PHYSICS-ADANET: LEARNING PHYSICAL LAWS VIA ADAPTIVE ARCHITECTURE ALIGNMENT ACROSS DIVERSE SCENARIOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in video generation aim to produce physically realistic content. However, existing methods often fail to maintain physical consistency under Out-of-Distribution (OOD) environmental changes. Motivated by this observation, we study the task of OOD generalization of physical laws in video generation, where a model must infer physically plausible motion in unseen environments by recomposing invariant physical laws from previously observed motion primitives, rather than extrapolating memorized visual patterns or numerical parameters. This setting poses two key challenges: extracting transferable physical structures from high-dimensional visual observations and adapting the alignment mechanism to diverse physical environments. To address these challenges, we propose Physics-AdaNet. Specifically, our approach employs Adaptive Physics Alignment (AdaAlign), which bridges the gap between latent representations and physical priors via temporal masking, fostering structural physical understanding rather than relying on statistical patterns. Meanwhile, to accommodate diverse physical environments, we incorporate Neural Architecture Search (NAS) to adaptively tune the alignment architecture for each scenario, further enhancing generalization capacity. Experiments on multiple benchmark tasks with structural OOD shifts demonstrate that Physics-AdaNet achieves the best kinematic accuracy across all OOD scenes, where the invariant governing law must be recomposed rather than recalled, while remaining competitive with state-of-the-art baselines on In-Distribution (ID) scenes and standard visual-fidelity metrics. More visualization results, frame-level qualitative results on real-world footage, and video demonstrations are provided in the appendix and on our project website at `https://physicsadanet.github.io/Physics-AdaNet/`.

1 INTRODUCTION

Recent advances in video generation (Ho et al., 2020; Peebles & Xie, 2023) have enabled models to produce highly realistic videos of dynamic physical scenarios. However, these models often fail to generalize beyond seen data distributions: while they can plausibly simulate familiar phenomena (e.g., falling or collisions), their physical consistency collapses in Out-of-Distribution (OOD) scenes where environmental conditions shift, e.g., producing inaccurate velocities (Kang et al., 2024).

Motivated by this observation, we investigate the problem of learning transferable physical laws for generative models from the novel perspective of OOD generalization. Building upon prior formulations of the OOD generalization problem (Kang et al., 2024; Lin et al., 2025), which construct OOD scenarios mainly by perturbing scalar physical parameters, we study a complementary and arguably more demanding setting in which the OOD shift is *structural*: the training distribution contains only decoupled motion primitives, and generalizing to the test distribution requires re-composing the invariant governing law rather than extrapolating a numerical parameter, thereby maintaining generalization even as environmental conditions shift.

Existing efforts to incorporate physical knowledge into generative models follow four paradigms: data-driven post-training, LLM/VLM-guided reasoning, physics-guided simulation and alignment-based representation learning, as detailed in Section 2. Data-driven methods merely fit the statistics

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.

of the training distribution, physics-guided methods are bound to expert-defined rules of specific phenomena, and LLM/VLM-guided methods externalize the knowledge to auxiliary large-scale models rather than internalizing it. Alignment-based methods (Zhang et al., 2025b; Wang et al., 2025c) are thus the most pertinent to our task; nevertheless, they maintain a static alignment architecture in varying physical environments, lacking the capacity for dynamic architectural adjustments tailored to scenario-dependent physical priors.

Thus, current approaches struggle with the OOD generalization of physical laws due to their inability to internalize intrinsic physical laws from data and their lack of dynamic adaptability across diverse environments and conditions. Consequently, the core challenges for this task can be summarized as follows: i) How can the invariant law be learned from observations biased toward a specific training environment, thereby enhancing OOD generalization performance during inference? ii) Considering the inherent diversity of physical laws in the real world, how can the model architecture be dynamically adjusted to optimize the learning effect for specific physical environments?

To bridge this gap, we propose Physics-AdaNet (Physics-aware Network with Adaptive Architecture Alignment), a novel framework that injects physical knowledge into generative models through alignment across diverse environments. Specifically, we introduce Adaptive Physics Alignment (AdaAlign), which uses a pre-trained visual encoder to align the generative network's intermediate feature distributions across both spatiotemporal dimensions, encouraging the model to capture invariant physical structures. Crucially, the alignment loss is refined for Image-to-Video (I2V) synthesis, preserving the initial conditional frames while strengthening the model's reasoning capacity upon them. Furthermore, to overcome the rigidity of manually-preset structures and adapt to shifting physical environments and conditions, we integrate a NAS-based Adaptive Projector, utilizing the scaled AdaAlign loss as a guiding score to dynamically identify the optimal projector configuration for each scenario, so that intermediate layers internalize the physical priors more effectively. Experiments on multiple benchmark tasks with structural OOD shifts confirm that our method advances the OOD generalization of physical laws, achieving the best kinematic accuracy in all OOD scenes while maintaining high-fidelity generation under drastic physical shifts.

Our core contributions are summarized as follows:

- We study the novel task of OOD generalization of physical laws in video generation, which requires models to infer kinematic trajectories in unseen scenes based on internalized physical laws, rather than relying on visual memorization during training.
- We propose Physics-AdaNet, a novel framework involving Adaptive Physics Alignment with a NAS-based Adaptive Projector. Our method incorporates prior knowledge into internal representations through alignment and dynamically optimizes the alignment architecture for specific environments.
- Experiments on multiple structural OOD benchmarks validate the effectiveness of Physics-AdaNet, which attains the best OOD kinematic accuracy among all compared methods, the metric that is decisive for evaluating physical-law generalization in OOD scenes, while remaining competitive on In-Distribution (ID) scenes and visual-fidelity metrics. Preliminary qualitative results on real-world footage further suggest its potential transferability beyond the synthetic benchmark.

2 Related Work

2.1 Physically-aware video generation

Despite advances in visual fidelity and temporal consistency, generating physically plausible dynamics remains a formidable challenge for current video models. Existing efforts to inject physical knowledge follow four paradigms. *Data-driven* methods leverage physics-rich or synthetic datasets together with post-training strategies (Ji et al., 2025; Li et al., 2025; Chen et al., 2025; Wang et al., 2025b; Le et al., 2025; Zhao et al., 2025), such as reinforcement learning from human preference (Ji et al., 2025), to implicitly bake physical common sense into model weights. However, these methods essentially fit the statistics of the training distribution rather than distilling the underlying governing laws, and thus degrade once test conditions fall outside the data coverage. *LLM/VLM-guided* methods (Xue et al., 2025; Yang et al., 2025; Tang et al., 2025; Lian et al., 2024; Zhang et al.,

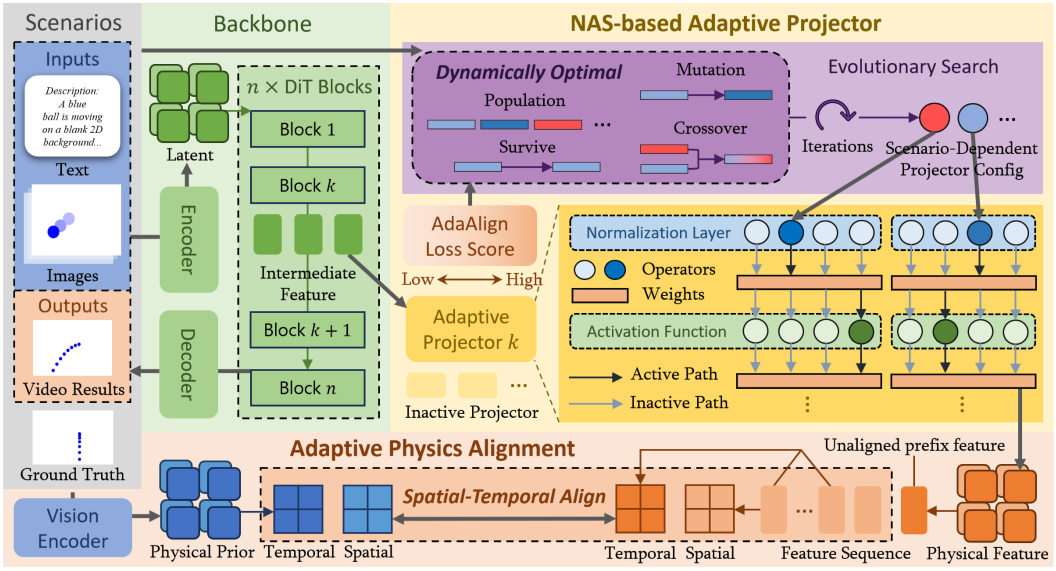


Figure 1: Overview of Physics-AdaNet. Our framework primarily consists of: (a) Adaptive Physics Alignment, which enforces physical consistency by aligning the intermediate features of the backbone with extracted physical priors from ground truth across both spatiotemporal dimensions to enhance OOD generalization; and (b) NAS-based Adaptive Projector, which employs an evolutionary algorithm to identify the optimal projector configuration for each scenario, activating the weights and operators along the corresponding path.

2025a; Motamed et al., 2025; Hsu et al., 2024; Lv et al., 2024; Lin et al., 2025) employ large-scale models as physics reasoners, evaluators, or prompt refiners, yet the knowledge resides in the external model, leaving the generator physically agnostic at inference and bounded by the LLMs' perception-level understanding of physics. *Physics-guided* methods integrate differentiable physics modules (Kandukuri et al., 2022; Link et al., 2022), external simulators (Liu et al., 2024), 3D representations (Wang et al., 2025a; Xie et al., 2025), or neural ODE modules (Yuan et al., 2025; Li et al., 2026) for explicit control, but are confined by the simulator's scope, require expert-defined parameters, and generalize poorly across phenomena. Most related to ours, *alignment-based* methods (Zhang et al., 2025b; Wang et al., 2025c) align the generator's intermediate features with representations from self-supervised video encoders, but target general Text-to-Video (T2V) generation with a fixed alignment architecture, leaving no mechanism to tailor it to scenario-dependent physical priors. In short, existing paradigms memorize, externalize, or only statically internalize physical knowledge, none of which alone suffices for OOD generalization across diverse physical environments, thereby motivating our adaptive alignment framework.

2.2 Neural Architecture Search

Neural Architecture Search (NAS) automates the design of neural networks to achieve optimal performance under constrained computational resources. Early approaches navigate the search space with Reinforcement Learning (Zoph & Le, 2016; Baker et al., 2016) or Evolutionary Algorithms (Real et al., 2017; Xie & Yuille, 2017), demonstrating that automated design can rival hand-crafted architectures, but at the cost of hundreds of GPU days. To improve search efficiency, weight-sharing methods treat the entire search space as a supernet whose parameters are shared between sub-architectures (Pham et al., 2018; Brock et al., 2017; Gong et al., 2019; Pasunuru & Bansal, 2019), with block-wise variants further mitigating the representation shift induced by parameter sharing (Li et al., 2020). More recent advancements (Liu et al., 2018; Luo et al., 2019; Xie et al., 2020), such as DARTS (Liu et al., 2018), have introduced differentiable approaches that relax the discrete search space in a continuous form to accelerate the search process.

3 METHOD

This section first formulates the problem of OOD generalization of physical laws in Section 3.1, and then presents the two components of Physics-AdaNet, namely the Adaptive Physics Alignment in Section 3.2 and the NAS-based Adaptive Projector in Section 3.3, with the overall framework illustrated in Figure 1.

