videos, supporting applications in content creation, simulation, and immersive media. However, visual realism alone does not guarantee physical correctness. Most T2V models are trained to reproduce motion patterns from large-scale video data, where motion is learned mainly through visual correlations rather than explicit physical dynamics (Guo et al., 2025; Chen et al., 2025). As a result, they may generate videos that appear plausible at first glance but violate basic physical rules, such as gravity, force-induced acceleration, collision response, and coherent transitions between different motion components.

To address physical inconsistency in T2V generation, recent works incorporate explicit physical priors to guide the generation process with physical dynamics constraints. NewtonGen (Yuan et al., 2025) is a representative one, which builds upon physics-informed neural ordinary differential equations (ODEs) (Chen et al., 2018; Raissi et al., 2019) and introduces a Neural Newtonian Dynamics (NND) module to model basic Newtonian motion. This design improves the physical plausibility and enables controllable single-object motion synthesis. However, existing physics-constrained T2V methods are still limited in complex real-world scenarios. First, they mainly focus on single-type fundamental motions and lack support for composite motion, where multiple motion components may occur in parallel, sequentially, or be interrupted by discrete dynamic events such as collisions (Gillman et al., 2025; 2026; Zhang et al., 2026). Second, their motion parameters are typically manually specified, which limits their practicality in open-world generation. Third, these methods have limited adaptability to novel physical laws. When encountering out-of-distribution environments, such as lunar gravity or low-friction surfaces, they usually require retraining the core dynamics module

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.

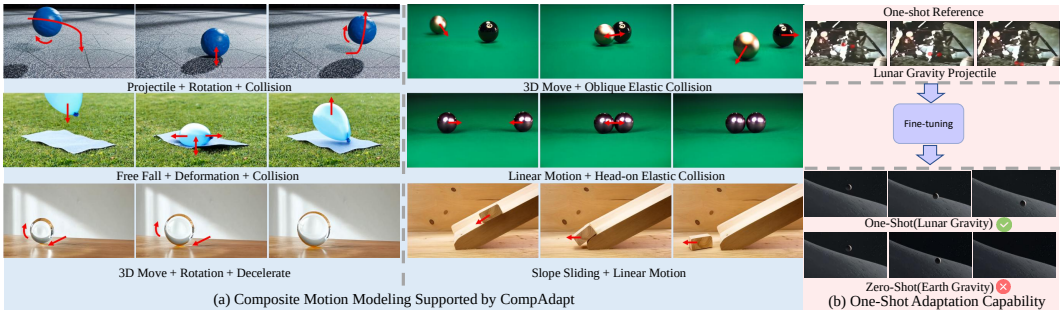

Figure 1: CompAdapt generates physically-consistent composite motion videos from texts, with (a) unified composite motion modeling and (b) efficient one-shot adaptation to novel physical laws.

with large-scale physics-clean synthetic data, lacking efficiency for adaptation (Chen et al., 2018; Rubanova et al., 2019; Greydanus et al., 2019; Cranmer et al., 2020). Therefore, physics-aware T2V generation requires a more flexible framework for handling complex real-world scenarios, which should flexibly compose multiple motion dynamics and rapidly adapt to unseen physical conditions.

Based on these observations, we propose CompAdapt, an adaptable physics-consistent T2V generation framework for complex real-world scenarios, as shown in Figure 1. CompAdapt provides three key capabilities. First, it represents complex real-world motion through parallel and sequential composition, supporting coupled motions, multi-stage transitions, and multi-object collisions beyond single-type fundamental motion. Second, it translates natural language prompts into structured physical semantics, including motion categories, temporal relations, and initial physical parameters. Third, it enables one-shot adaptation to unseen physical laws through dynamics-aware prior matching, allowing efficient adaptation to conditions such as lunar gravity or low-friction surfaces.

Specifically, CompAdapt realizes these capabilities through a language-to-dynamics-to-video pipeline. A text-to-physical parser first converts the input prompt into a standardized motion sequence and initial physical state. Based on an augmented 10-dimensional physical state representation, neural ODE-based dynamics modules predict object-level motion trajectories. The Motion Type Composite Module (MTC) then composes parallel and multi-stage motions into a coherent physical trajectory, while the Multi-Object Interaction Module (MOT) handles discrete multi-object interactions such as collisions. The predicted trajectory is further used to guide a physics-aware video generator with state-dependent latent feature propagation. For novel physical environments, dynamics-aware prior matching selects and adapts the most compatible dynamics module from a single observation.

We evaluate CompAdapt on physics-focused benchmarks covering motion parsing, composite motion generation, multi-object collision, and adaptation to unseen physical laws. Extensive experiments show that CompAdapt consistently improves physical consistency over both general T2V models and physics-constrained baselines. In particular, it better handles composite motions and collision dynamics, and can adapt to novel physical environments from one sample. These results demonstrate the effectiveness of CompAdapt for physics-consistent T2V generation in complex real-world scenarios.

The main contributions are summarized as follows:

- We propose CompAdapt, a physics-consistent T2V generation framework, which integrates prompt-based physical parameter parsing, composite dynamics modeling, and one-shot adaptation to novel physical laws for producing videos in complex physical scenarios.
- We introduce a text-to-physical parsing and composite motion modeling scheme. It converts prompts into structured physical semantics and composes parallel, sequential, and collision motions into coherent physical trajectories.
- We design a dynamics-aware one-shot adaptation strategy for unseen physical laws. By matching a single observation to the most compatible pre-trained dynamics module, CompAdapt can handle novel physical conditions with limited updates.
- We conduct extensive experiments on physics-focused T2V generation benchmarks. The results demonstrate that CompAdapt achieves stronger physical consistency than general T2V models and physics-constrained baselines, while maintaining high visual fidelity.

2 Related Work

2.1 Physics-Aware and Motion-controllable Video Generation

Recent text-to-video models can generate visually impressive videos, but they often violate basic physical laws (Huang et al., 2023; 2024). Existing physics-aware generation methods improve physical consistency through simulation-based pipelines, embedded dynamics priors, or post-hoc optimization (Zhang et al., 2024; Liu et al., 2024a; Xie et al., 2024; Zhang et al., 2026; Guo et al., 2025; Chen et al., 2025). NewtonGen (Yuan et al., 2025) is closely related to our work, as it introduces Neural Newtonian Dynamics for controllable physical motion generation. However, it mainly focuses on fundamental motion types and relies on manually specified physical parameters, limiting its applicability to composite motions, collisions, and unseen physical environments. Motion-controllable video generation provides another important direction for improving generation controllability. Representative methods guide video generation with trajectories, drag signals, camera motion, or motion prompts (Guo et al., 2023; Yin et al., 2023; Geng et al., 2025; Shi et al., 2024; Li et al., 2025; Liao et al., 2025; Ling et al., 2025; He et al., 2024a;b). These methods improve motion alignment but generally do not explicitly model physical dynamics. We build upon Wan-Move (Chu et al., 2025) and introduce physics-aware feature propagation guided by predicted physical states.

2.2 Neural Dynamics Modeling and Physical Evaluation

Neural dynamics models, including PINNs, Neural ODEs, Hamiltonian Neural Networks, and Lagrangian Neural Networks, have been widely studied for modeling physical systems (Raissi et al., 2019; Chen et al., 2018; Rubanova et al., 2019; Greydanus et al., 2019; Cranmer et al., 2020). Relational and graph-based dynamics models further extend this idea to multi-object interaction modeling (Kipf et al., 2018; Sanchez-Gonzalez et al., 2020). Despite their progress, applying neural dynamics to open-ended T2V generation remains difficult due to the need for text-conditioned parameter parsing, composite motion composition, and adaptation to novel physical laws. We address these limitations with text-to-physical parsing, composite dynamics modeling, and one-shot adaptation.

3 Methodology

Given a text prompt that describes a physical scene and its intended motion, our goal is to generate a visually realistic and physically consistent video, which presents four challenges. First, a free-form prompt does not directly provide the physical quantities required by a dynamics model, such as velocity, acceleration, mass, or collision timing. Second, real-world motion often contains multiple components that occur in parallel, unfold sequentially, or interact through discrete events such as collisions. Third, trajectory-guided video generators tend to produce texture distortion under complex motion. This occurs because naively replicating first-frame features along a motion path ignores how deformation and rotation modulate visual appearance across frames. Fourth, pre-trained dynamics modules learned under standard physical conditions no longer match when the environment changes, such as under lunar gravity or on low-friction surfaces, calling for adaptation to novel physical laws. To address these challenges, we propose CompAdapt, as shown in Fig. 2. CompAdapt first converts the input text into structured physical semantics, then models composite physical motion into trajectories, and finally renders a physics-aware video guided by the predicted trajectory. In addition, we introduce a one-shot adaptation paradigm for novel physical laws. We first formulate the physical state representation in Sec. 3.1 and then detail the text-to-physical parser in Sec. Sec. 3.2. Next, we describe composite motion modeling in Sec. 3.3. Finally, we present the physics-aware video generation pipeline in Sec. 3.4 and the one-shot adaptation paradigm in Sec. 3.5.