3.1 PROBLEM DEFINITION

In this section, we study the problem of learning transferable physical laws for generative models from the novel perspective of OOD generalization. We specifically explore how a generative model can internalize generalizable physical principles rather than merely fitting superficial statistical correlations. This distinction is critical: while the governing physical laws themselves are invariant, their visual manifestations in video data are highly sensitive to environmental conditions, necessitating a model that can generalize beyond the observed training manifold.

Based on this, we formally define the problem of the OOD generalization of physical laws, which is shown in Figure 2 (a). Let $f_\theta(y \mid x)$ denote a conditional generative model, and let $P(x, y)$ be the joint distribution of the physical conditions x and the video sequences y used during training. The optimal parameters θ_P^* of model f_θ are obtained by minimizing the expected risk:

$$\theta_P^* = \arg\min_\theta \ \mathbb{E}_{x,y\sim P}\left[\mathcal{L}\left(f_\theta(y \mid x), y\right)\right] \tag{1}$$

Subsequently, we define the structural OOD scene of physical laws:

Definition 3.1 (Structural OOD scene of physical laws)**.** *Let $P(x, y)$ and $Q(x, y)$ be the training and target distributions with $Q(x, y) \neq P(x, y)$. The pair (P, Q) constitutes a structural OOD scene of physical laws if and only if:*

$$P(F(x, y) = 0) = Q(F(x, y) = 0) = 1, \tag{2}$$

$$\exists\, i:\ \mathrm{Var}_{x\sim P}[x_i] = 0 \ \wedge\ \mathrm{Var}_{x\sim Q}[x_i] > 0, \tag{3}$$

$$\mathcal{R}_Q\left(\theta_P^*\right) > \min_\theta \mathcal{R}_Q(\theta), \tag{4}$$

where $F(x, y) = 0$ is the constraint induced by the governing law (Eqs. (5)–(8)), x_i denotes the i-th component of the physical condition x, $\mathcal{R}_Q(\theta) = \mathbb{E}_{x,y\sim Q}\left[\mathcal{L}\left(f_\theta(y \mid x), y\right)\right]$ and θ_P^* is defined in Eq. (1).

In the OOD generalization of physical laws, distribution shifts occur when physical conditions (e.g., gravitational acceleration, initial velocity, or object mass) vary between training and inference: the joint distribution changes from P to Q, while the underlying mechanism F remains constant and serves as the fundamental basis for generalization. In particular, Eq. (3) requires the shift to activate a condition variable frozen during training (e.g., $v_{0,x}$ in Task 1 and v_b in Task 2), so that test dynamics must be re-composed from functional primitives observed only in decoupled form under P, rather than extrapolated along an already-varying parameter. The core challenge therefore lies in learning the invariant law F from environment-specific observations $P(x, y)$.

3.1.1 TASK 1: NEWTON'S LAW OF UNIVERSAL GRAVITATION

We evaluate the OOD generalization of physical laws across two distinct scenarios, the first of which focuses on Newton's law of universal gravitation under the near-surface approximation. Specifically, we set a constant gravitational acceleration g in all video samples while initializing a ball with stochastic radius r, random color c, and varying initial velocity v_0 at random coordinates (x_0, y_0). For the training set, the velocity v_0 is restricted to the vertical axis, resulting in a vertical motion, which can be formulated as:

$$y(t) = y_0 + v_{0,y}t + \frac{1}{2}gt^2, \quad x(t) = x_0 \tag{5}$$

In contrast, the test set introduces an OOD challenge by assigning initial velocities v_0 in arbitrary directions. This shifts the dynamics to unconstrained projectile motion, defined by:

$$y(t) = y_0 + v_{0,y}t + \frac{1}{2}gt^2, \quad x(t) = x_0 + v_{0,x}t \tag{6}$$

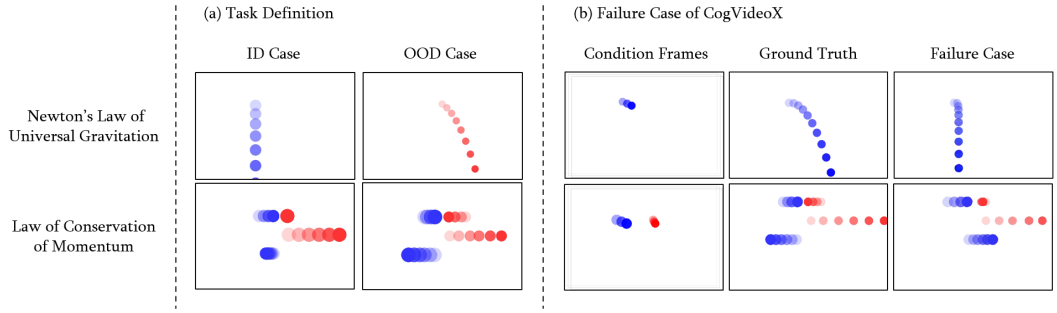


Figure 2: Overview of the benchmark tasks and representative failure cases. (a) Task definition: ID and OOD scenes of the universal gravitation task (top; vertical motion → projectile motion) and the conservation of momentum task (bottom; collision with a stationary vs. moving target). Momentum trajectories are segmented at the collision point, and the right-hand ball's post-collision trajectory is vertically translated for visualization only. (b) Failure cases of CogVideoX under the OOD settings defined in (a): it fails to sustain horizontal velocity in the gravitation task and fails the velocity exchange in the momentum task. Condition frames are offset diagonally for visualization only.

Although both scenes are governed by Newton's law of universal gravitation, their visual manifestations diverge significantly due to the introduction of the horizontal displacement term $\Delta x = v_{0,x}t$. Mathematically, while the model is exposed only to vertical trajectories, the ID governing equations already require the internalization of the linear temporal component $v_{0,y}t$. The two terms $v_{0,y}t$ and $v_{0,x}t$ exhibit an identical functional form, yet the OOD task requires the model to generalize this linear relationship from the y-direction to the x-direction. If a model merely fits the statistical correlations of the training manifold, it will inevitably struggle to project the restricted vertical dynamics onto the broader complexity of arbitrary projectile paths.

3.1.2 Task 2: Law of Conservation of Momentum

Furthermore, we design another task for the law of conservation of momentum. Specifically, two balls a and b are initialized with stochastic radii r_a and r_b, distinct colors c_a and c_b, and random horizontal positions satisfying $x_a + r_a \leq x_b - r_b$ along a shared vertical coordinate y. For the training set, ball b remains stationary while ball a is assigned a horizontal rightward velocity v_a. Consequently, after a duration of $t = (x_b - x_a - r_a - r_b)/v_a$, the two balls undergo a perfectly elastic collision. The velocities after collision are determined by the laws of conservation of momentum and kinetic energy, formulated as follows:

$$v_a' = \frac{m_a - m_b}{m_a + m_b} v_a, \quad v_b' = \frac{2m_a}{m_a + m_b} v_a \tag{7}$$

Following the impact, both balls proceed with constant linear motion at these updated velocities. In contrast, the test set introduces a more complex OOD configuration where ball b is assigned a horizontal leftward velocity $v_b < 0$. The collision occurs at $t = (x_b - x_a - r_a - r_b)/(v_a - v_b)$, following which the velocities are updated according to:

$$v_a' = \frac{(m_a - m_b)\, v_a + 2m_b v_b}{m_a + m_b}, \quad v_b' = \frac{(m_b - m_a)\, v_b + 2m_a v_a}{m_a + m_b} \tag{8}$$

Mathematically, the OOD scene incorporates two additional components: $\Delta v_a' = \frac{2m_b v_b}{m_a + m_b}$ and $\Delta v_b' = \frac{(m_b - m_a) v_b}{m_a + m_b}$. These formulations are functionally identical to the velocity transformation rules encountered in the ID training set. However, the OOD task requires the model to generalize the physical mapping originally learned for ball a acting on a stationary ball b to a reciprocal scene where ball b also possesses initial velocity. A model that merely memorizes visual patterns rather than understanding collision mechanics would fail to generalize from the simplified stationary-target case to these intricate dual-motion interactions.

3.1.3 DESIGN RATIONALE OF THE BENCHMARK

Unlike prior protocols whose OOD scenes are constructed by mildly perturbing scalar physical parameters (Kang et al., 2024), our OOD settings introduce structural shifts in the sense of Eq. (3), forcing a rearrangement of learned physical laws rather than a simple numerical extrapolation. Strictly isolating projectile motion and dual-motion collisions from training is therefore an intentional design choice to prevent shortcut learning: it evaluates whether the model generalizes to unseen trajectories from the decoupled physical laws it has implicitly learned during training. Figure 2 (b) illustrates representative failure cases under this protocol: CogVideoX fails to maintain horizontal velocity in the universal gravitation scenario and struggles with velocity exchange in the conservation of momentum scenario, suggesting the reliance on superficial statistical memorization rather than the underlying physical law. These observations highlight the two aforementioned challenges. To bridge this gap, we propose Physics-AdaNet, which systematically integrates physical alignment with architectural plasticity.

3.2 ADAPTIVE PHYSICS ALIGNMENT

To address the first challenge, we propose Adaptive Physics Alignment (AdaAlign), which injects transferable physical laws into the generative process by aligning the model's latent representations with structured physical priors. The alignment objective, termed Token Relation Distillation (TRD) loss $\mathcal{L}_{\text{TRD}}$, is decoupled into spatial and temporal components as $\mathcal{L}_{\text{TRD}}(s,t) = \mathcal{L}_{\text{Spatial}}(s,t) + \mathcal{L}_{\text{Temporal}}(s,t)$, so that both static object properties and dynamic motion patterns are captured:

$$\begin{aligned}\mathcal{L}_{\text{Spatial}}(s,t) &= \frac{1}{l(hw)^2}\sum_{d=s}^{t-1}\sum_{i,j=1}^{hw}\left\|h_{\text{spatial}}^{d,i,j} - y_{\text{spatial}}^{d,i,j}\right\|_2 \\ \mathcal{L}_{\text{Temporal}}(s,t) &= \frac{1}{l(l-1)(hw)^2}\sum_{d,e=s,d\neq e}^{t-1}\sum_{i,j=1}^{hw}\left\|h_{\text{temp}}^{d,e,i,j} - y_{\text{temp}}^{d,e,i,j}\right\|_2\end{aligned} \tag{9}$$

where $f \times h \times w$ is the token count over f temporal and $h \times w$ spatial positions, $0 \le s < t \le f$ denote the start and end temporal indices, and $l = t - s$ is the length of the temporal window. Specifically, y_{spatial} and y_{temp} represent the pairwise similarity matrices that characterize the relational structure of the feature space. These matrices are computed as follows:

$$y_{\text{spatial}}^{d,i,j} = \frac{y^{d,i}\cdot y^{d,j}}{\|y^{d,i}\|\,\|y^{d,j}\|}, \quad y_{\text{temp}}^{d,e,i,j} = \frac{y^{d,i}\cdot y^{e,j}}{\|y^{d,i}\|\,\|y^{e,j}\|} \tag{10}$$

where $y^{d,i}$ denotes the feature of the i-th patch in the d-th latent frame, extracted from a pre-trained visual encoder that provides essential physical priors. The alignment targets for the diffusion model, h_{spatial} and h_{temp}, are derived from the model's intermediate latent features using an identical computational procedure. By minimizing the discrepancy between these relational matrices, the generative model is guided to internalize the structural and temporal dependencies inherent in the physical priors. Consequently, the learned representations are transferred more reliably to an unseen distribution $Q(x,y)$ during inference, improving the physical consistency of the videos synthesized even under structural distribution shifts.

Tailored to the specific requirements of Image-to-Video (I2V) physical reasoning, we further introduce a temporal masking mechanism. This mechanism is motivated by the observation that initial frames typically serve as fixed control conditions, which do not require alignment with evolving physical laws. Thus, we selectively exempt these prefix frames from the alignment objective to prevent the interference of static spatial constraints with dynamic motion learning. The total training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda\mathcal{L}_{\text{TRD}}(k:f) \tag{11}$$

where $\mathcal{L}_{\text{diff}}$ is the standard diffusion loss, λ is a balancing hyperparameter, and k denotes the number of masked prefix frames, set to 1 in our experiments to account for the initial image prompt. Masking the conditioning frames preserves the initial states while compelling the generative backbone to model the predictive dynamics governed by the underlying physical laws, which is critical for generalization in OOD scenes.

3.3 NAS-BASED ADAPTIVE PROJECTOR

To address the second challenge, we introduce the NAS-based Adaptive Projector, which enables the alignment architecture to adapt across diverse physical environments and to internalize the invariant governing law F from high-dimensional visual data more effectively. Neural Architecture Search (NAS) aims to automate the design of neural networks to achieve optimal performance under constrained computational resources. Formally, we define the search for an optimal projector as a bi-level optimization problem:

$$a^* = \arg\min_{a \in \mathcal{S}} \mathcal{L}_{\text{val}}\left(w^*(a), a\right), \quad \text{where } w^*(a) = \arg\min_{w \in \mathcal{W}(a)} \mathcal{L}_{\text{train}}(w, a) \tag{12}$$

where $\mathcal{S}$ denotes the predefined search space, a represents a candidate architecture, and $w^*(a)$ indicates the weights that minimize the loss of training $\mathcal{L}_{\text{train}}$ in the parameter space $\mathcal{W}(a)$ given the fixed architecture a. The goal is to find an architecture a^* that minimizes the validation loss $\mathcal{L}_{\text{val}}$. Specifically, we search for the feature projectors that align the intermediate layers of the generative network with the features of the pre-trained visual encoder. The search space comprises: i) the layer index a from which intermediate features z_a are extracted; ii) the normalization method n; iii) the projector depth c, together with the weighted module type $type_i$ and the activation function act_i of each layer; and iv) a characteristic dimensionality dim shared by all modules to ensure architectural compatibility. The full parameter ranges and operator categories are shown in Table 3 in Appendix E.

However, exhaustively evaluating every candidate in the search space $\mathcal{S}$ is computationally prohibitive, as $|\mathcal{S}|$ grows combinatorially with the depth of the projector c. We therefore adopt a weight-sharing supernet that instantiates all candidate alignment modules within a single backbone denoising network, with identical modules on different paths sharing parameters. This reduces the number of modules to be trained from exponential to polynomial in c, and the bi-level optimization problem can accordingly be decoupled into two separate phases:

$$a^* = \arg\min_{a \in \mathcal{S}} \mathcal{L}_{\text{val}}\left(w^*(a), a\right), \quad \text{where } w^*(a) = \arg\min_{w \in \mathcal{W}_{\text{super}}} \mathbb{E}_{a \sim \Gamma(\mathcal{S})}\left[\mathcal{L}_{\text{train}}(w, a)\right] \tag{13}$$

where $a \sim \Gamma(\mathcal{S})$ is the prior architecture distribution for the search space $\mathcal{S}$, and $\mathcal{W}_{\text{super}}$ denotes the trainable weights of the supernet, which contains all the architectures in the search space $\mathcal{S}$ and shares the set of parameters with them. Specifically, during each training step, we sample the architecture parameters and operators (e.g., normalization method n and activation functions act_i) according to the stochastic distribution Γ, which is set to a uniform distribution in our experiments. This process activates a single path within the supernet corresponding to a specific candidate architecture for that iteration.

Once the supernet converges, the relative ranking of sub-architectures within the supernet serves as a reliable proxy for their standalone performance. We therefore adopt the AdaAlign loss $\mathcal{L}_{\text{TRD}}$, computed during the alignment process, as the criterion for selecting the sub-architecture for full-scale training. However, raw alignment losses are not directly comparable across depths since feature representations at different layers vary in scale and contribute differently to the final physical consistency. We thus scale the raw loss according to the relative depth of the candidate alignment layer within the backbone:

$$\mathcal{L}_{\text{score}} = \left(\alpha\left(\frac{a_{\text{total}} - a_{\min}}{a_{\text{total}} - a_{\text{cur}}} - 1\right) + 1\right)\mathcal{L}_{\text{TRD}} \tag{14}$$

where a_{total} denotes the total number of transformer blocks in the generative backbone ($a_{\text{total}} = 30$ for CogVideoX-5B), a_{cur} is the index of the candidate alignment layer, $a_{\min} = 12$ is the minimum index of the layer in the search space, and α is a balancing hyperparameter. This scaling normalizes the alignment signal across depths, and its effect on the selected layers is analyzed in Appendix E.

Finally, we search the space with an evolutionary strategy. Starting from a random population, the candidates are iteratively refined through selection, crossover, and mutation guided by $\mathcal{L}_{\text{score}}$, which identifies near-optimal projector architectures for each physical scenario without the prohibitive cost of full enumeration. We do not couple the global alignment weight λ with the evolutionary search, as this would expand the search space exponentially and render joint optimization intractable. Instead, λ is fixed beforehand through small-scale preliminary experiments in Appendix A.

Table 1: Quantitative results across two tasks, grouped by ID and OOD scenes. ↓ and ↑ denote the direction of better performance. Bold indicates the best performance, while underline represents the second-best result. All metrics report mean ± standard deviation over 3 independent runs.

Task 1: Newton's Law of Universal Gravitation						
Type	Method	$v_{x,\mathrm{err}}$ ↓	FVD↓	SSIM↑	PSNR↑	LPIPS↓
ID	CogVideoX	0.1961±0.0046	**18.8121**±2.3761	0.9908±0.0002	32.7264±0.0700	**0.0334**±0.0003
	PISA	**0.1349**±0.0048	27.3103±0.8051	**0.9916**±0.0001	32.7287±0.1401	0.0780±0.0014
	VideoREPA	0.1722±0.0041	39.8954±0.9638	0.9907±0.0002	33.4267±0.1494	0.0858±0.0004
	Physics-AdaNet	0.1854±0.0053	28.2793±1.4998	0.9915±0.0001	**34.3286**±0.1303	0.0613±0.0008
OOD	CogVideoX	7.6781±0.1778	18.3885±1.6394	0.9880±0.0004	29.4198±0.2541	0.0510±0.0013
	PISA	7.5062±0.1868	30.7405±1.3926	0.9874±0.0002	28.1841±0.3772	0.0994±0.0010
	VideoREPA	7.2038±0.1526	28.7635±1.5872	0.9871±0.0004	29.1744±0.1611	0.0704±0.0009
	Physics-AdaNet	**4.3210**±0.1355	**17.2379**±1.4154	**0.9902**±0.0004	**33.3071**±0.2001	**0.0494**±0.0015

Task 2: Law of Conservation of Momentum							
Type	Method	$v_{b,\mathrm{pre},\mathrm{err}}$ ↓	$v_{a,\mathrm{post},\mathrm{err}}$ ↓	FVD↓	SSIM↑	PSNR↑	LPIPS↓
ID	CogVideoX	2.4649±0.0689	3.0130±0.0907	18.3242±1.5775	0.9837±0.0002	27.4478±0.3309	0.0653±0.0012
	PISA	1.9925±0.0740	2.4819±0.0704	16.0173±1.1721	**0.9866**±0.0003	**29.9170**±0.3333	0.0634±0.0013
	VideoREPA	**1.8750**±0.0587	**1.9204**±0.0369	22.2317±1.1689	0.9852±0.0005	28.5860±0.3386	0.0791±0.0009
	Physics-AdaNet	2.1603±0.0579	1.9862±0.0692	**14.3536**±0.6371	0.9842±0.0004	28.4910±0.1902	**0.0463**±0.0009
OOD	CogVideoX	6.8255±0.1975	9.9163±0.2326	30.9228±1.0505	0.9752±0.0010	22.7242±0.1339	0.0943±0.0028
	PISA	6.5456±0.1489	9.1287±0.2431	35.1895±2.3744	0.9742±0.0004	22.6834±0.2679	0.0970±0.0012
	VideoREPA	7.1548±0.1646	10.5501±0.2669	**25.3459**±2.5363	**0.9754**±0.0006	**23.1745**±0.1294	**0.0842**±0.0018
	Physics-AdaNet	**5.1170**±0.1593	**7.5075**±0.2703	29.8604±2.3684	0.9753±0.0007	23.0088±0.1913	0.0888±0.0023

4 EXPERIMENTS

We evaluate Physics-AdaNet on both the ID and OOD test sets of the two tasks. As traditional generative metrics often fail to capture structural physical consistency, our evaluation prioritizes kinematic accuracy over low-level pixel statistics. We measure the ℓ_2 error between velocity sequences extracted from the generated videos and their theoretical ground truth. The complete definitions are given in Appendix B. We also report FVD (Unterthiner et al., 2018), SSIM (Wang et al., 2004), PSNR, and LPIPS (Zhang et al., 2018) as standard visual-fidelity metrics. Sections 4.1 and 4.2 present quantitative and qualitative comparisons against state-of-the-art baselines, and Section 4.3 presents ablation studies; the detailed experimental setup is deferred to Appendix A.

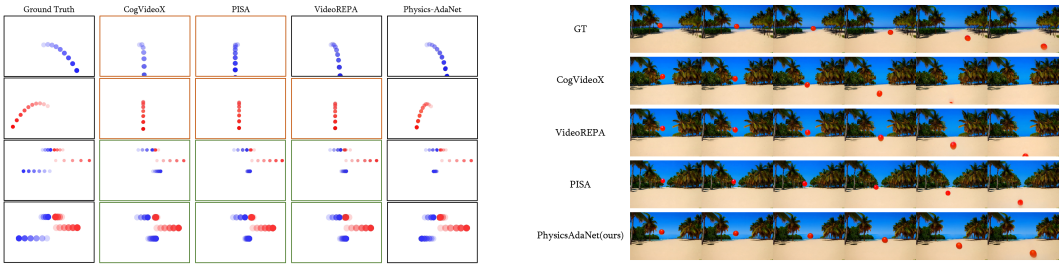
4.1 QUANTITATIVE RESULTS

We compare our method against three baselines across different paradigms: (i) CogVideoX (Yang et al., 2024), directly fine-tuned on our task-specific datasets; (ii) PISA (Li et al., 2025), a data-driven method originally designed for 3D physical video generation, which we adapt to our 2D benchmark by fine-tuning it on CogVideoX with its optical-flow and semantic-segmentation supervision (depth maps are omitted as they are unavailable in 2D); and (iii) VideoREPA (Zhang et al., 2025b), the most relevant alignment-based method, representing static representational alignment strategies.

The quantitative results are summarized in Table 1. In the OOD setting, Physics-AdaNet outperforms all baselines in every kinematic metric, with margins exceeding the reported $\pm 1\sigma$ range and remains optimal or competitive in visual metrics. In the ID scenes, Physics-AdaNet remains competitive with the baselines across all metrics, indicating that adaptive alignment preserves in-distribution performance while its gains concentrate on the OOD scenes.