3.1 Physical State Representation

NewtonGen (Yuan et al., 2025) introduces a physics-informed neural ODE framework for modeling single-object continuous motion, where the physical state is represented as a 9-dimensional latent vector, *i.e.*,

$$\mathbf{Z} = [x, y, v_x, v_y, \theta, \omega, s, l, a].$$

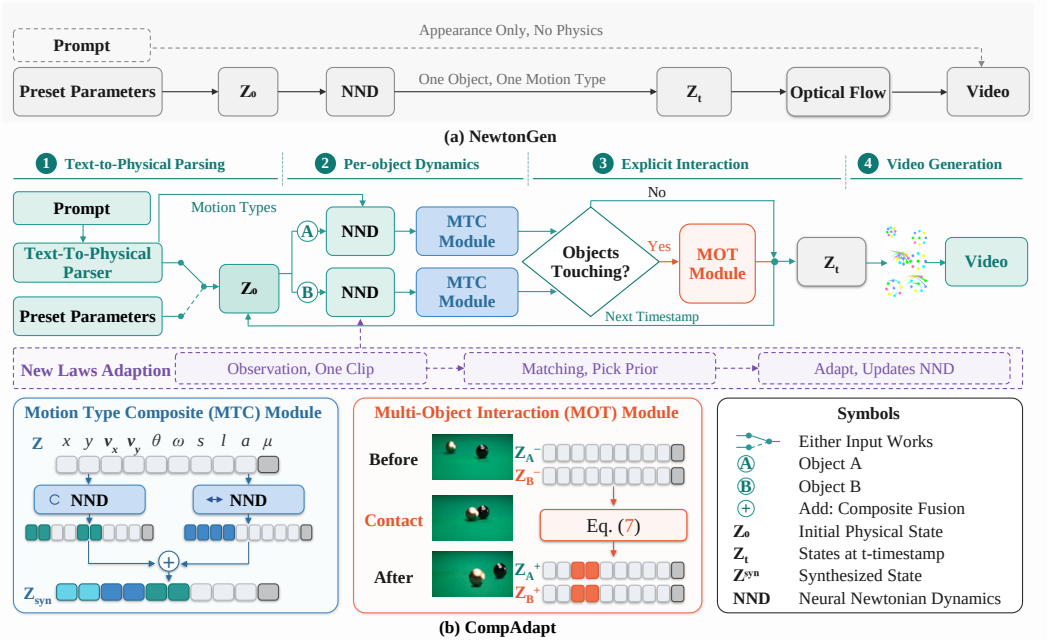

Figure 2: Overview of CompAdapt. (a) NewtonGen uses one NND with preset parameters for a single object and motion type. (b) CompAdapt adds text-to-physical parsing, composite motion modeling (MTC), multi-object interaction (MOT), and one-shot adaptation. The resulting state $Z(t)$ guides video generation.

(x, y) and (v_x, v_y) are the 2D centroid position and velocity, θ and ω the angle and angular velocity, and s, l, and a the shortest side, longest side, and projected area. Because momentum transfer during multi-object collisions depends on mass, a quantity not determined by kinematics or geometry alone, we augment the physical state representation with an explicit mass dimension μ, *i.e.*,

$$\mathbf{Z}' = \{\mathbf{Z}, \mu\}.$$

For notational simplicity, we use $\mathbf{Z}$ to denote $\mathbf{Z}'$ hereafter. Following NewtonGen (Yuan et al., 2025), the temporal dynamics of each element z in $\mathbf{Z}$ follows a second-order ODE that combines linear physics terms with a nonlinear residual, *i.e.*,

$$a_z \ddot{z} + b_z \dot{z} + c_z z + d_z + \mathrm{MLP}(\mathbf{Z}) = 0, \tag{1}$$

where a_z, b_z, c_z, d_z are learnable parameters capturing the dominant Newtonian dynamics, and an MLP models the unknown residuals. Given an initial state $\mathbf{Z}_0$, the future state $\mathbf{Z}_t$ is obtained via a neural ODE integrator (Chen et al., 2018), *i.e.*,

$$\mathbf{Z}_t = \mathbf{Z}_0 + \int_{t_0}^{t} \mathrm{Func}(\mathbf{Z}(\tau))\, d\tau. \tag{2}$$

$\mathrm{Func}(\mathbf{Z}(\tau))$ denotes the set of individual $\frac{\mathrm{d}z}{\mathrm{d}t}$ ODEs governing each element, and $\mathbf{Z}_0 = \mathbf{Z}(t_0)$ is the known initial physical state at time t_0.

3.2 Text-to-Physical Parser

Real-world composite motion is inherently hierarchical, *i.e.*, motions at a timestamp are executed in parallel, while consecutive timestamps unfold sequentially. Translating such motion from a free-form text prompt into a structured physical state further requires two distinct capabilities, *i.e.*, recognizing motion semantics (*e.g.*, category and temporal order) and estimating numerical physical parameters (*e.g.*, velocity and mass). We design a dual-LLM parser based on Qwen2.5-7B-Instruct (Yang et al., 2024), *i.e.*, $\mathrm{LLM}_{\mathrm{seq}}$ and $\mathrm{LLM}_{\mathrm{param}}$, to decouple these objectives. Let P denote the input text prompt. $\mathrm{LLM}_{\mathrm{seq}}$ extracts the motion components and determines their temporal relations, where each motion

is classified into one of 12 predefined categories, *e.g.*, uniform motion, accelerated motion, and collision. It outputs a set of motion sequences $\mathcal{S}$ for N objects, *i.e.*,

$$\mathcal{S} = \text{LLM}_{\text{seq}}(\text{P}) = \{\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \ldots, \mathbf{S}^{(N)}\}, \tag{3}$$

where $\mathbf{S}^{(n)} = (\mathcal{M}_1^{(n)}, \ldots, \mathcal{M}_T^{(n)})$ is the motion sequence for the n-th object, and T is the number of timestamps. $\mathcal{M}_i^{(n)}$ is the synchronous motion components in the i-th timestamp, which can be written as,

$$\mathcal{M}_i^{(n)} = \{m_{i,1}^{(n)}, \ldots, m_{i,o_i}^{(n)}\}. \tag{4}$$

$m_{i,j}^{(n)}$ is the j-th component and o_i is the number of components at timestamp i. $\text{LLM}_{\text{param}}$ takes P and $\mathcal{S}$ as inputs and estimates the physical state $\mathcal{Z}_0$ for the initial timestamp, which can be written as,

$$\mathcal{Z}_0 = \text{LLM}_{\text{param}}(\text{P}, \mathcal{S}) = \{\mathbf{Z}_0^{(1)}, \mathbf{Z}_0^{(2)}, \ldots, \mathbf{Z}_0^{(N)}\}. \tag{5}$$

$\mathbf{Z}_0^{(n)}$ is the physical state for the n-th object.

3.3 Composite Motion Modeling

Given per-object motion sequences $\mathbf{S}^{(n)}$ and their initial physical states $\mathbf{Z}_0^{(n)}$, we instantiate the corresponding executable neural dynamics for each object. Specifically, each motion component $m_{i,j}^{(n)} \in \mathcal{M}_i^{(n)}$ activates its category-specific pre-trained NND module $\mathcal{D}_{\text{cat}(m_{i,j}^{(n)})}$, which takes $\mathbf{Z}_0^{(n)}$ as the initial condition and propagates it into a time-continuous physical state $\mathbf{Z}_{i,j}^{(n)}(t)$. To compose $\mathbf{Z}_{i,j}^{(n)}(t)$ into physically consistent object-level trajectories, we introduce a unified compositional framework in which the Motion Type Composite Module (MTC) fuses concurrent motion components at each timestamp, and the Multi-Object Interaction Module (MOT) resolves discrete multi-object contacts when interactions occur.

Motion Type Composite Module. When multiple motion components act on an object simultaneously, their corresponding NND modules operate on mostly disjoint subsets of $\mathbf{Z}$-space variables, so simultaneous motions such as rotation paired with deceleration naturally superpose at the kinematic level within the physical state representation. Furthermore, since all components share the same initial state $\mathbf{Z}_{\text{init}}$, this composition rule is strictly commutative. This additivity and commutativity motivate an additive composition scheme for parallel motion components. Concretely, at timestamp t_i the object is subject to o_i parallel components $\mathcal{M}_i^{(n)}$. Summing these rollouts would count the force-free motion once per component (*e.g.* the $v_0 t$ drift is duplicated under concurrent accelerations). We therefore count it once and add each component's particular response. Let F_{free} be the force-free field, which keeps the linear and angular velocities constant and the shape fixed (Newton's first law on every block), with rollout $\mathbf{Z}_{\text{free}}(t) = \Phi_t^{F_{\text{free}}}(\mathbf{Z}_{\text{init}})$. The particular response $\mathbf{Z}_{i,j}^{(n)}(t) - \mathbf{Z}_{\text{free}}(t)$ then vanishes at $t = 0$. The composed state is then

$$\mathbf{Z}^{\text{syn}}(t) = \mathbf{Z}_{\text{free}}(t) + \sum_{j=1}^{o_i} \mathbf{r} \odot \big(\mathbf{Z}_{i,j}^{(n)}(t) - \mathbf{Z}_{\text{free}}(t)\big). \tag{6}$$

Here $\odot$ is element-wise multiplication, and $\mathbf{r} = [1, 1, 1, 1, 1, 1, 1, 1, 1, 0]$ zeros the mass coordinate, which does not change under composition. On disjoint blocks Eq. (6) reduces to the displacement sum. On shared coordinates with affine-linear dynamics it gives the exact free-plus-particular superposition. Under the temporal chaining paradigm, $\mathbf{Z}_{\text{init}}$ is set to $\mathbf{Z}_0^{(n)}$ at the first timestamp, and to the synthesized terminal state $\mathbf{Z}^{\text{syn}}(t_{i-1})$ from the preceding timestamp for all subsequent steps. This ensures continuous trajectory concatenation across the full sequence.