4.2 QUALITATIVE RESULTS

Figure 3 (a) presents representative sequences in which ball trajectories are extracted across sampled frames for qualitative comparison. In the universal gravitation task, CogVideoX and PISA fail to extrapolate horizontal motion beyond the initial frames. VideoREPA exhibits limited extrapolation but remains clearly inferior to ours. Only our model produces substantial lateral displacement, resulting



(a) OOD scenes of the two benchmark tasks.

(b) A real-world universal gravitation scene (ball).

Figure 3: Qualitative comparison on the OOD scenes of the two benchmark tasks (a) and a real-world universal gravitation scene (b). (a) Columns show the ground truth and the four compared methods; rows present two representative scenes for each task, with the universal gravitation task on top and the conservation of momentum task at the bottom. Orange and green boxes mark generalization failures in the two tasks, respectively. (b) Each model is conditioned on the same initial frame of real-world footage, and columns show uniformly sampled frames of the generated sequences.

in a physically plausible parabolic trajectory. In the conservation of momentum task, our model captures the uniform leftward motion of the ball b in the first scene, whereas other methods stay nearly stationary after the control frames. Our method therefore localizes the collision point more precisely and predicts a physically correct post-collision direction of the ball a, while all baselines predict the opposite direction. The second scene shows the same trend, with our approach producing both the correct direction and more accurate velocities for the ball a after collision. Beyond the two benchmark tasks, our method can be applied directly to real-world scenes. As shown in Figure 3 (b), conditioned on the same initial frame of real-world footage, Physics-AdaNet successfully extrapolates the projectile trajectory of the real-world object, sustaining horizontal motion where the objects generated by all baselines fall to the ground prematurely. Additional visual results are provided in the appendix and on our project website, including a preliminary qualitative study on real-world scenes with frame-level comparisons.

4.3 Ablation Study

To evaluate the contribution of each core component, we conduct ablation experiments under both ID and OOD scenes, with full results reported in Table 2 in Appendix C. In OOD scenes, both AdaAlign and the NAS-based Projector contribute substantially to generalization, and deviating from the optimal number of masked frames in either direction degrades kinematic fidelity; in ID scenes, all configurations perform similarly, with the mask strategy incurring only a slight trade-off in visual quality. Moreover, the projector searched on one task degrades when retrained on the other, confirming that the searched architectures are scenario-dependent rather than universally optimal. Further analysis of NAS can be found in Appendices D and E.

5 Conclusion

We have studied the OOD generalization of physical laws in video generation, where a model must re-compose invariant governing laws under structural distribution shifts rather than recall memorized visual patterns. This poses two fundamental challenges: internalizing invariant physical laws from environment-specific visual data and adapting the alignment architecture to diverse physical environments. Physics-AdaNet addresses them by coupling Adaptive Physics Alignment (AdaAlign), which distills relational physical priors into the generative backbone through spatiotemporal alignment with temporal masking, with a NAS-based Adaptive Projector that searches for the optimal alignment architecture of each scenario. On multiple benchmark tasks with structural OOD shifts, Physics-AdaNet achieves the best kinematic accuracy in every OOD scene while remaining competitive with state-of-the-art baselines on ID scenes and visual-fidelity metrics, with preliminary qualitative evidence of transfer to real-world footage.

REPRODUCIBILITY STATEMENT

We provide a complete description of the benchmark construction and data-sampling distributions in Section 3.1 and Appendix A, the full search space in Table 3, the searched optimal architectures in Appendix E, a time-complexity analysis of the NAS overhead in Appendix D, and all evaluation-metric definitions in Appendix B. All quantitative results are reported over three independent random seeds. Anonymous source code and video results are included in the supplementary materials.

AI USE STATEMENT

In this work, we used generative AI tools to assist with organizing experimental results and polishing the writing of this paper. We have not used generative AI tools for any other tasks, including ideation, method design, experimental implementation, and data analysis. All AI-assisted content was carefully reviewed and verified by the authors against the original experimental records. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. *arXiv preprint arXiv:1611.02167*, 2016.

Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Smash: one-shot model architecture search through hypernetworks. *arXiv preprint arXiv:1708.05344*, 2017.

Harold Haodong Chen, Haojian Huang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Hierarchical fine-grained preference optimization for physically plausible video generation, 2025. URL https://arxiv.org/abs/2508.10858.

Xinyu Gong, Shiyu Chang, Yifan Jiang, and Zhangyang Wang. Autogan: Neural architecture search for generative adversarial networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3224–3234, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Hao-Yu Hsu, Zhi-Hao Lin, Albert Zhai, Hongchi Xia, and Shenlong Wang. Autovfx: Physically realistic video editing from natural language instructions, 2024. URL https://arxiv.org/abs/2411.02394.

Sihui Ji, Xi Chen, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Physmaster: Mastering physical representation for video generation via reinforcement learning, 2025. URL https://arxiv.org/abs/2510.13809.

Rama Kandukuri, Jan Achterhold, Michael Moeller, and Joerg Stueckler. Physical representation learning and parameter identification from video using differentiable physics. *International Journal of Computer Vision*, 130:1934–1957, 2022.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? post-training newton's laws with verifiable rewards. *arXiv preprint arXiv:2512.00425*, 2025.

Changlin Li, Jiefeng Peng, Liuchun Yuan, Guangrun Wang, Xiaodan Liang, Liang Lin, and Xiaojun Chang. Block-wisely supervised neural architecture search with knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1989–1998, 2020.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv preprint arXiv:2503.09595*, 2025.

Shiqian Li, Ruihong Shen, Yaoyu Tao, Chi Zhang, and Yixin Zhu. Neural force field: Few-shot learning of generalized physical reasoning, 2026. URL https://arxiv.org/abs/2502.08987.

Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models, 2024. URL https://arxiv.org/abs/2309.17444.

Wang Lin, Liyu Jia, Wentao Hu, Kaihang Pan, Zhongqi Yue, Wei Zhao, Jingyuan Chen, Fei Wu, and Hanwang Zhang. Reasoning physical video generation with diffusion timestep tokens via reinforcement learning, 2025. URL https://arxiv.org/abs/2504.15932.

Martin Link, Max Schwarz, and Sven Behnke. Predicting physical object properties from video. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2022.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation, 2024. URL https://arxiv.org/abs/2409.18964.

Renqian Luo, Fei Tian, Tao Qin, Enhong Chen, and Tie-Yan Liu. Neural architecture optimization, 2019. URL https://arxiv.org/abs/1808.07233.

Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning, 2024. URL https://arxiv.org/abs/2311.12631.

Saman Motamed, Minghao Chen, Luc Van Gool, and Iro Laina. Travl: A recipe for making video-language models better judges of physics implausibility, 2025. URL https://arxiv.org/abs/2510.07550.

Ramakanth Pasunuru and Mohit Bansal. Continual and multi-task architecture search, 2019. URL https://arxiv.org/abs/1906.05226.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *International conference on machine learning*, pp. 4095–4104. PMLR, 2018.

Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V Le, and Alexey Kurakin. Large-scale evolution of image classifiers. In *International conference on machine learning*, pp. 2902–2911. PMLR, 2017.

Hao Tang, Kevin Ellis, Suhas Lohit, Michael J Jones, and Moitreya Chatterjee. Programmatic video prediction using large language models. *arXiv preprint arXiv:2505.14948*, 2025.

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

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, 2025a. URL https://arxiv.org/abs/2509.20358.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, Yuhui Yin, and Xiaodan Liang. Wisa: World simulator assistant for physics-aware text-to-video generation, 2025b. URL https://arxiv.org/abs/2503.08153.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking, 2023. URL https://arxiv.org/abs/2303.16727.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Zijun Wang, Panwen Hu, Jing Wang, Terry Jingchen Zhang, Yuhao Cheng, Long Chen, Yiqiang Yan, Zutao Jiang, Hanhui Li, and Xiaodan Liang. Prophy: Progressive physical alignment for dynamic world simulation. *arXiv preprint arXiv:2512.05564*, 2025c.

Lingxi Xie and Alan Yuille. Genetic cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1379–1388, 2017.

Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. Snas: Stochastic neural architecture search, 2020. URL https://arxiv.org/abs/1812.09926.

Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang. Physanimator: Physics-guided generative cartoon animation, 2025. URL https://arxiv.org/abs/2501.16550.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation, 2025. URL https://arxiv.org/abs/2412.00596.

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, et al. Vlipp: Towards physically plausible video generation with vision and language informed physical prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12360–12370, 2025.

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.

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. *arXiv preprint arXiv:2509.21309*, 2025.

Ke Zhang, Cihan Xiao, Jiacong Xu, Yiqun Mei, and Vishal M. Patel. Think before you diffuse: Infusing physical rules into video diffusion, 2025a. URL https://arxiv.org/abs/2505.21653.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025b.

Qi Zhao, Xingyu Ni, Ziyu Wang, Feng Cheng, Ziyan Yang, Lu Jiang, and Bohan Wang. Synthetic video enhances physical fidelity in video synthesis, 2025. URL https://arxiv.org/abs/2503.20822.

Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

A EXPERIMENTAL SETUP

Following Kang et al. (2024), we synthesize two Box2D datasets, each containing 8000 training videos and 100 videos per ID/OOD test split, with 33 frames per video at a spatial resolution of 720×480. The ball radii are sampled from $\mathcal{U}(0.5, 1.5)$ for geometric diversity. In the universal gravitation task, we set $g = 4.0$; the vertical velocity v_y of the training samples is drawn from $\mathcal{U}(-5.0, 5.0)$, while the OOD set adds the horizontal velocity $v_x \sim \mathcal{U}([-5.0, -2.0] \cup [2.0, 5.0])$. In the conservation of momentum task, ball a moves with $v_{a,x} \sim \mathcal{U}(2.0, 5.0)$ against a stationary ball b during training, and the OOD set additionally assigns $v_{b,x} \sim \mathcal{U}(-5.0, -2.0)$ to ball b.

Physics-AdaNet is built upon the CogVideoX-5B I2V model and trained with full-parameter Supervised Fine-Tuning (SFT) on the transformer backbone, so that the physical laws are learned directly from the target environments rather than relying on the pre-trained prior. The alignment targets in Eq. (10) are produced by a VideoMAEv2 encoder (Wang et al., 2023) with its standard public checkpoint, which always observes ground-truth frames. As shown in Figure 4, the hyperparameter λ is determined through a preliminary experiment without architecture search, trained for the same 8000 steps as the main experiments with all other settings kept strictly identical, and set to 0.35 and 0.65 for the two tasks, respectively. Critically, λ is selected on a held-out validation split of the training distribution that is completely disjoint from both test sets, so no information about them is involved in the selection of the hyperparameters.

The supernet is first trained for 16000 steps until the shared weights converge. We then apply an evolutionary search with population size $p = 50$ for $m = 10$ iterations and scaling factor $\alpha = 0.8$. In each iteration, $k_1 = 5$ elites are preserved, while mutation and crossover parents are drawn from the top $k_2 = 20$ candidates. Each candidate architecture is evaluated on 100 validation videos with $\mathcal{L}_{\text{score}}$ as the fitness metric, and the selected architecture is retrained as a standalone model for 8000 steps.