Multi-Object Interaction Module. While the MTC handles smooth continuous motion, collisions introduce instantaneous, discontinuous state transitions that fall outside the scope of standard continuous NND evolution. To capture these discrete events, we introduce a collision dynamics module $\mathcal{D}_{\text{col}}$, a separately trained NND that maps pre-collision physical states to post-collision physical states, which can be written as,

$$(\mathbf{Z}_{t_c}^{(A,+)},\ \mathbf{Z}_{t_c}^{(B,+)}) = \mathcal{D}_{\text{col}}(\mathbf{Z}_{t_c}^{(A,-)},\ \mathbf{Z}_{t_c}^{(B,-)}). \tag{7}$$

$\mathbf{Z}_{t_c}^{(A,-)}$ and $\mathbf{Z}_{t_c}^{(B,-)}$ denote the pre-collision physical states of objects A and B at collision time t_c. $\mathbf{Z}_{t_c}^{(A,+)}$ and $\mathbf{Z}_{t_c}^{(B,+)}$ are the post-collision states. The contact time t_c is the first timestamp at which the minimum distance between the two convex object shapes falls below a threshold ϵ (Eq. (22)); otherwise each object continues to evolve independently. After a collision, the two objects continue from $\mathbf{Z}_{t_c}^{(A,+)}$ and $\mathbf{Z}_{t_c}^{(B,+)}$ until the next collision or the end of the motion sequence. Formal derivations of both rules, with exactness conditions, invariant preservation, and a coupling error bound, are given in Sec. C.

3.4 Physics-Aware Video Generation

The composite motion modeling module outputs a full sequence of Z-space physical states with guaranteed physical consistency. As the rendering branch of our decoupled framework, the video generation module translates these abstract kinematic states into realistic video frames, allowing independent optimization of physical accuracy and visual quality.

We build our rendering pipeline on Wan-Move (Chu et al., 2025), a trajectory-conditioned video diffusion model, as per-frame centroid trajectories can be directly projected from physical states. Its core motion guidance mechanism relies on latent feature replication. Given the first-frame latent feature map $\mathbf{F}_0$ and per-frame point coordinates (u_f, v_f) projected from each $\mathbf{Z}_f$, it copies feature values from the initial trajectory position to all subsequent frames in a hard manner, formulated as,

$$\mathbf{F}_f(u_f, v_f, :) = \mathbf{F}_0(u_0, v_0, :). \tag{8}$$

Here $\mathbf{F}_f$ denotes the latent condition feature map at frame f. However, this global hard copying scheme only models translational displacement and leaves rotation- and scaling-induced distortion uncorrected. To mitigate this issue, we inject physical geometric priors into the diffusion feature propagation process. Specifically, we design two complementary modules to implement this idea. One is a spatial warp module dedicated to explicit geometric transformation modeling, and the other is an adaptive blending module for fine-grained artifact compensation.

Spatial Warp Module. Instead of global feature copying, we adopt spatially transformed local patch propagation. For each decoder block, a local latent patch $\mathbf{H}_0$ is extracted from $\mathbf{F}_0$ centered at the initial object position (u_0, v_0). For frame f, this patch is warped according to the transformation matrix, *i.e.*,

$$\mathbf{M}_f = \mathbf{R}(\theta_f - \theta_0) \cdot \mathrm{diag}(\tfrac{s_f}{s_0}, \tfrac{l_f}{l_0}), \tag{9}$$

where $\mathbf{M}_f$ denotes the spatial transformation matrix for frame f, $\mathbf{R}(\cdot)$ is the 2D rotation matrix, and $\mathrm{diag}(\cdot)$ constructs a diagonal scaling matrix. This design explicitly models rotation and scaling during feature propagation.

Adaptive Blending Module. To further compensate for artifacts from latent-space scaling and out-of-plane appearance changes, we introduce a lightweight MLP that takes the relative dimensional change vector $\boldsymbol{\rho}_f = (\frac{s_f}{s_0}, \frac{l_f}{l_0}, \frac{a_f}{a_0})$ as input and predicts a channel-wise gate $\mathbf{G}$ to perform adaptive feature blending. The blending operation is formulated as,

$$\mathbf{h}_f = \mathbf{G} \odot \mathbf{M}_f(\mathbf{H}_0) + (\mathbf{1} - \mathbf{G}) \odot \tilde{\mathbf{h}}_f, \tag{10}$$

where $\mathbf{h}_f$ is the final fused feature for frame f, and $\tilde{\mathbf{h}}_f$ is the native diffusion feature at the current denoising step.

3.5 One-Shot Adaptation to Novel Physical Laws

One-shot adaptation starts from the observed motion. SAM2 (Ravi et al., 2024) tracks the object, and we read the state sequence $\{\hat{\mathbf{Z}}_f\}_{f=0}^{F}$ from its masks and trajectory. Adapting a dynamics module to a novel physical environment must satisfy two core requirements. First, selecting a strong initialization that matches the target dynamics to improve convergence efficiency and adaptation reliability. Second, preserving the learned physical prior during fine-tuning to avoid overfitting to a single observation. To address these requirements, we perform a two-step adaptation based on the extracted observation sequence with the Prior Matching Module and the Optimization Loss.

Prior Matching Module. After that, we perform dynamics-aware prior matching to select the most suitable pre-trained dynamics module. Let $\{\mathcal{D}_k\}_{k=1}^{K}$ denote the set of pre-trained single-object dynamics modules, where each module $\mathcal{D}_k$ is parameterized by ϕ_k^0. Starting from the observed initial state $\hat{\mathbf{Z}}_0$, each module predicts a trajectory, *i.e.*,

$$\tilde{\mathbf{Z}}_f^{(k)} = \mathcal{D}_k(\hat{\mathbf{Z}}_0, f; \phi_k^0), \quad f = 0, \ldots, F. \tag{11}$$

We select the module $\mathcal{D}_{k^*}$ whose predicted trajectory best matches $\{\hat{\mathbf{Z}}_f\}$, which provides a better initialization for one-shot adaptation. It can be written as,

$$k^* = \arg\min_k \frac{1}{F} \sum_{f=0}^{F} \left\| \tilde{\mathbf{Z}}_f^{(k)} - \hat{\mathbf{Z}}_f \right\|_2^2. \tag{12}$$

Optimization Loss. Finally, we fine-tune $\mathcal{D}_{k^*}$ on the extracted trajectory with the proposed optimization loss, which can be written as,

$$\mathcal{L}_{\text{adapt}} = \frac{1}{F} \sum_{f=0}^{F} \left\| \mathcal{D}_{k^*}(\hat{\mathbf{Z}}_0, f; \phi) - \hat{\mathbf{Z}}_f \right\|_2^2 + \lambda \left\| \phi - \phi_{k^*}^0 \right\|_2^2. \tag{13}$$

$\phi_{k^*}^0$ denotes the original pre-trained parameters, and λ controls the regularization strength. The first term fits the observed trajectory, while the second term prevents the adapted model from drifting too far from the learned dynamics prior. CompAdapt employs the updated dynamics module to generate videos under the new physical condition.

4 Experiments

4.1 Experimental Settings

Datasets. We construct four physics-focused datasets (*i.e.*, *Text2Phys*, *CollidePhys*, *CompoPhys*, and *OODPhys*) from physical reasoning benchmarks (*i.e.*, PHYRE (Bakhtin et al., 2019), Physion (Bear et al., 2021), and I-PHYRE (Li et al., 2023)) and large-scale T2V datasets (*i.e.*, Panda-70M (Chen et al., 2024b) and OpenVid-1M (Nan et al., 2025)), following the physics-clean simulation protocol in NewtonGen (Yuan et al., 2025). *Text2Phys* contains 10,000 text prompts with structured physical annotations, which are used to train the text-to-physical parser. *CollidePhys* has 1,000 two-object elastic-collision videos, 800 for training and 200 for testing the collision module. *CompoPhys* has 1,500 composite-motion videos, 1,200 for training and 300 for testing, with per-frame states for parallel, sequential, and hybrid compositions. *OODPhys* contains 34 controlled out-of-distribution scenarios and three real-world videos covering lunar free-fall, low-friction curling, and a damped pendulum. The scenarios are organized into four difficulty tiers by their structural gap to the pre-trained dynamics modules: 22 near-domain (parameter shift), 5 mid-domain (partial structure shift), 5 far-domain (structural OOD), and 2 adversarial (capacity limit). Scenario definitions and the sampling protocol are provided in Sec. E.

Implementation Details. We fine-tune a Qwen2.5-7B-Instruct (Yang et al., 2024) with LoRA as the dual-branch text-to-physical parser. The learning rate, LoRA rank and batch size are set to 2×10^{-5}, 8, and 8 respectively. For the dynamics modules, we use the pre-trained models from NewtonGen (Yuan et al., 2025). The collision-specific module is trained on *CollidePhys* with AdamW optimizer (Loshchilov & Hutter, 2017) for 15,000 training steps. The learning rate and batch size are set to 1×10^{-4} and 32 respectively. The proposed physics-aware video generation are trained on *CollidePhys* and *CompoPhys* with the learning rate of 5×10^{-6}. All experiments are conducted on a single NVIDIA RTX A6000 GPU.

Evaluation Configurations. We measure physical consistency with the Physical Invariance Score (PIS) of NewtonGen (Yuan et al., 2025), the relative stability of a motion invariant, normalized to $[0, 1]$. Visual fidelity uses FID, FVD, PSNR, SSIM, and EPE (Huang et al., 2023; Unterthiner et al., 2018; Huang et al., 2024; Liu et al., 2024b), covering spatial quality, temporal quality, and trajectory error.

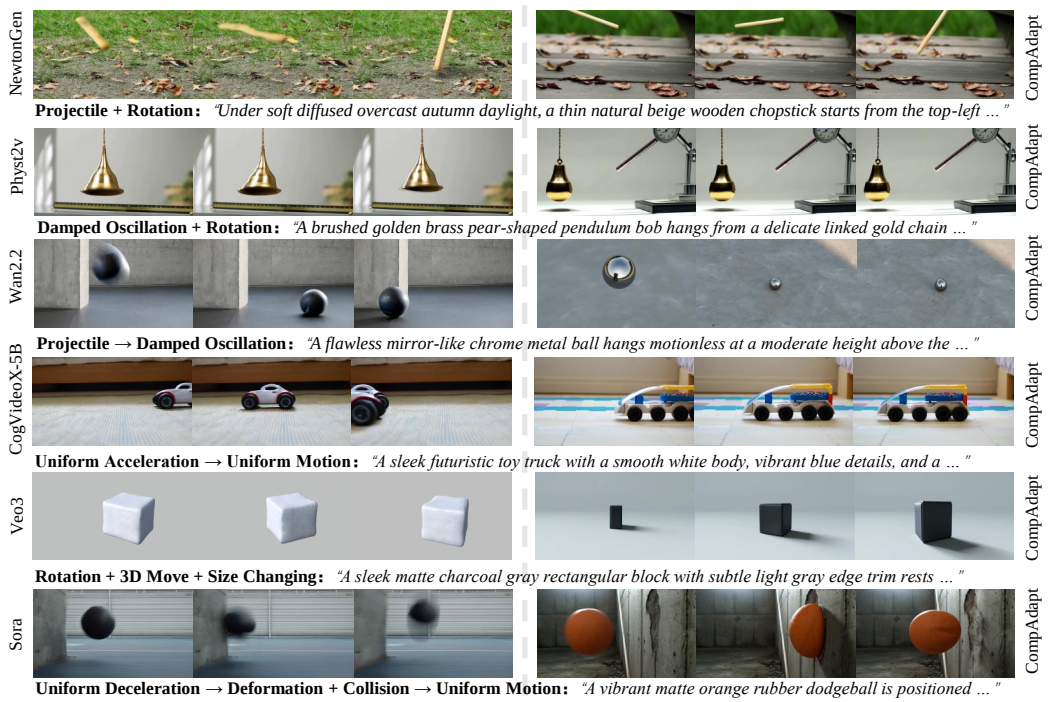

Figure 3: Qualitative comparison results on some representative physical motion scenarios. Our CompAdapt generates videos with accurate physical motion and consistent appearance.

4.2 Comparison with State-of-the-Art Methods

Comparison Configurations. General T2V baselines are Sora (Brooks et al., 2024), Veo3 (Wiedemer et al., 2025), CogVideoX-5B (Yang et al., 2025), Wan2.2 (Wan et al., 2025), and vanilla Wan-Move (Chu et al., 2025). Physics-constrained baselines are retrained NewtonGen (Yuan et al., 2025) and PhysT2V (Xue et al., 2025).

Quantitative Results. We evaluate physical consistency following the standard protocol of NewtonGen (Yuan et al., 2025) on 9 motion categories, covering 7 composite motions and 2 elastic collisions, and report per-parameter PIS in Table 1 (best and second-best highlighted). CompAdapt obtains the highest PIS on every reported parameter. General T2V models score lower because they rely on visual correlations rather than explicit dynamics. Retrained NewtonGen ranks second on composite motions, and the remaining gap reflects explicit composite motion modeling; on elastic collisions it cannot produce valid videos, whereas CompAdapt stays close to the simulation reference.

Qualitative Results. In Fig. 3, general T2V models produce implausible trajectories, and physics-constrained baselines blur textures. CompAdapt keeps both the physical trajectory and the object appearance.

4.3 Ablation Studies

Effect of Composite Motion Modeling. The *w/o MTC* variant runs every component sequentially, the *w/o Temporal* variant runs every component synchronously, and the *w/o Dual LLM* variant uses one LLM for both decomposition and parameter prediction. Table 2 shows that each removal lowers average PIS, so parallel composition, temporal chaining, and the decoupled parser contribute complementary information.

Effect of Physics-Aware Video Generation. Table 3 shows that the spatial warp and adaptive blend improve visual quality over both Go-with-the-Flow (Burgert et al., 2025) and untuned Wan-Move (Chu et al., 2025). Corresponding frames appear in Fig. 4.

Effect of One-shot Adaptation. For each *OODPhys* scenario, we fine-tune the prior-matched module on a single reference (Eq. (13)) and evaluate on the remaining trajectories. Table 4 shows

Table 1: Per-parameter quantitative comparison of physical consistency (PIS). Reference is calculated on simulated ground-truth videos. * denotes retrained version.

Motion Type	PIS↑	Methods							
		Reference	Sora	Veo3	CogVideoX-5B	Wan2.2	PhysT2V	NewtonGen*	Ours
Composite Motions									
Projectile + Synchronous Rotation	v_x	0.9988	0.6548	0.7621	0.5392	0.6395	0.3474	0.8914	**0.9803**
	a_y	0.9487	0.5723	0.6187	0.4230	0.5571	0.3567	0.7802	**0.8189**
	ω	0.9829	0.4267	0.5285	0.3380	0.3425	0.4119	0.8905	**0.9289**
Uniform Acceleration + Synchronous Scaling	a_x	0.8489	0.3437	0.5033	0.4010	0.3077	0.5012	0.6012	**0.6568**
	Δ_r	0.8501	0.2840	0.4167	0.4774	0.1972	0.3410	0.6105	**0.6362**
Circular Motion → Post-collision Linear	ω	0.9933	0.6391	0.7726	0.5392	0.4677	0.6012	0.8807	**0.9788**
	v	0.9972	0.5349	0.6395	0.4774	0.5285	0.4310	0.9784	**0.9830**
Oblique Parabola → Damped Oscillation	v_x	0.9988	0.6370	0.7747	0.5392	0.6187	0.4010	0.8914	**0.9803**
	a_y	0.9402	0.2841	0.3494	0.3083	0.3516	0.4418	0.5107	**0.5240**
Approaching + Rotation Post-collision Rebound	Δ_l	0.7388	0.2911	0.4583	0.3026	0.5013	0.5932	0.6214	**0.6472**
	ω	0.9836	0.4267	0.5285	0.6596	0.3425	0.7842	0.8593	**0.8827**
	v_y	0.9986	0.6510	0.8384	0.6690	0.8481	0.8913	0.9205	**0.9371**
	a_y	0.9487	0.3567	0.5571	0.4230	0.5723	0.7662	0.7802	**0.8189**
Uniform Acceleration + Scaling → Uniform Motion	a_x	0.8489	0.3437	0.5033	0.4010	0.3077	0.5012	0.6012	**0.6568**
	Δ_r	0.8501	0.2840	0.4167	0.4774	0.1972	0.3410	0.6105	**0.6362**
	v	0.9972	0.5349	0.6395	0.4774	0.5285	0.4310	0.9784	**0.9830**
Damped Oscillation + Rotation → Uniform Translation	a_y	0.9402	0.2841	0.3494	0.3083	0.3516	0.4418	0.5107	**0.5240**
	ω	0.9836	0.4267	0.5285	0.6596	0.3425	0.7842	0.8601	**0.8831**
	v	0.9972	0.5349	0.6395	0.4774	0.5285	0.4310	0.9784	**0.9830**
Elastic Collision Motions									
2-object Direct Elastic Collision	v_{x1}	0.9891	0.4812	0.4675	0.4563	0.4704	0.5217	-	**0.9315**
	v_{y1}	0.9876	0.4784	0.4591	0.4485	0.4672	0.5182	-	**0.9287**
	v_{x2}	0.9884	0.4905	0.4703	0.4617	0.4736	0.5279	-	**0.9294**
	v_{y2}	0.9869	0.4847	0.4656	0.4552	0.4691	0.5225	-	**0.9276**
2-object Oblique Elastic Collision	v_{x1}	0.9857	0.4773	0.4602	0.4531	0.4665	0.5154	-	**0.9182**
	v_{y1}	0.9842	0.4715	0.4564	0.4493	0.4627	0.5108	-	**0.9157**
	v_{x2}	0.9851	0.4802	0.4637	0.4575	0.4689	0.5191	-	**0.9169**
	v_{y2}	0.9838	0.4756	0.4598	0.4514	0.4643	0.5146	-	**0.9148**

Table 2: Effect of composite motion modeling.

Methods	PIS ↑	FID ↓	FVD ↓	PSNR ↑	SSIM ↑
w/o MTC	0.615	18.7	102.3	15.1	0.55
w/o Temporal	0.589	19.2	105.7	14.8	0.54
w/o Dual LLM	0.724	16.8	96.2	16.3	0.58
Ours	0.862	15.0	91.4	17.2	0.61

Table 3: Comparison of video generators.

Methods	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	EPE ↓
Go-with-the-Flow	17.5	96.7	14.9	0.56	3.2
Wan-Move	16.3	93.5	16.5	0.58	2.9
Ours	15.0	91.4	17.2	0.61	2.9

high consistency in the near- and mid-domain. Far-domain scenarios define an ambitious target for extending the module library, and per-scenario results with detailed regime analyses are provided in Secs. D and E.

Table 4: One-shot adaptation on *OODPhys* (PIS ↑; mean and std over 5 reference trajectories).

Difficulty Tier	# Scenarios	Avg. Var-PIS	Avg. Std
Near-domain (parameter shift)	22	0.888	0.013
Mid-domain (partial structure shift)	5	0.854	0.015
Far-domain (structural OOD)	5	0.515	0.042
Adversarial (capacity limit)	2	0.730	0.009
Lunar free-fall (NASA footage)	1	0.938	0.012
Curling low-friction sliding (sports broadcast)	1	0.921	0.015
Damped pendulum (physics laboratory)	1	0.907	0.013

5 Conclusion

We presented CompAdapt, which brings structured physical semantics and explicit dynamics into text-to-video generation for composite motion, collisions, and unseen physical laws. Experiments show stronger physical consistency than general T2V and physics-constrained methods while preserving visual fidelity. The planar, multi-object focus admits analytical guarantees.

AI USE STATEMENT

In this work, we did not use generative AI tools for any of the tasks that require disclosure under the ICLR 2027 AI Policy for Authors. The composite motion formulation, the dynamics and collision modules, the one-shot adaptation objective, the benchmark construction, and the interpretation of the experimental results are our own.