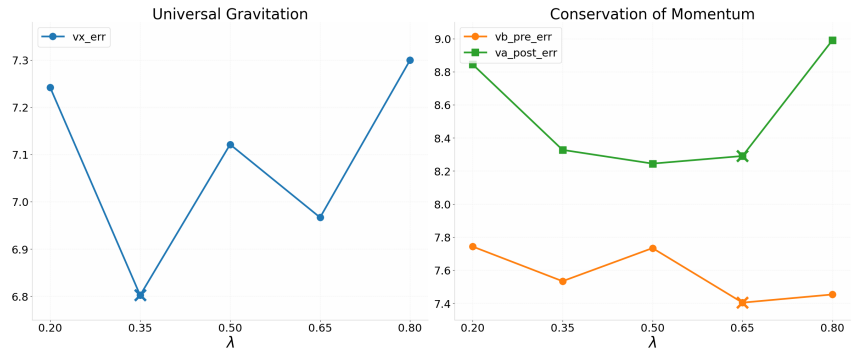


Figure 4: Impact of the hyperparameter λ on kinematic accuracy across two physical scenarios, measured on a held-out validation split that is disjoint from both the ID and OOD test sets. Left: the universal gravitation task. Right: the conservation of momentum task.

B EVALUATION METRIC DEFINITIONS

Kinematic accuracy is measured as the velocity error v_{err} between the extracted trajectories and their theoretical ground truth. Object centroids $\mathbf{p}_t = (x_t, y_t)$ are localized by pixel-level segmentation, and instantaneous velocities v_t are obtained by temporal differencing.

In the universal gravitation task, we compute the ℓ_2 error between the extracted horizontal velocity sequence v_x and its theoretical ground truth $\hat{v}_{x,0}$:

$$v_{x,\text{err}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(v_{x,t} - \hat{v}_{x,0}\right)^2} \tag{15}$$

This metric directly tests cross-dimensional generalization, i.e., whether the constant-velocity principle learned from vertical motion transfers to horizontal uniform motion.

For the conservation of momentum task, we first identify the collision frame $t_{\text{col}} = \arg\min_t \|\mathbf{p}_{a,t} - \mathbf{p}_{b,t}\|_2$ and partition the video into pre-collision and post-collision segments, then compare the pre-collision velocity of ball b and the post-collision velocity of ball a against the theoretical values $\hat{v}_x$ given by the conservation of momentum:

$$v_{b,\text{pre,err}} = \sqrt{\frac{1}{t_{\text{col}}} \sum_{t \le t_{\text{col}}} \left(v_{x,b,t} - \hat{v}_{x,b,\text{pre}}\right)^2}, \quad v_{a,\text{post,err}} = \sqrt{\frac{1}{T - t_{\text{col}} + 1} \sum_{t \ge t_{\text{col}}} \left(v_{x,a,t} - \hat{v}_{x,a,\text{post}}\right)^2} \tag{16}$$

These two metrics target the generalization from single-object scenes to dual-moving collisions. In the OOD setting where ball b moves leftward instead of staying stationary, $v_{b,\text{pre,err}}$ measures the transfer of uniform-motion rules across objects, while $v_{a,\text{post,err}}$ reflects mastery of the momentum-transfer term $\frac{2m_b v_b}{m_a+m_b}$. The remaining term $\frac{(m_b-m_a)v_b}{m_a+m_b}$ is less informative, since the independently and identically sampled masses make its coefficient zero-centered and often close to zero, unlike the strictly positive momentum-transfer term.

C Detailed ablation results

Table 2 reports the complete quantitative results of the ablation study discussed in Section 4.3. All configurations share the same training protocol as the main experiments, with results averaged over 3 independent runs. The baseline without NAS-Projector adopts the default hand-crafted configuration, i.e., the heavy 3-layer MLP projector from VideoREPA (Zhang et al., 2025b). Its gap to the full model in OOD scenes shows that hand-designed projectors are sub-optimal for a given task.

To further verify that the searched projector is scenario-dependent rather than universally optimal, we conduct a cross-task architecture transfer experiment. Specifically, we take the optimal architecture identified by the evolutionary search on one task (Table 4) and retrain it as a standalone model on the other task under the identical training protocol, denoted as *w/ projector from Task i* when the architecture is searched on Task i and retrained on the other task. As shown in Table 2, transferring the projector across tasks degrades the OOD kinematic accuracy relative to the projector searched on the target task, indicating that distinct physical scenarios favor different alignment architectures. This confirms that per-scenario architecture adaptation, rather than a single hand-crafted or cross-task-shared design, is necessary for effective physical-law internalization.

D Time complexity analysis of NAS

To quantify the training overhead introduced by NAS, we provide a time-complexity analysis. Since the per-step cost matches that of baseline training and the selection overhead of the evolutionary search is marginal, the total complexity is governed by the step counts alone. Conventional training costs $\Theta(n)$, where n is the number of steps to convergence, while our method costs $\Theta(T+mpv+n)$: T is the number of supernet training steps, which depends on n and $|\mathcal{S}|$; mpv is the cost of the evolutionary search, with m iterations, population size p, and v evaluation videos per candidate; and the final n is the retraining of the selected architecture.

Based on our experimental analysis, we set $n = 8000$ as the convergence steps of the baseline. Despite the extensive search space, the supernet converges efficiently due to its projection-based NAS design, resulting in $T = 16000$. With search parameters $m = 10$, $p = 50$, and $v = 100$, and because candidate evaluation requires only forward passes, which empirically is about $\frac{1}{30}$ of a full training step, the total cost amounts to $16000 + [50 + (10-1) \times (50-5)] \times 100/30 + 8000 \approx 25500$ equivalent training steps, where the $k_1 = 5$ elites preserved after the first generation are excluded from redundant evaluation. Thus, our method involves a roughly three-fold increase in training overhead over the baseline $n = 8000$. We contend that this one-time training investment is well-justified: the resulting model incurs no additional inference cost over the baseline while achieving significantly superior OOD kinematic accuracy.

Furthermore, Figure 5 presents the evolution of the best loss score and the mean population loss score across generations. We observe that the population-mean loss drops rapidly, entering a convergence phase after the second iteration. Meanwhile, the best loss score follows a smoother downward

Table 2: Ablation study of Physics-AdaNet, grouped by ID and OOD scenes. ↓ and ↑ denote the direction of better performance. Bold indicates the best performance, while underline represents the second-best result. All metrics report mean ± standard deviation over 3 independent runs. *w/o NAS-Projector* replaces the searched projector with the default hand-crafted 3-layer MLP projector from VideoREPA. *w/ projector from Task 1/2* retrains the optimal architecture searched on the other task under the identical protocol, measuring the cross-task transferability of the searched projector.

Task 1: Newton's Law of Universal Gravitation						
Type	Configuration	$v_{x,\mathrm{err}}$ ↓	FVD↓	SSIM↑	PSNR↑	LPIPS↓
ID	w/o AdaAlign	0.1961±0.0046	**18.8121**±2.3761	0.9908±0.0002	32.7264±0.0700	**0.0334**±0.0003
	mask $k = 0$ frames	0.2401±0.0077	20.7070±1.1693	**0.9917**±0.0002	**34.5845**±0.2376	0.0447±0.0008
	mask $k = 2$ frames	0.1946±0.0042	28.2713±1.1427	0.9913±0.0003	32.0610±0.1157	0.0737±0.0012
	w/o NAS-Projector	0.2047±0.0056	30.9490±1.7786	0.9912±0.0002	31.6306±0.1409	0.0830±0.0005
	w/ projector from Task 2	0.2002±0.0049	30.2452±1.6273	0.9911±0.0002	32.8517±0.1622	0.0762±0.0007
	Physics-AdaNet	**0.1854**±0.0053	28.2793±1.4998	0.9915±0.0001	34.3286±0.1303	0.0613±0.0008
OOD	w/o AdaAlign	7.6781±0.1778	18.3885±1.6394	0.9880±0.0004	29.4198±0.2541	0.0510±0.0013
	mask $k = 0$ frames	4.8969±0.1355	**16.3627**±2.2348	0.9878±0.0003	28.8025±0.1260	**0.0478**±0.0017
	mask $k = 2$ frames	5.8854±0.1581	18.8367±3.9577	0.9874±0.0004	32.3927±0.2627	0.0562±0.0016
	w/o NAS-Projector	6.8020±0.1501	25.7122±1.3023	0.9871±0.0003	29.2660±0.2919	0.0685±0.0008
	w/ projector from Task 2	6.4588±0.1476	21.9826±1.8773	0.9875±0.0003	30.5441±0.2241	0.0611±0.0011
	Physics-AdaNet	**4.3210**±0.1355	17.2379±1.4154	**0.9902**±0.0004	**33.3071**±0.2001	0.0494±0.0015

Task 2: Law of Conservation of Momentum							
Type	Configuration	$v_{b,\mathrm{pre},\mathrm{err}}$ ↓	$v_{a,\mathrm{post},\mathrm{err}}$ ↓	FVD↓	SSIM↑	PSNR↑	LPIPS↓
ID	w/o AdaAlign	2.4649±0.0689	3.0130±0.0907	18.3242±1.5775	0.9837±0.0002	27.4478±0.3309	0.0653±0.0012
	mask $k = 0$ frames	2.3139±0.0349	2.1936±0.0592	**12.0219**±0.8464	**0.9863**±0.0003	**29.8594**±0.1521	**0.0403**±0.0010
	mask $k = 2$ frames	2.0406±0.0651	2.1608±0.0537	14.9905±1.4431	0.9849±0.0005	29.2093±0.2902	0.0611±0.0018
	w/o NAS-Projector	**2.0012**±0.0651	**1.8511**±0.0343	21.4611±0.5891	0.9847±0.0003	29.0918±0.2160	0.0652±0.0017
	w/ projector from Task 1	2.2871±0.0554	2.1746±0.0612	18.7288±1.2246	0.9840±0.0003	28.1267±0.2426	0.0584±0.0013
	Physics-AdaNet	2.1603±0.0579	1.9862±0.0692	14.3536±0.6371	0.9842±0.0004	28.4910±0.1902	0.0463±0.0009
OOD	w/o AdaAlign	6.8255±0.1975	9.9163±0.2326	30.9228±1.0505	0.9752±0.0010	22.7242±0.1339	0.0943±0.0028
	mask $k = 0$ frames	6.1294±0.1570	8.8746±0.3794	**23.2312**±5.2655	0.9749±0.0005	22.7424±0.1079	**0.0792**±0.0022
	mask $k = 2$ frames	6.5752±0.0973	9.0193±0.2847	26.8610±1.8833	0.9748±0.0009	22.5454±0.1173	0.0807±0.0021
	w/o NAS-Projector	7.4039±0.2475	8.2445±0.1951	31.8781±1.3837	0.9742±0.0004	22.4283±0.3555	0.1140±0.0015
	w/ projector from Task 1	6.6457±0.1838	8.5127±0.2495	31.2657±2.1435	0.9746±0.0006	22.6739±0.1961	0.0985±0.0020
	Physics-AdaNet	**5.1170**±0.1593	**7.5075**±0.2703	29.8604±2.3684	**0.9753**±0.0007	**23.0088**±0.1913	0.0888±0.0023

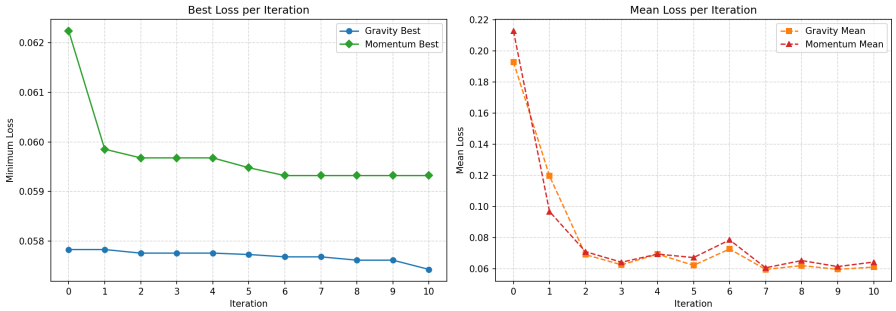


Figure 5: Evolutionary search curves showing the best and population-mean loss scores across two tasks. The x-axis represents the generation number, including the initial population with index 0, while the y-axis indicates the loss score.

trajectory, indicating that the evolutionary process continues to discover architectures with superior alignment performance throughout subsequent generations.