Among the tasks for which disclosure is recommended, we used generative AI tools to aid and polish the writing of this manuscript, specifically to edit the text for readability and to reformat the manuscript and its reference list for the ICLR 2027 template. We have reviewed all AI-assisted work: each edit was checked against our intended technical meaning, and no experimental result, claim, or reference was generated or substantively altered by an AI tool. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The model formulation and training objectives are described in Section 3, while the datasets, evaluation protocols, baselines, and implementation settings are documented in Section 4 and the appendix. The appendix additionally reports physical-state encoding details, parser evaluation, a theoretical justification of composite motion modeling, per-scenario one-shot adaptation results, benchmark construction, implementation details, and extended ablations to support reproduction and verification of the reported findings.

REFERENCES

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross B. Girshick. Phyre: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, pp. 5083–5094, 2019.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *ACM SIGGRAPH Asia 2024 Conference Papers*, 2024.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiau-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin A. Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. *CoRR*, abs/2106.08261, 2021.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI technical report, 2024. URL `https://openai.com/index/video-generation-models-as-world-simulators/`.

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*, pp. 13–23, 2025.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *CoRR*, abs/2310.19512, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320, 2024a.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pp. 6572–6583, 2018.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13320–13331, 2024b.

Yongfan Chen, Xiuwen Zhu, and Tianyu Li. A physical coherence benchmark for evaluating video generation models via optical flow-guided frame prediction. *CoRR*, abs/2502.05503, 2025.

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, Yu Liu, Yingya Zhang, and Yujiu Yang. Wan-move: Motion-controllable video generation via latent trajectory guidance. *CoRR*, abs/2512.08765, 2025.

Miles D. Cranmer, Sam Greydanus, Stephan Hoyer, Peter W. Battaglia, David N. Spergel, and Shirley Ho. Lagrangian neural networks. *CoRR*, abs/2003.04630, 2020.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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. *CoRR*, abs/2505.19386, 2025.

Nate Gillman, Yinghua Zhou, Zitian Tang, Evan Luo, Arjan Chakravarthy, Daksh Aggarwal, Michael Freeman, Charles Herrmann, and Chen Sun. Goal force: Teaching video models to accomplish physics-conditioned goals. *CoRR*, abs/2601.05848, 2026.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, pp. 15353–15363, 2019.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. *CoRR*, abs/2505.00337, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *CoRR*, abs/2307.04725, 2023.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *CoRR*, abs/2404.02101, 2024a.

Xuehai He, Shuohang Wang, Jianwei Yang, Xiaoxia Wu, Yiping Wang, Kuan Wang, Zheng Zhan, Olatunji Ruwase, Yelong Shen, and Xin Eric Wang. Mojito: Motion trajectory and intensity control for video generation. *CoRR*, abs/2412.08948, 2024b.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *CoRR*, abs/2210.02303, 2022a.

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, pp. 8633–8646, 2022b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. *CoRR*, abs/2311.17982, 2023.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *CoRR*, abs/2411.13503, 2024.

Thomas N. Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard S. Zemel. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2693–2702, 2018.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. *CoRR*, abs/2503.16421, 2025.

Shiqian Li, Kewen Wu, Chi Zhang, and Yixin Zhu. I-phyre: Interactive physical reasoning. *CoRR*, abs/2312.03009, 2023.

Xinyao Liao, Xianfang Zeng, Liao Wang, Gang Yu, Guosheng Lin, and Chi Zhang. Motionagent: Fine-grained controllable video generation via motion field agent. *CoRR*, abs/2502.03207, 2025.

Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. In *International Conference on Learning Representations*, 2025.

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

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond H. Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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. In *International Conference on Learning Representations*, 2025.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

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. *arXiv preprint arXiv:2408.00714*, 2024.

Yulia Rubanova, Tian Qi Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, pp. 5321–5331, 2019.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 8459–8468, 2020.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. *CoRR*, abs/2401.15977, 2024.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations*, 2023.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025.

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 (CVPR)*, pp. 4389–4398, 2024.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang,

Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

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. *CoRR*, abs/2308.08089, 2023.

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H. Chan. Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. *CoRR*, abs/2509.21309, 2025.

Qiyuan Zhang, Biao Gong, Shuai Tan, Zheng Zhang, Yujun Shen, Xing Zhu, Yuyuan Li, Kelu Yao, Chunhua Shen, and Changqing Zou. Physrvg: Physics-aware unified reinforcement learning for video generative models. *CoRR*, abs/2601.11087, 2026.

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. *CoRR*, abs/2404.13026, 2024.

APPENDIX

The content of the supplementary material involves:

- Physical parameter encoding in the 10D state in Sec. **A**.
- More evaluation of the text-to-physical parser in Sec. **B**.
- A theoretical justification of composite motion modeling in Sec. **C**.
- Extended one-shot adaptation evaluation in Sec. **D**.
- More details of the *OODPhys* benchmark in Sec. **E**.
- The Physical Invariance Score and implementation details in Sec. **F**.
- Generalization to open-ended prompts in Sec. **G**.
- Additional ablations, collision-time detection, and the video-generator comparison in Sec. **H**.
- Full text prompts of the qualitative comparison in Sec. **I**.

A PHYSICAL PARAMETER ENCODING IN THE 10D STATE

The 10D state explicitly represents per-frame kinematics, while material and environmental properties (friction, damping, viscosity, stiffness, *etc*.) are implicitly encoded in the learnable parameters of the NND modules. To verify that such properties are genuinely learned rather than absorbed by trajectory fitting, we probe the parameter encoding at two levels. At the *parameter level*, we train NND modules from scratch under systematically varying values of 8 physical quantities and compute the Pearson correlation between the learned ODE parameters and the ground-truth physical values. At the *trajectory level*, for each quantity we train a module on 40 trajectories under a fixed parameter value and evaluate it on (i) 10 unseen trajectories with the same parameter but different initial conditions and (ii) 10 trajectories under a distinctly different parameter value, reporting the PIS gap to the ground-truth trajectories and the MSE ratio between the incorrect- and correct-parameter settings.

Table 5: Correlation between learned ODE parameters and ground-truth physical values across 8 probed quantities (each probed at 7 values, 50 trajectories per value).

Physical Quantity	Encoding Mechanism	Pearson $\lvert r\rvert$	Encoding
Friction μ_f	6 linear ODE parameters jointly	0.999	Strong
Oscillation frequency f	second-order stiffness term	1.000	Strong
Size expansion rate	size ODE constant term β_s	0.989	Strong
Area conservation under compression	emergent $\alpha_s + \alpha_l \approx 0$	0.980	Strong
Gravitational acceleration g	constant acceleration term c_y	0.914	Strong
Damping coefficient	dedicated damping parameter	0.817	Moderate
Fluid viscosity η	velocity damping term b_y	0.754	Moderate
Spring stiffness k	position- and velocity-dependent terms	0.630	Moderate

Table 6: Trajectory-level generalization of the implicitly encoded parameters. PIS Gap is the difference between the PIS of the correct-parameter module and the ground-truth trajectories on unseen initial conditions; MSE Ratio compares the incorrect-parameter setting against the correct-parameter setting.

Physical Quantity	PIS Gap ↓	MSE Ratio ↑	Tier
Gravitational acceleration g	0.003	174×	Precise encoding
Size expansion rate	0.045	82×	Precise encoding
Area conservation under compression	0.000	54×	Precise encoding
Spring stiffness k	0.022	51×	Solid encoding
Oscillation frequency f	0.002	38×	Solid encoding
Friction μ_f	0.126	6×	Solid encoding
Damping coefficient	0.001	3×	Weak specificity
Fluid viscosity η	0.128	1.6×	Weak specificity

Three interpretable encoding patterns emerge from the probe. First, *direct mapping*, where a one-to-one correspondence exists between a physical quantity and an ODE parameter (*e.g*., the expansion

rate maps to β_s, and the oscillation frequency maps to the stiffness term). Second, *multi-parameter joint encoding*, where a single property is distributed across several ODE components (*e.g.*, friction is jointly represented by 6 linear parameters, consistent with friction affecting resultant acceleration). Third, *automatic constraint discovery*, where the NND satisfies conservation laws without explicit supervision (*e.g.*, area conservation under compression emerges as $\alpha_s + \alpha_l \approx 0$, and the residual MLP self-zeros for linear motions). **Scope and extensions.** As a tractable and well-defined instantiation, the framework is developed on planar dynamics, where the state blocks admit exact analytical treatment, and the collision module realizes two-body contacts through a closed-form state map. This choice deliberately factors out contact-surface mechanics, three-dimensional forces, and articulated bodies, which the compositional state-block design is well positioned to incorporate by adding coordinates—a natural direction for extension. The dominant numerical limits at high parameter values are set by ODE solver precision and frame sampling rather than model capacity, and are likewise relieved by finer integration.

B TEXT-TO-PHYSICAL PARSER EVALUATION

We evaluate the dual-LLM parser in isolation on a held-out test set of 2,000 annotated samples from *Text2Phys*, covering all 12 motion categories and three composite structures (parallel, sequential, and hybrid), with an average of 2.3 motion components per sample. We compare against zero-shot Qwen2.5-7B-Instruct (Yang et al., 2024) and a single-LLM fine-tuned baseline that performs both decomposition and parameter prediction.

Motion decomposition. As shown in Table 7, the dedicated LLM_{seq} branch achieves 99.1% sequence-level accuracy and 99.4% per-motion category accuracy, outperforming the single-LLM baseline by 20.7 percentage points on sequence-level accuracy. For samples with three or more sequential stages, sequence-level accuracy remains 98.3%.

Table 7: Motion decomposition accuracy on the *Text2Phys* test set (2,000 samples).

Method	Sequence-level Acc. ↑	Per-motion Category Acc. ↑
Zero-shot Qwen2.5-7B-Instruct	54.7%	70.2%
Single-LLM fine-tuned baseline	78.4%	86.9%
Ours (dual-LLM parser)	**99.1%**	**99.4%**

Physical parameter prediction. The $\text{LLM}_{\text{param}}$ branch adopts a two-stage design in which the LLM performs semantic-level recognition and a deterministic post-processing program generates the final numerical values. The LLM distinguishes three input scenarios: explicit numerical values (extracted verbatim), qualitative adjectives (*e.g.*, "high speed", "heavy"; mapped to *small/medium/large* labels and sampled from the corresponding physical intervals), and absent descriptions (outputting a *none* token and sampling from a globally reasonable range). Correctness is defined as $\leq 1\%$ relative error for explicit values, exact label match for adjectives, and correct *none* detection for absent descriptions. As shown in Table 8, the parser exceeds 99% accuracy in all three scenarios, providing a reliable foundation for downstream dynamics modeling. Qualitative examples: for "A small steel ball falls freely from a height of 2 meters with an initial horizontal velocity of 3 m/s", the parser outputs `y=2.0, vx=3.0, mass=small`; for "A heavy wooden block slides quickly down a rough slope", it outputs `mass=large, initial_velocity=large`; for "A ball performs projectile motion and then bounces off the wall", it correctly outputs `mass=none, initial_velocity=none`.

Table 8: Physical parameter parsing accuracy on 2,000 mixed test samples.

Method	Overall ↑	Explicit Num. ↑	Adjective Class. ↑	No-Description ↑
Zero-shot Qwen2.5-7B-Instruct	51.8%	70.0%	46.5%	40.5%
Single-LLM fine-tuned baseline	81.7%	88.0%	81.5%	75.5%
Ours	**99.3%**	**99.8%**	**99.2%**	**99.0%**

C Theoretical Justification of Composite Motion Modeling

We justify the additive composition in Eq. (6) and the collision map in Eq. (7) from first principles. Each component module is trained to reproduce the corresponding component dynamics, as quantified by the Physical Invariance Score. We then show that the composition introduces no additional error and preserves every conserved quantity. Write the component dynamics as a smooth vector field $\dot{Z} = F_j(Z)$ with flow $\Phi_t^{F_j}$, and recall that forces add, $\dot{Z} = f = \sum_j F_j$. The state partitions into the translation block $\{x, y, v_x, v_y\}$, rotation block $\{\theta, \omega\}$, shape block $\{s, l, a\}$, and the mass coordinate $\{\mu\}$. Let P_j be the diagonal projector onto the block acted on by component j.

Exactness under block decoupling. Call F_j *block-supported* if it acts only on its own block and depends only on coordinates within it, *i.e.*, $P_{j^\perp} F_j \equiv 0$, $\partial F_j / \partial z_k = 0$ $(k \notin \mathcal{B}_j)$. If the blocks are pairwise disjoint and every F_j is block-supported, then

$$\Phi_t^f(Z_{\text{init}}) = Z_{\text{init}} + \sum_j \mathbf{r} \odot \big(\Phi_t^{F_j}(Z_{\text{init}}) - Z_{\text{init}}\big). \tag{14}$$

On block $\mathcal{B}_m$, all F_j with $j \neq m$ vanish identically, so their flows leave the block at its initial value. The right-hand side of Eq. (14) restricted to $\mathcal{B}_m$ therefore equals $\Phi_t^{F_m}|_{\mathcal{B}_m}$. Since F_m depends only on $\mathcal{B}_m$, this block evolves autonomously and $f|_{\mathcal{B}_m} = F_m|_{\mathcal{B}_m}$, and by uniqueness of ODE solutions it equals $\Phi_t^f|_{\mathcal{B}_m}$. The mask keeps μ fixed.

Linear superposition on shared coordinates. When components share coordinates and the dynamics is affine linear, $\dot{x} = Ax + b$ with $b = \sum_j b_j$, the exact composite is the free evolution counted once plus the sum of the particular responses,

$$x(t) = e^{At}x_0 + \sum_j x_{p,j}(t), \qquad \dot{x}_{p,j} = Ax_{p,j} + b_j, \quad x_{p,j}(0) = 0, \tag{15}$$

whereas the naive displacement sum repeats the free drift once per component (*e.g.* two concurrent constant accelerations with nonzero initial velocity duplicate the $v_0 t$ term). The particular-solution form, which a force-type module can be trained to output, is exact for all linear systems, including linear (Stokes) drag, damped and driven oscillators, and gyroscopic (Coriolis-type) coupling. The same result extends to the linearized dynamics of every system near a stable equilibrium: the quadratic expansion and the normal-mode transform $Ka = \omega^2 Ma$ diagonalize those dynamics into independent harmonic oscillators, covering small-angle pendulums, coupled springs, and elastic deformation modes, while the higher-order nonlinear residual is controlled by Eq. (20).

Preservation of invariants. If a quantity I is conserved by every component, $\nabla I \cdot F_j = 0$, it is conserved by the composite,

$$\frac{\mathrm{d}}{\mathrm{d}t} I = \nabla I \cdot \sum_j F_j = 0; \tag{16}$$

block-supported fields likewise preserve each block's own invariant (momentum of the free block, angular momentum of the torque-free block, oscillator energy, and the area $a = \kappa s l$ with $\dot{a}/a = \dot{s}/s + \dot{l}/l = 0$ under compression). Consequently the composition adds zero marginal PIS degradation and preserves every accurate module exactly.

Collision map and the necessity of mass. Let $\mathbf{n}$ be the contact normal directed from A to B, and $\boldsymbol{\ell}_A, \boldsymbol{\ell}_B$ the vectors from the centers of mass to the contact point, with planar cross product $\boldsymbol{\ell} \times \mathbf{n} = \ell_x n_y - \ell_y n_x$. Under the impulse pair $-J\mathbf{n}, +J\mathbf{n}$ (Newton's third law), the impulse–momentum theorem gives

$$\begin{aligned} \mathbf{v}_A^+ &= \mathbf{v}_A^- - \frac{J}{\mu_A}\mathbf{n}, & \mathbf{v}_B^+ &= \mathbf{v}_B^- + \frac{J}{\mu_B}\mathbf{n}, \\ \omega_A^+ &= \omega_A^- - \frac{J}{I_A}(\boldsymbol{\ell}_A \times \mathbf{n}), & \omega_B^+ &= \omega_B^- + \frac{J}{I_B}(\boldsymbol{\ell}_B \times \mathbf{n}). \end{aligned} \tag{17}$$

The contact-point normal approach speed and the effective inverse mass are

$$u^- = \mathbf{n}\cdot\mathbf{v}_A^- + \omega_A^-(\boldsymbol{\ell}_A \times \mathbf{n}) - \mathbf{n}\cdot\mathbf{v}_B^- - \omega_B^-(\boldsymbol{\ell}_B \times \mathbf{n}), \quad W = \frac{1}{\mu_A} + \frac{1}{\mu_B} + \frac{(\boldsymbol{\ell}_A \times \mathbf{n})^2}{I_A} + \frac{(\boldsymbol{\ell}_B \times \mathbf{n})^2}{I_B}, \tag{18}$$

with $u^+ = u^- - JW$. The restitution law $u^+ = -e\,u^-$ uniquely determines $J = (1+e)u^-/W$, since $W > 0$. The equal-and-opposite impulses act at one point, so total linear and angular momentum are conserved. Summing the kinetic-energy change over both bodies yields

$$T^+ - T^- = -\frac{(1-e^2)(u^-)^2}{2W} \leq 0, \tag{19}$$

which vanishes for the elastic case $e = 1$. At fixed contact geometry the map is affine linear in the pre-impact velocities, so the learned module $\mathcal{D}_{\text{col}}$ can represent it exactly. Because J depends on the masses through W, the mass coordinate is indispensable—identical pre-impact kinematics with different masses yield different post-impact states. The fixed-wall case follows at $\mu_B \to \infty$.

Error bound under coupling. If the fields are L-Lipschitz on the state region traversed and the omitted coupling together with the module error satisfies $\|f - \sum_j F_j\| \leq \eta$ throughout the interval, the Grönwall inequality bounds the composition error by

$$\|\hat{Z}(t) - Z(t)\| \leq \frac{\eta}{L}\big(e^{Lt} - 1\big), \tag{20}$$

which vanishes as the modules become accurate, the coupling is modeled, and the segment length shrinks. Temporal chaining accumulates this bound segment by segment. Additive composition is therefore exact for the quasi-independent families above—rigid-body kinematics, together with linear dissipation and small oscillations when force modules output the particular responses of Eq. (15)—and controlled by Eq. (20) when components are coupled, as quantified for the nonholonomic rolling constraint $v = \omega r$ in Sec. H. A constraint-aware extension follows by introducing Lagrange multipliers, $M\ddot{q} = \mathbf{F} - A^T\lambda$ with $A\ddot{q} = b$.

D Extended One-Shot Adaptation Evaluation

Protocol. Each *OODPhys* scenario contains 10 trajectories (84 frames, $\mathrm{d}t$=0.01 s). For one-shot adaptation, we extract the reference trajectory, perform dynamics-aware prior matching over the 12 pre-trained modules by trajectory MSE (Eq. (13) initialization), fine-tune the matched module on the single reference with the regularized loss in Eq. (13), and evaluate on the remaining 9 trajectories. This is repeated for 5 different reference selections to obtain Var-PIS and the standard deviation. For temporal extrapolation, the module is trained on the first 28 frames (one third) of the reference and evaluated on the full 84-frame sequence.

Prior matching behavior. The prior selected by trajectory MSE frequently differs from the semantically matching module: lunar-gravity projectiles are best matched by the 3D-motion module, damped pendulums by the deformation module, viscous settling by the 3D-motion module, and spring oscillators by the acceleration or damped-oscillation modules (ZS-PIS column of Table 9). This validates dynamics-aware selection over semantic matching: the matched prior alone already provides a strong initialization, and fine-tuning then adapts it to the target environment.