E Optimal architecture configuration of NAS

In this section, we provide the detailed definition of the search space $\mathcal{S}$ and the specific architectural configurations derived from our evolutionary search algorithm. The definition of the search space $\mathcal{S}$ is given in Table 3, and the optimal architecture specifications identified for the two distinct scenarios are summarized in Table 4. These configurations represent the best candidates within the defined search space $\mathcal{S}$.

Table 3: Search space configuration for the Adaptive Projector.

Hyperparameter	Search Candidates
Alignment Layer (a)	$\mathcal{A} = \{12, 15, 18, 21, 24\}$
Normalization Method (n)	$\mathcal{N} = \{\text{None}, \text{BN}, \text{LN}, \text{RMS}\}$
Projector Depth (c)	$\mathcal{C} = \{2, 3, 4, 5\}$
Hidden Dimension (dim)	$\mathcal{D} = \{1024, 2048, 3072, 4096\}$
Module Type ($type_i$)	$\mathcal{W} = \{\text{Linear}, \text{Self-Attn}, \text{Conv1D}\}$
Activation Function (act_i)	$\mathcal{F} = \{\text{Identity}, \text{GeLU}, \text{SiLU}, \text{SwiGLU}\}$

Table 4: Optimal architecture specifications for two scenarios identified via evolutionary search.

Hyperparameter	Universal Gravitation	Conservation of Momentum
Alignment Layer (a)	18	12
Normalization Method (n)	BatchNorm	RMSNorm
Projector Depth (c)	2	2
Hidden Dimension (dim)	1024	2048
Module Type ($type_i$)	(Linear, Linear)	(Linear, Linear)
Activation Function (act_i)	(SiLU, Identity)	(SwiGLU, Identity)

Notably, the architectures discovered by NAS converge to a consistent and structurally simple pattern: a shallow two-layer projector built solely from Linear modules, with a normalization layer and a modest hidden dimension. Identifying this simplicity is itself a non-trivial outcome of the search. This diverges from conventional default settings: despite the inclusion of Self-Attn and Conv1D in the search space, complex modules consistently underperform those built solely from Linear layers, suggesting that they introduce unnecessary inductive biases that interfere with the projection of feature distributions. At roughly $3\times$ the standard training cost, NAS recovers such lightweight designs in a data-driven manner without human bias, which facilitate the extraction of abstract information and impose more effective constraints on the model's feature distribution.

Sensitivity of the scaling coefficient α. Table 5 reports the scaled fitness score $\mathcal{L}_{\text{score}}$ (Eq. (14)) of each candidate alignment layer under different values of α in the universal gravitation scenario, with bold entries marking the layer selected by NAS under each α. Without scaling ($\alpha = 0.0$), the score decreases monotonically with layer depth and the search degenerates to the deepest candidate ($a_{\text{cur}} = 24$), confirming the order-of-magnitude imbalance of the raw $\mathcal{L}_{\text{TRD}}$ across depths. As α increases, this depth-induced bias is progressively counteracted: under our default $\alpha = 0.8$, the search recovers the intermediate layer $a_{\text{cur}} = 18$, matching the optimal configuration in Table 4. The scaling mechanism thus does not simply favor deeper or shallower layers, but provides a more faithful estimate of each candidate's architectural efficiency.

Table 5: Effect of the scaling coefficient α on $\mathcal{L}_{\text{score}}$ across alignment-layer indices a_{cur} in the universal gravitation scenario. Bold denotes the layer selected by NAS under each α.

a_{cur}	$\mathcal{L}_{\text{TRD}}$ ($\alpha = 0.0$)	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1.0$
12	0.10239	0.10239	0.10239	**0.10239**
15	0.08703	0.09748	0.10096	0.10444
18	0.07065	0.09184	**0.09891**	0.10597
21	0.05632	0.09010	0.10137	0.11263
24	**0.03891**	**0.08560**	0.10116	0.11673

F MORE VISUALIZATION RESULTS

As a supplement to the main text, this section provides additional qualitative visualizations in Figures 6, 7, 8, and 9. The corresponding video demonstrations are available on our project website at `https://physicsadanet.github.io/Physics-AdaNet/`.

For the ID scenes not discussed in the main text, our approach remains competitive with the baselines. Notably, in the ID scenes of the conservation of momentum scenario highlighted with green borders, our method predicts more accurate velocity directions of the ball a where other models fail entirely. Consistent with the OOD observations, these ID cases indicate that conventional models can struggle to capture the principles of velocity exchange under the law of conservation of momentum, underscoring the value of adaptive alignment even within the training distribution.

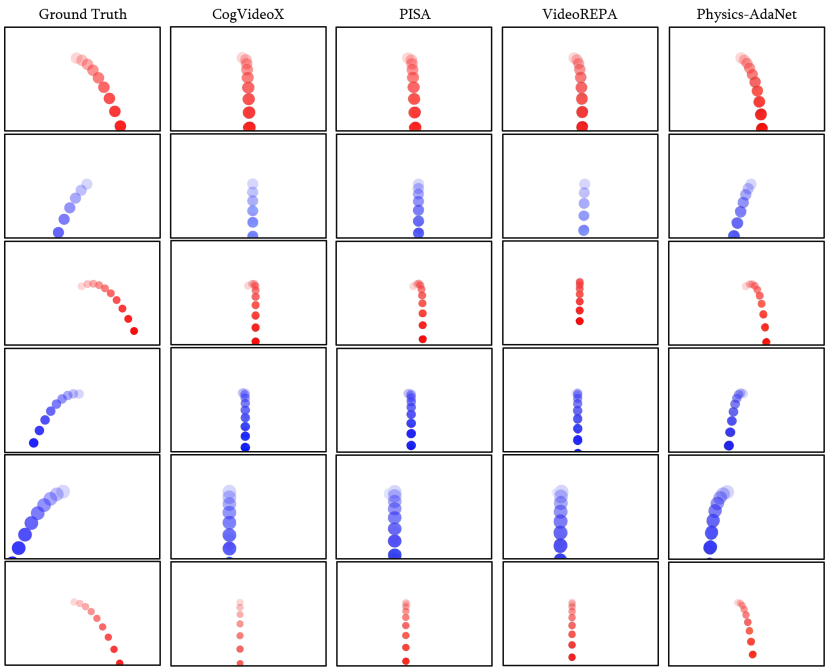


Figure 6: More qualitative comparison of physical law generalization for the OOD settings of universal gravitation.

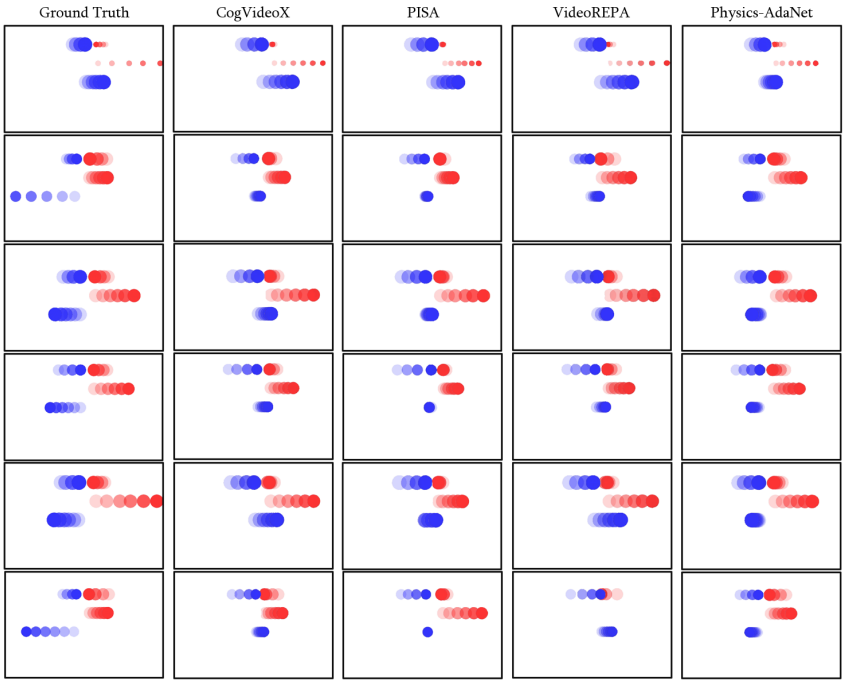


Figure 7: More qualitative comparison of physical law generalization for the OOD settings of conservation of momentum.

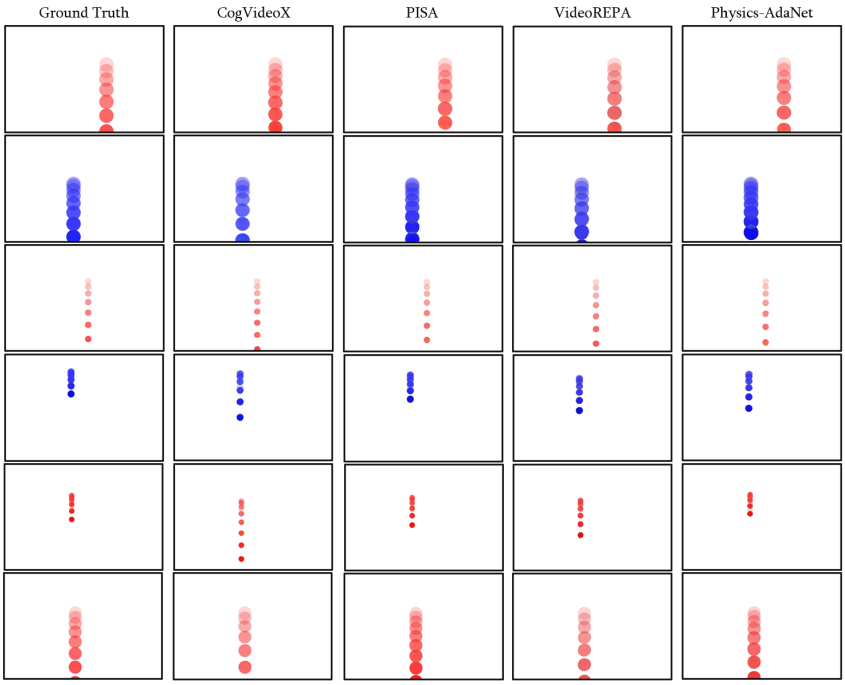


Figure 8: More qualitative comparison of physical law generalization for the ID settings of universal gravitation.