Sample sensitivity. Across the 34 scenarios, 24 have a standard deviation below 0.02 and 16 below 0.01, showing that one-shot adaptation is largely insensitive to the reference sample in near-domain and mid-domain scenarios. The four scenarios with standard deviations above 0.03 (outward and tight spirals, pure rotation, and the combined lunar-plus-low-friction-plus-size-change scenario) are exactly those without a close pre-trained prior or with multi-factor parameter shifts.

Temporal extrapolation. Training on the first third of frames and evaluating on the full sequence yields a PIS drop below 0.02 in 12 of 34 scenarios, indicating that adaptation learns reusable dynamical patterns rather than memorizing the reference trajectory. Moderate drops (0.10–0.16) appear in five scenarios: the two spring oscillators, the two deceleration scenarios, and super gravity, where the early segment carries less information about the full dynamics. In five further scenarios (pure rolling $\times 2$, rotation $\times 2$, and the double pendulum), PIS_{out} exceeds PIS_{in}, *i.e.*, the module is conservative under limited early information and remains stable in the extrapolated window.

Behavior under extreme shifts. At g=980 (100$\times$ gravity), ω=15, or a=25, the matched prior lies far outside its pre-training range. Physical Invariance Score and absolute trajectory error then report complementary aspects: the structural invariant remains moderately stable (0.83 for extreme gravity), while scale-dependent absolute error grows—confirming that PIS diagnoses invariant structure rather

than scale accuracy, and that extreme regimes call for extending the parameter range of the module library.

Real-world validation. We validate the full adaptation pipeline on three public real-world physics videos: a NASA lunar free-fall experiment, a curling broadcast with low-friction sliding, and a damped-pendulum physics laboratory recording. For each video, SAM2 (Ravi et al., 2024) segments and tracks the target object, the state-extraction pipeline (`physical_encoder`) converts the track into a physical state sequence, and one-shot adaptation proceeds exactly as in simulation. Var-PIS is computed with the extracted real trajectory as reference, under natural imaging noise and tracking errors. As shown in Table 4 of the main text, the adapted modules reach Var-PIS above 0.90 in all three cases with standard deviations below 0.016, confirming that the adaptation remains effective beyond controlled simulation.

E Details of the OODPhys Benchmark

Scenario tiers and overview. The expanded *OODPhys* benchmark contains 34 controlled scenarios spanning seven dynamical categories and 3 real-world videos. The categories cover shifted gravity, friction, damping, and thrust, coupled rolling, and structurally novel mechanisms including spring oscillation, spiral motion, and chaotic double pendulums. Scenarios are grouped into four tiers by the structural gap between the target dynamics and the 12 pre-trained dynamics modules, as summarized in Table 4 of the main text. The per-scenario results are listed in Table 9.

Two scenarios deserve explanation. For standard and extreme deceleration, GT-PIS is bounded at 1.000 and 0.600 because the monitored invariant (constant acceleration) is inherently less stable for strongly decelerating motion, and Var-PIS closely tracks GT-PIS, *i.e.*, adaptation reaches the theoretically attainable consistency. For oscillating size, the ground-truth trajectory itself scores 0.106 on the size-change invariant, so the adapted module (0.955) exceeds the ground-truth score on this metric; we retain the scenario to document metric behavior rather than adaptation quality.

Dynamics of the new scenario families. The far- and mid-domain scenarios introduce dynamical mechanisms absent from the pre-trained module library: spring-mass oscillation, $m\ddot{x} + kx = 0$, with solution $x(t) = A\cos(\omega t + \varphi)$ and $\omega = \sqrt{k/m}$, whose invariant is the total spring energy $\frac{1}{2}k(x - x_{\text{eq}})^2 + \frac{1}{2}mv^2$; viscous settling under the three-force balance $m\ddot{y} = (m - \rho_f V)g - 6\pi\eta r\dot{y}$, converging to the terminal velocity $v_t = 2r^2(\rho_o - \rho_f)g/(9\eta)$; non-uniform gravity $g(y) = g_0\,(R/(R+y))^2$, which generalizes the constant-g assumption; pure rolling with the no-slip constraint $v = \omega r$, yielding $a = g(\sin\theta - \mu_r\cos\theta)/(1 + I/(mr^2))$ and $a = \frac{2}{3}g(\sin\theta - \mu_r\cos\theta)$ for a solid cylinder; constrained spiral motion $r(t) = r_0 + \dot{r}t$ with constant angular velocity; and the chaotic double pendulum, a 4-DOF nonlinear coupled system that motivates extending the state representation beyond the current ten dimensions.

Generation protocol. Each scenario contains 10 trajectories of 84 frames simulated at $\mathrm{d}t = 0.01\,\mathrm{s}$ by integrating the target dynamics with randomized initial conditions (initial positions, velocities, and object sizes). The 12 pre-trained single-object modules (uniform motion, acceleration, deceleration, parabolic motion, parabolic motion with rotation, circular motion, rotation, damped oscillation, slope sliding, size changing, deformation, and 3D motion) form the candidate set for prior matching.

F Implementation Details

Physical Invariance Score (PIS). Following NewtonGen (Yuan et al., 2025), the Physical Invariance Score measures the relative stability of a motion-invariant physical quantity q over a trajectory and is normalized to $[0, 1]$, *i.e.*,

$$\mathrm{PIS} = \frac{1}{1 + \mathrm{std}(q)/(|\mathrm{mean}(q)| + \varepsilon)}. \tag{21}$$

Each scenario is evaluated only on the invariant associated with its own physical law (*e.g.*, horizontal velocity and energy for projectiles, pendulum energy for pendulum-like motions, the $v = \omega r$ coupling for pure rolling), so PIS reflects how well the generated trajectory obeys the target law.

Hyperparameters and reproducibility. Table 10 summarizes the key hyperparameters of each trainable component. All experiments are conducted on a single NVIDIA RTX A6000 GPU. All

Table 9: Per-scenario results of one-shot adaptation on the 34 controlled *OODPhys* scenarios. GT-PIS is the PIS of the ground-truth simulation, *i.e.*, the attainable upper bound. ZS-PIS is the PIS of the best matched pre-trained module without fine-tuning (prior matching only). Var-PIS and Std are the mean and standard deviation of the test PIS over 5 reference-sample selections. $\mathrm{PIS}_{\mathrm{in}}$ and $\mathrm{PIS}_{\mathrm{out}}$ are the PIS on the seen first third and the extrapolated remaining window when training on the first third of frames only.

Scenario	Modified physics	GT-PIS	ZS-PIS	Var-PIS	Std	$\mathrm{PIS}_{\mathrm{in}}$	$\mathrm{PIS}_{\mathrm{out}}$
Near-domain: parameter shift (22)							
Lunar gravity projectile	g=1.63	1.000	0.972	0.982	0.005	0.981	0.969
High gravity projectile	g=12	1.000	0.971	0.961	0.018	0.990	0.982
Super gravity projectile	g=30	1.000	0.974	0.891	0.009	0.962	0.815
Low-friction slope	μ_f=0.01	1.000	0.998	0.982	0.015	0.993	0.982
Standard slope	μ_f=0.1	1.000	0.998	0.978	0.011	0.997	0.992
Lunar slope	g=1.63	1.000	0.998	0.913	0.027	0.991	0.977
Low-damping pendulum	d=0.05	0.989	0.947	0.895	0.006	0.961	0.928
Standard-damping pendulum	d=0.8	0.880	0.947	0.896	0.006	0.961	0.929
Overdamped pendulum	d=3.0	0.833	0.947	0.901	0.007	0.961	0.930
Negative-damping pendulum	$d = -0.5$	0.878	0.947	0.894	0.006	0.961	0.928
Lunar pendulum	g=1.63	0.857	0.947	0.936	0.024	0.961	0.931
Standard acceleration	a=4	1.000	0.997	0.942	0.005	0.987	0.964
Extreme acceleration	a=15	1.000	0.997	0.916	0.001	0.989	0.968
Standard deceleration	a=8	1.000	0.997	0.574	0.028	0.937	0.792
Extreme deceleration	a=25	0.600	0.997	0.574	0.028	0.935	0.779
Pure rotation	ω=1.5	1.000	0.805	0.753	0.042	0.796	0.895
Fast rotation	ω=15	1.000	0.823	0.737	0.006	0.780	0.836
Constant size expansion	β_s shift	1.000	0.999	0.969	0.001	0.995	0.989
Oscillating size	periodic β_s	0.106	0.999	0.955	0.006	0.992	0.981
Standard thrust	a=3	1.000	0.990	0.987	0.002	0.990	0.984
Strong thrust	a=8	1.000	0.990	0.987	0.002	0.990	0.984
Combined lunar g + low friction + size change	multi-factor	1.000	0.998	0.904	0.032	0.991	0.977
Mid-domain: partial structure shift (5)							
Viscous settling	η=1.0	0.736	0.959	0.940	0.029	0.981	0.975
Viscous settling, high viscosity	η=5.0	0.621	0.959	0.925	0.021	0.919	0.893
Non-uniform gravity	$g(y)$=$g_0(R/(R+y))^2$	1.000	0.930	0.825	0.017	0.971	0.940
Pure rolling on slope	μ_r=0.01	0.899	0.844	0.794	0.004	0.837	0.903
Pure rolling, high friction	μ_r=0.1	0.909	0.844	0.784	0.002	0.837	0.903
Far-domain: structural OOD (5)							
Spring oscillator	k=10	0.798	0.507	0.483	0.011	0.705	0.584
Stiff spring oscillator	k=50	0.395	0.488	0.515	0.018	0.728	0.600
Outward spiral	$\dot{r}$=0.5	0.562	0.565	0.423	0.082	0.417	0.317
Tight spiral	$\dot{r}$=2.0	0.569	0.700	0.459	0.083	0.369	0.365
Large-amplitude circular motion	large radius	0.776	0.761	0.693	0.017	0.746	0.724
Adversarial: capacity limit (2)							
Chaotic double pendulum	4-DOF coupling	0.647	0.643	0.629	0.000	0.628	0.633
100× extreme gravity	g=980	1.000	0.946	0.830	0.017	0.966	0.942

data generation, adaptation, and evaluation pipelines are deterministic, and every experiment uses fixed random seeds. To ensure reproducibility, we will release the code, the datasets (*Text2Phys*, *CollidePhys*, *CompoPhys*, and the expanded *OODPhys*), the evaluation scripts, the state-extraction pipeline, and the generated videos upon acceptance.

G Generalization to Open-Ended Prompts

To test generalization beyond the controlled benchmark vocabulary, we collect 50 open-ended prompts with diverse object categories, materials, lighting, and scene contexts, with zero overlap with the original test set, and generate 10 videos per prompt. As shown in Table 11, the average PIS is 0.859 versus 0.862 on the *CompoPhys* test set, and FID/FVD remain essentially unchanged at 15.0/91.5 versus 15.0/91.4. Physical guidance therefore preserves both physical consistency and visual quality across diverse visual contexts, consistent with the strictly decoupled design of physical-semantics extraction and visual rendering: the parser is trained to filter out appearance modifiers, and the video backbone renders all visual content. On the same prompts, the parser's decomposition accuracy decreases by at most 0.2% relative to the in-distribution test set.

Table 10: Key hyperparameters of the trainable components.

Component	Setting	Value
Dual-LLM parser	Backbone	Qwen2.5-7B-Instruct + LoRA
	LoRA rank / learning rate / batch size	8 / 2×10^{-5} / 8
	Epochs	3
	Decoding	greedy, $\leq$ 512 tokens
Collision module $\mathcal{D}_{\text{col}}$	Optimizer	AdamW
	Learning rate / schedule	1×10^{-4} / cosine annealing
	Training steps / batch size	15,000 / 32
Physics-aware video generator	Learning rate	5×10^{-6}
	Batch size	1
	Training data	*CollidePhys* + *CompoPhys*
One-shot adaptation	Optimizer	AdamW
	Learning rate / schedule	5×10^{-3} / cosine annealing
	Epochs / trajectory length	100 / 84 frames ($\mathrm{d}t$=0.01 s)

Table 11: Physical consistency and visual quality on 50 open-ended prompts versus the *CompoPhys* test set.

Setting	PIS ↑	FID ↓	FVD ↓
CompoPhys test set	0.862	15.0	91.4
Open-ended prompts (50)	0.859	15.0	91.5

H Additional Ablations

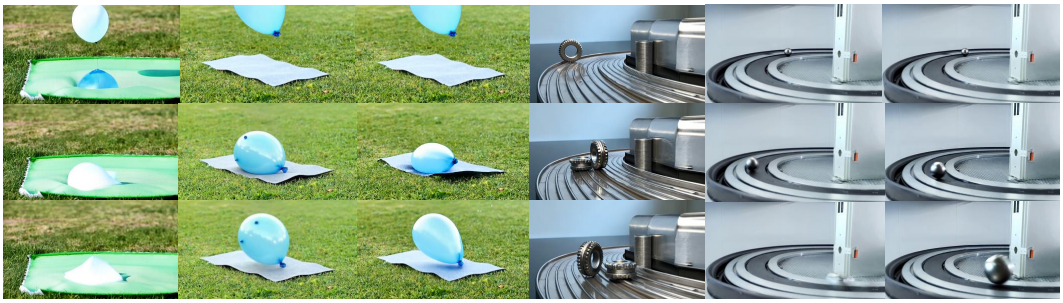

Figure 4: Qualitative comparison of different video generators. Our method maintains consistent visual appearance while ensuring accurate physical motion and sharper texture details.

Collision-time detection. Each object is modeled as a convex shape $\mathcal{K}$ parameterized by its centroid $\mathbf{c}_t = (x_t, y_t)$, orientation θ_t, and dimensions (s, l) extracted from $\mathbf{Z}_t$. The minimum boundary distance used in Sec. 3.3 is

$$d_t = \max_{\|\mathbf{d}\|=1} \left[\mathbf{d} \cdot (\mathbf{c}_t^{(B)} - \mathbf{c}_t^{(A)}) - h^{(A)}(\mathbf{d}) - h^{(B)}(-\mathbf{d}) \right], \tag{22}$$

where $h^{(n)}(\mathbf{d}) = \max_{\mathbf{x}\in\mathcal{K}^{(n)}} \mathbf{d}\cdot\mathbf{x}$ is the support function. A collision is registered at $t_c = t$ when $d_t \leq \epsilon$, and $\mathcal{D}_{\text{col}}$ is applied via Eq. (7). When $d_t > \epsilon$, each object continues under the NND temporal chaining framework.

Mass dimension and the MOT. The mass dimension μ and the collision-aware MOT are both prerequisite for physically correct collision modeling: without mass, the collision solver cannot enforce momentum conservation, and without the solver, colliding objects pass through each other with no interaction. As shown in Table 12, removing the mass dimension (9D state) reduces the average collision PIS from 0.923 to 0.689, and removing the MOT reduces it to 0.487, comparable to general T2V models.

Table 12: Ablations on multi-object elastic collisions (PIS ↑).

Variant	Direct Collision	Oblique Collision	Avg
w/o mass (9D state)	0.694	0.683	0.689
w/o MOT	0.492	0.481	0.487
Ours (10D state + MOT)	**0.929**	**0.916**	**0.923**

Scope of additive composition in the MTC. As shown in Sec. C, additive composition is exact when parallel components act on disjoint subsets of the state space, which covers the dominant composite motions in the benchmark—rotation paired with deceleration, scaling paired with translation. On shared coordinates, the free-evolution-corrected form of Eq. (6) is likewise exact for affine-linear dynamics, counting the force-free evolution once and adding each component's particular response. A residual appears only when the coupling is genuinely nonlinear or constrained. The scheme then supplies a well-defined approximation whose residual is governed by the Grönwall bound (Eq. (20)). Pure rolling illustrates this regime: the no-slip relation $v = \omega r$ links the translational and rotational blocks, so independent rollouts approximate rather than exactly realize the constraint. The residual is small, with adapted Var-PIS 0.784–0.794 against 0.978–0.982 for uncoupled slopes under comparable shifts, and 0.737 for rotation beyond the trained angular-velocity range (ω=15). These measurements map the route forward: quasi-independent components already compose exactly, while constraint-aware extensions (Sec. C) target coupled rolling, dense contact, and articulated chains.

I Full Text Prompts of the Qualitative Comparison

Due to the limited space, each row of Fig. 3 in the main paper only displays a truncated prefix of its text prompt. For completeness and reproducibility, Tab. 13 re-typesets all six prompts in full, listed in the same top-to-bottom order as the rows of Fig. 3. Within each row, all compared methods are driven by exactly the same prompt, and no per-method prompt engineering is applied.

Table 13: Full text prompts used in the qualitative comparison of Fig. 3. The rows follow the top-to-bottom order of the figure, and the truncated prefixes shown in the figure are reproduced here in full.

Motion Type	Full Text Prompt
(1) Projectile + Synchronous Rotation	"Under soft diffused overcast autumn daylight, a thin natural beige wooden chopstick starts from the top-left corner of the frame at a height of 0.9 meters above the deck. It is tossed diagonally downward at a 35-degree angle with a leisurely initial speed, traveling along a smooth parabolic path while rotating at a barely noticeable, glacial clockwise pace, barely turning a right angle throughout its entire descent. Below it lies a rustic weathered gray wooden deck scattered with curled, faded brown and amber fallen leaves, with a hazy backdrop of out-of-focus vibrant green bushes."
(2) Damped Oscillation + Rotation	"A brushed golden brass pear-shaped pendulum bob hangs from a delicate linked gold chain, initially held motionless at its leftmost swing position. It begins a gentle, rhythmic simple pendulum motion, swinging smoothly from side to side, and spins slowly and continuously clockwise as it moves. The scene is set on a pristine white laboratory surface, with a sleek silver precision gauge instrument featuring a circular dial and graduated scale resting to the right."
(3) Projectile $\rightarrow$ Damped Oscillation	"A flawless mirror-like chrome metal ball hangs motionless at a moderate height above the ground. It accelerates downward in a smooth vertical free fall, touches the surface, and then begins a soft, rhythmic damped simple harmonic oscillation."
(4) Uniform Acceleration $\rightarrow$ Uniform Motion	"A sleek futuristic toy truck with a smooth white body, vibrant blue details, and a sunny yellow roof rests motionless on the soft pastel-colored interlocking foam play mat. It begins accelerating steadily to the left, its black wheels spinning faster and faster with gentle momentum, before reaching a constant, smooth cruising speed and moving uniformly across the mat toward the left side of the frame."
(5) Rotation + 3D Move + Size Changing	"A sleek matte charcoal gray rectangular block with subtle light gray edge trim rests motionless at the center of a clean empty space. It glides steadily forward in fluid 3D motion toward the viewer, spins slowly and continuously clockwise around its central vertical axis, and grows larger evenly in all dimensions as it approaches. The scene features a seamless matte light gray floor and background, with soft directional lighting casting a gentle elongated shadow that shifts and scales in sync with the block's movement."
(6) Uniform Deceleration $\rightarrow$ Deformation + Collision $\rightarrow$ Uniform Motion	"A vibrant matte orange rubber dodgeball is positioned in the left half of an empty industrial concrete room, initially traveling steadily straight toward the right wall at a moderate constant speed. It strikes the textured gray concrete wall with a firm impact, squishes noticeably into an oblate shape at the moment of collision, then bounces cleanly off the wall and continues moving straight to the left at an identical steady speed, with motion blur emphasizing its continuous movement."