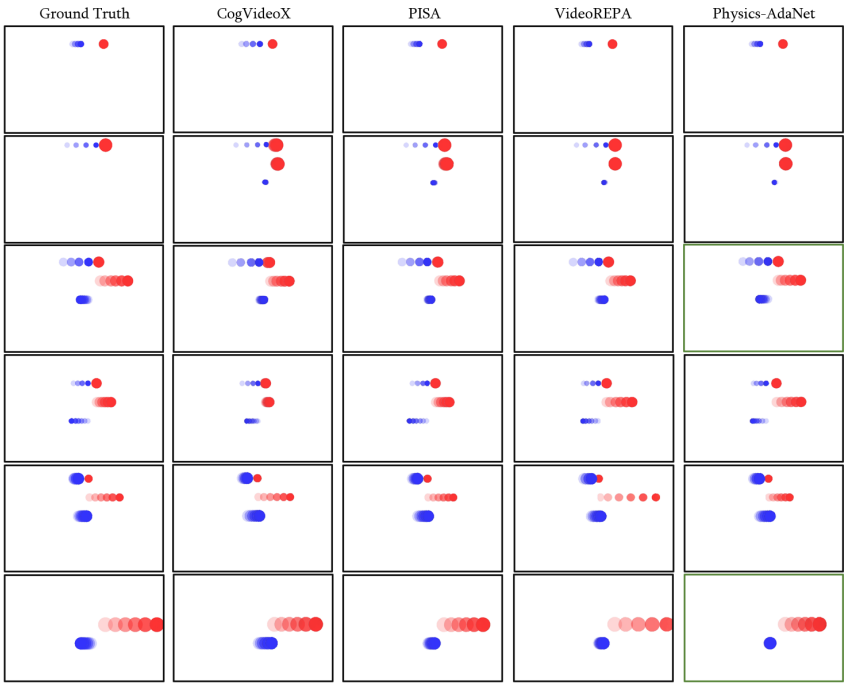


Figure 9: More qualitative comparison of physical law generalization for the ID settings of conservation of momentum.

G MORPHOLOGICAL GENERALIZATION

To assess the morphological generalization of the proposed framework, we further conduct inference under geometric domain shifts, replacing the spherical objects in both ID and OOD scenes with cubes while keeping the physical laws and experimental configurations unchanged. Table 6 summarizes the quantitative results, and Figures 10, 11, 12, and 13 present the corresponding qualitative visualizations.

Under this setting, the absolute metrics of all methods degrade while their relative ordering remains consistent with the main experiments: our framework attains the best kinematic accuracy in OOD scenes with competitive or better video-quality metrics, and remains competitive in ID scenes while achieving the best results on several metrics. One exception is the elevated FVD in the OOD universal gravitation scene, which stems from a distinct diagonal deformation of the cubes—absent in baseline outputs—as they transition from vertical motion to projectile motion. This deformation is a by-product of the correct lateral displacement that all baselines fail to produce, rather than a degradation in visual fidelity itself, as visualized in Figure 10.

Qualitatively, VideoREPA shows partial extrapolative capability in the gravitation task, whereas our framework generalizes more robustly in OOD settings. In the conservation of momentum task, our method better preserves the initial velocity of the right-hand object in OOD scenes and predicts the post-collision direction and magnitude of the left-hand object substantially more accurately. Finally, all methods exhibit some shape degradation on cubes relative to the ground truth, with sharp corners tending to become rounded, suggesting a shared limitation in reproducing geometric features unseen during training.

Table 6: Quantitative results of morphological generalization across two tasks, grouped by ID and OOD scenes. ↓ and ↑ denote the direction of better performance. Bold indicates the best performance, while underline represents the second-best result. All metrics report mean ± standard deviation over 3 independent runs.

Task 1: Newton's Law of Universal Gravitation

Type	Method	$v_{x,\mathrm{err}}$ ↓	FVD↓	SSIM↑	PSNR↑	LPIPS↓
ID	CogVideoX	0.2317±0.0037	**29.6394**±2.1868	0.9906±0.0003	32.6077±0.1441	0.0764±0.0006
	PISA	**0.1466**±0.0053	35.2098±2.0170	<u>0.9911</u>±0.0002	31.7686±0.1472	**0.0407**±0.0009
	VideoREPA	0.2645±0.0072	35.9971±2.7443	0.9906±0.0002	<u>32.8658</u>±0.1476	0.0550±0.0012
	Physics-AdaNet	<u>0.2245</u>±0.0051	<u>30.1507</u>±1.9599	**0.9919**±0.0002	**34.8353**±0.2613	<u>0.0413</u>±0.0006
OOD	CogVideoX	8.0224±0.1195	**30.6627**±2.8809	<u>0.9873</u>±0.0004	<u>28.9659</u>±0.1691	0.0626±0.0013
	PISA	8.0786±0.1969	<u>37.1427</u>±2.9158	0.9864±0.0003	27.2461±0.4113	0.0746±0.0014
	VideoREPA	<u>7.7554</u>±0.2033	40.2472±2.7412	0.9858±0.0002	28.2147±0.1311	<u>0.0476</u>±0.0014
	Physics-AdaNet	**5.2899**±0.1309	62.0545±3.3722	**0.9890**±0.0005	**31.7201**±0.3889	**0.0460**±0.0023

Task 2: Law of Conservation of Momentum

Type	Method	$v_{b,\mathrm{pre,err}}$ ↓	$v_{a,\mathrm{post,err}}$ ↓	FVD↓	SSIM↑	PSNR↑	LPIPS↓
ID	CogVideoX	3.7775±0.0607	2.9691±0.0478	29.2412±1.6795	<u>0.9858</u>±0.0003	27.6391±0.1704	<u>0.0611</u>±0.0005
	PISA	<u>2.2615</u>±0.0480	**2.3894**±0.0421	**22.0108**±1.6232	**0.9885**±0.0004	**30.2353**±0.0802	0.0712±0.0005
	VideoREPA	3.8914±0.0722	2.6662±0.0435	34.3384±1.2189	0.9854±0.0002	<u>27.9192</u>±0.0921	0.0659±0.0006
	Physics-AdaNet	**1.5991**±0.0273	<u>2.6460</u>±0.0450	<u>27.0909</u>±1.1072	0.9762±0.0002	25.9073±0.0734	**0.0518**±0.0005
OOD	CogVideoX	7.8750±0.1810	9.7604±0.2433	26.7584±1.9826	0.9757±0.0006	22.4051±0.3391	<u>0.0883</u>±0.0025
	PISA	<u>6.2009</u>±0.1460	<u>8.4719</u>±0.2565	**22.4459**±2.9217	0.9752±0.0012	22.4780±0.1866	0.0945±0.0021
	VideoREPA	6.9088±0.2434	10.1350±0.3459	26.7852±1.7413	**0.9762**±0.0007	**23.0343**±0.3483	0.0899±0.0015
	Physics-AdaNet	**5.5002**±0.1683	**7.0736**±0.2090	<u>25.0525</u>±2.0497	<u>0.9760</u>±0.0009	<u>22.7668</u>±0.2238	**0.0867**±0.0016

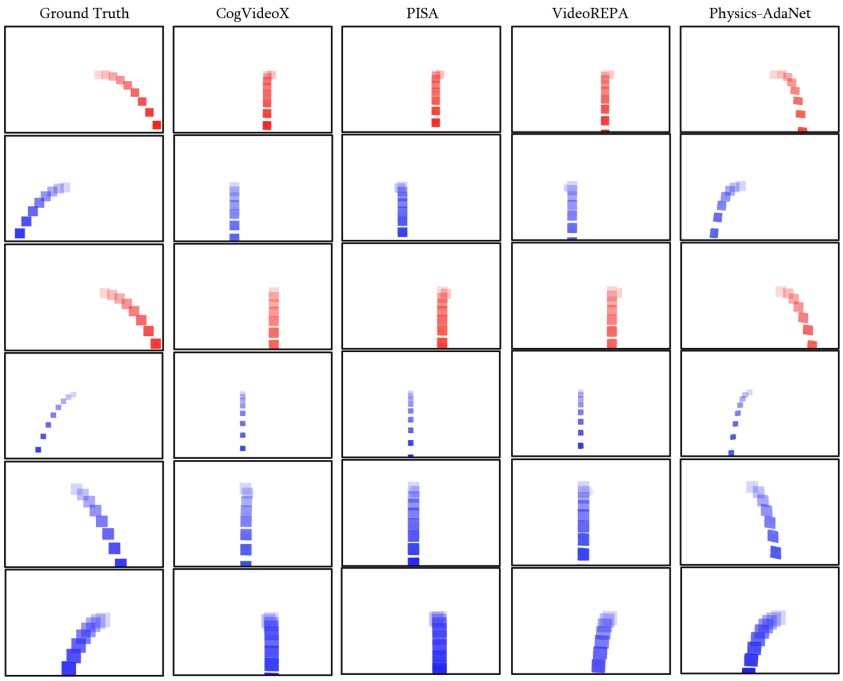


Figure 10: More qualitative comparison of morphological generalization for the OOD settings of universal gravitation.

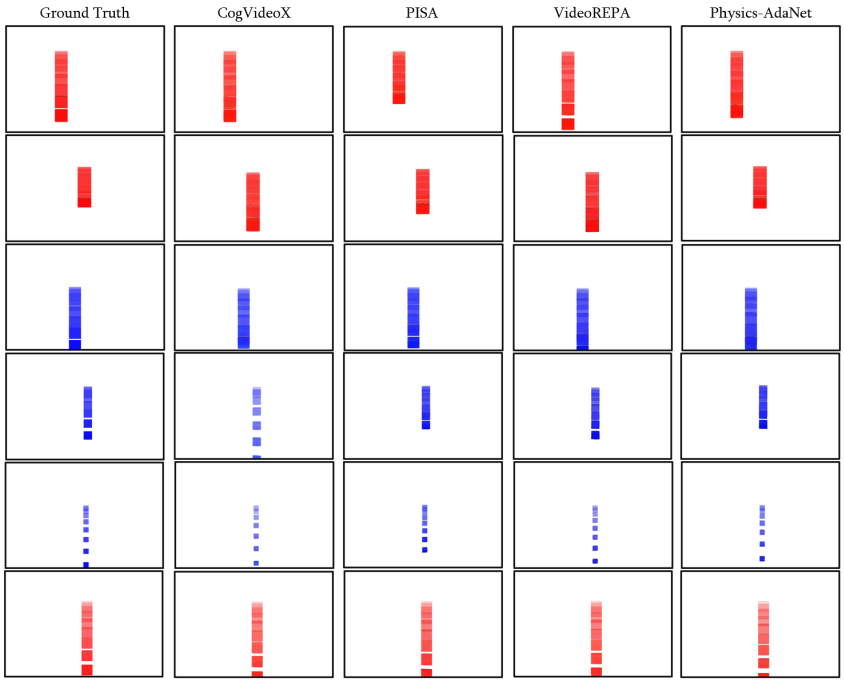


Figure 11: More qualitative comparison of morphological generalization for the ID settings of universal gravitation.

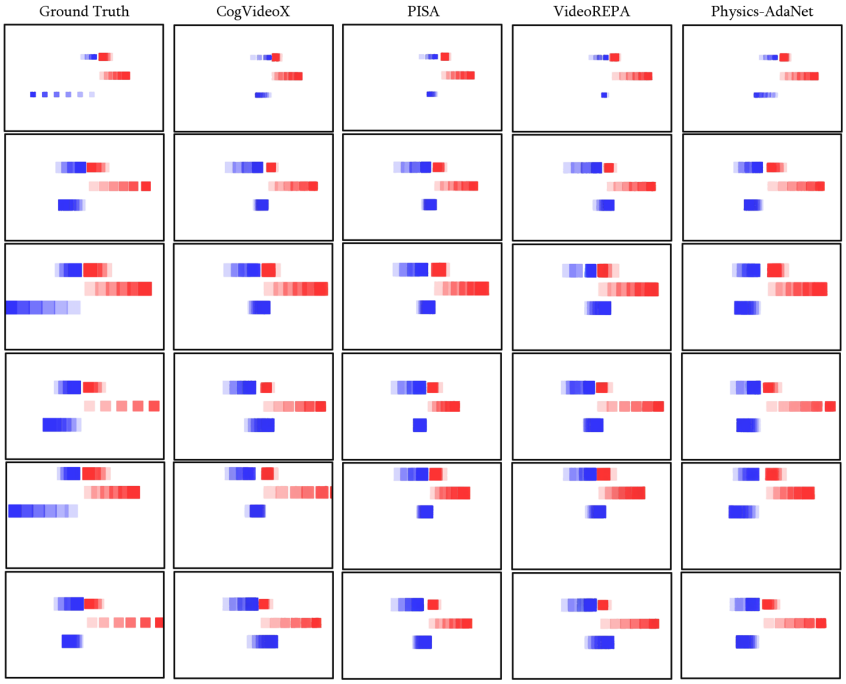


Figure 12: More qualitative comparison of morphological generalization for the OOD settings of conservation of momentum.

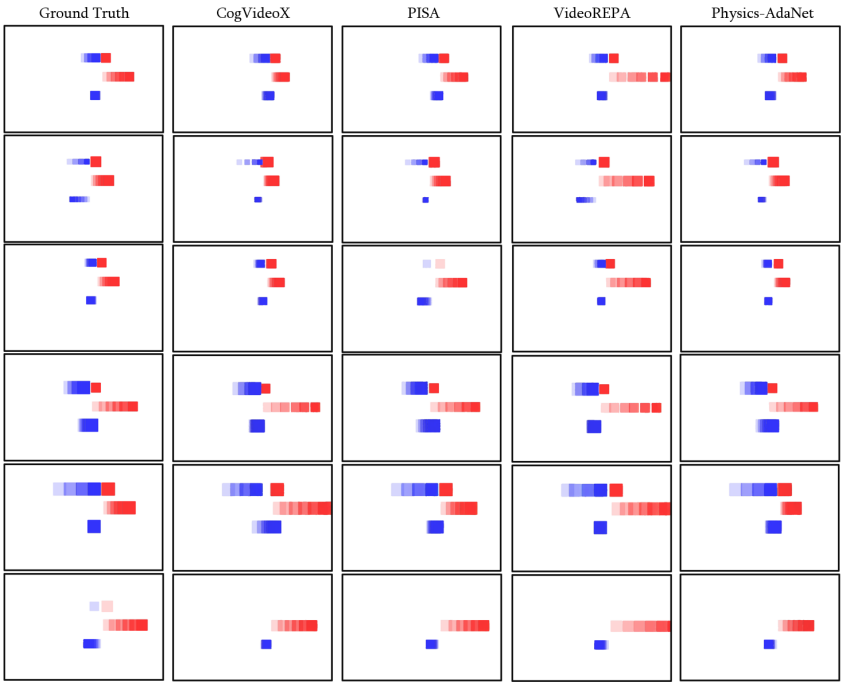


Figure 13: More qualitative comparison of morphological generalization for the ID settings of conservation of momentum.

H QUALITATIVE EVALUATION ON REAL-WORLD SCENES

To examine whether the alignment learned on synthetic data transfers beyond the simulated benchmark, we additionally conduct a qualitative study on real-world scenes. Following NewtonGen (Yuan et al., 2025), we use the previously generated 2D videos as motion maps, giving trajectory guidance to warp the initial noise, and each model generates the subsequent frames conditioned on the warped noise together with the same initial frame. We then uniformly sample frames from the generated sequences for visual comparison. Since precise ground-truth physical parameters (gravitational acceleration in the camera frame, object masses, and pre-/post-collision velocities) are not available for the real-world footage, this study is qualitative: we visually inspect whether the generated trajectories remain physically plausible rather than reporting kinematic errors.

Complementing the real-world universal gravitation scene (ball) shown in Figure 3 (b) of the main text, Figures 14–16 present the frame-level qualitative results for the real-world universal gravitation and conservation of momentum scenes with different object shapes (ball and cube), respectively. Across the inspected scenes, Physics-AdaNet consistently produces the most physically plausible trajectories among the compared methods. In the gravitation scenes, our method successfully extrapolates the projectile trajectory of the real-world object, sustaining horizontal motion where all baselines either stall or drop prematurely. In the collision scenes, our method yields noticeably more accurate post-collision velocity exchange, whereas the baselines fail to even faithfully generate the two objects and their collision interaction. Notably, in the gravitation scene with a cube, our method is the only one that reproduces the rotational shape change of the falling object observed in the ground truth, while all baselines keep the cube's orientation fixed throughout the sequence.

Limitations. Our benchmark remains deliberately simplified, excluding rotation, friction, deformation, and multi-body or fluid interactions, and our quantitative evaluation is confined to synthetic 2D videos, which provide the exact ground-truth kinematics that real-world footage cannot. The qualitative study above serves as a first step beyond this setting, and extending Physics-AdaNet to richer physical interactions with quantitative evaluation is a natural direction for future work.

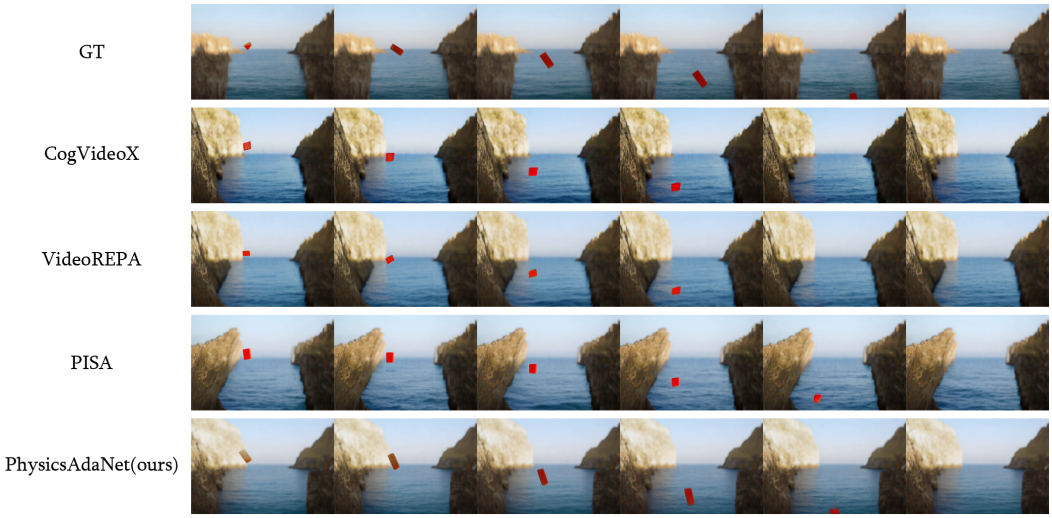


Figure 14: Qualitative results on a real-world universal gravitation scene (cube). Each model is conditioned on the same initial frame. Columns show uniformly sampled frames of the generated sequences.

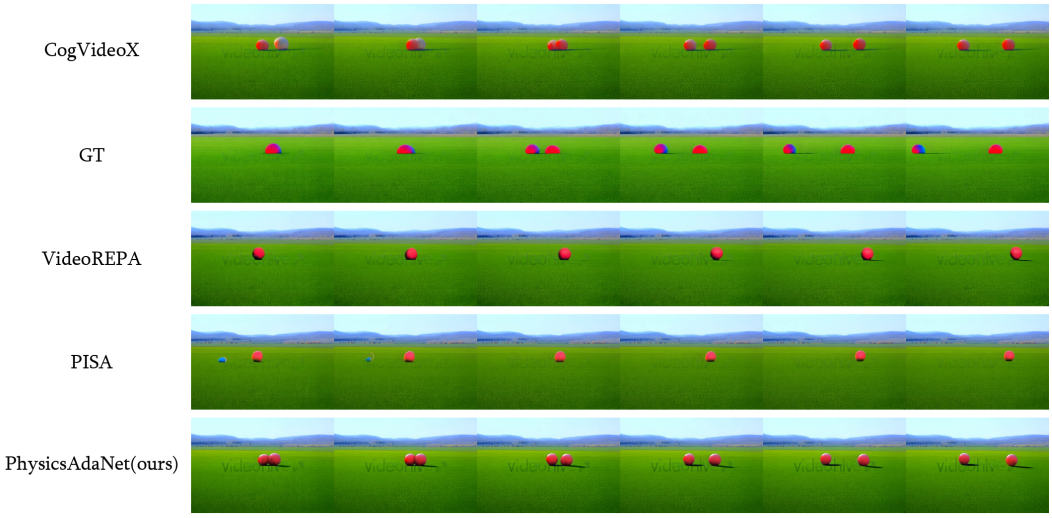


Figure 15: Qualitative results on a real-world conservation of momentum scene (balls). Each model is conditioned on the same initial frame. Columns show uniformly sampled frames of the generated sequences.

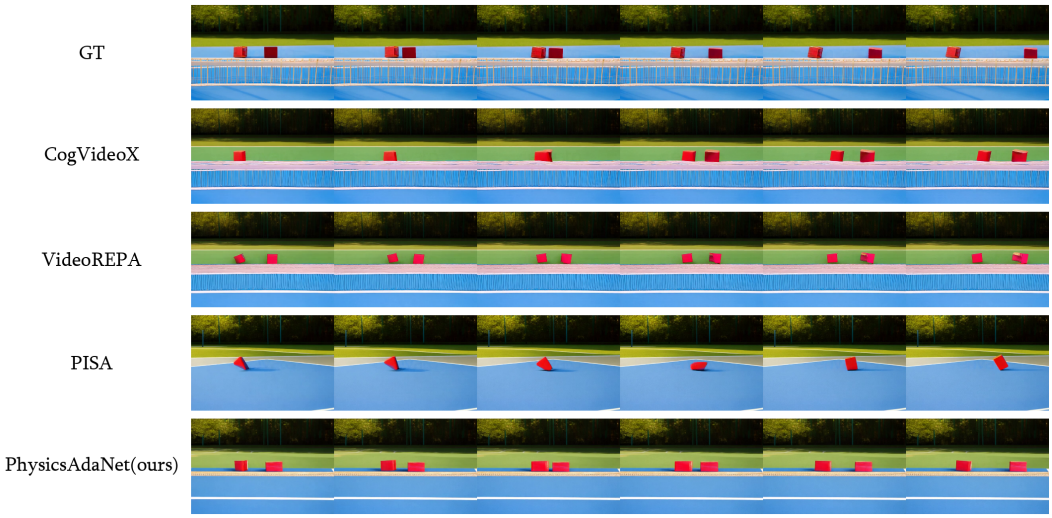


Figure 16: Qualitative results on a real-world conservation of momentum scene (cubes). Each model is conditioned on the same initial frame. Columns show uniformly sampled frames of the generated sequences.