How Far Is Video Generation from World Model: A Physical Law Perspective

Bingyi Kang[*1] **Yang Yue**[*1,2]
Rui Lu[2] **Zhijie Lin**[1] **Yang Zhao**[1] **Kaixin Wang**[3] **Gao Huang**[2] **Jiashi Feng**[1]
[*] *Equal Contribution (in alphabetical order)*
[1] Bytedance Research [2] Tsinghua University [3] Technion
Website: https://phyworld.github.io

Abstract

OpenAI's Sora highlights the potential of video generation for developing world models that adhere to fundamental physical laws. However, the ability of video generation models to discover such laws purely from visual data without human priors can be questioned. A world model learning the true law should give predictions robust to nuances and correctly extrapolate on unseen scenarios. In this work, we evaluate across three key scenarios: in-distribution, out-of-distribution, and combinatorial generalization. We developed a 2D simulation testbed for object movement and collisions to generate videos deterministically governed by one or more classical mechanics laws. This provides an unlimited supply of data for large-scale experimentation and enables quantitative evaluation of whether the generated videos adhere to physical laws. We trained diffusion-based video generation models to predict object movements based on initial frames. Our scaling experiments show perfect generalization within the distribution, measurable scaling behavior for combinatorial generalization, but failure in out-of-distribution scenarios. Further experiments reveal two key insights about the generalization mechanisms of these models: (1) the models fail to abstract general physical rules and instead exhibit "case-based" generalization behavior, *i.e.*, mimicking the closest training example; (2) when generalizing to new cases, models are observed to prioritize different factors when referencing training data: color > size > velocity > shape. Our study suggests that scaling alone is insufficient for video generation models to uncover fundamental physical laws, despite its role in Sora's broader success.

1 Introduction

Foundation models (Bommasani et al., 2021) have emerged remarkable capabilities by scaling the model and data to an unprecedented scale (Brown, 2020; Kaplan et al., 2020). As an example, OpenAI's Sora (Brooks et al., 2024) not only generates high-fidelity and surreal videos, but also has sparked a new surge of interest in studying world models (Yang et al., 2023).

> "*Scaling video generation models is a promising path towards building general purpose simulators of the physical world.*" — Sora Report (Brooks et al., 2024)

World simulators are receiving broad attention from robotics (Yang et al., 2023) and autonomous driving (Hu et al., 2023) for the ability to generate realistic data and accurate simulations. These models are required to comprehend fundamental physical laws to produce data that extends beyond the training corpus and to guarantee precise simulation. However, it remains an open question whether video generation can discover such rules merely by observing videos, as Sora does. We aim to provide a systematic study to understand the critical role and limitation of scaling in physical law discovery.

It is challenging to determine whether a video model has learned a law instead of merely memorizing the data. Since the model's internal knowledge is inaccessible, we can only infer the model's understanding by examining its predictions on unseen scenarios, *i.e.*, its generalization ability. We propose a categorization (Figure 1) for comprehensive evaluation based on the relationship between training and testing data in this paper. *In-distribution* (ID) generalization assumes that training and testing data are independent and identically distributed (*i.i.d.*). *Out-of-distribution* (OOD)

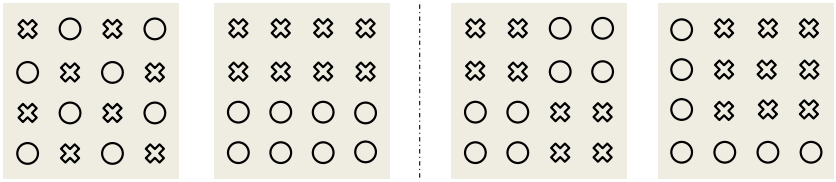

Figure 1: Categorization of generalization patterns.○ denotes training data. × denotes testing data.

generalization, on the other hand, refers to the model's performance on testing data that come from a different distribution than the training data, particularly when latent parameters fall outside the range seen during training. Human-level physical reasoning can easily extrapolate OOD and predict physical processes without encountering the exact same scenario before. Additionally, we also examine a special OOD capacity called *combinatorial* generalization, which assesses whether a model can combine two distinct concepts in a novel way, a trait often considered essential for foundation models in advancing toward artificial general intelligence (AGI) (Du & Kaelbling, 2024).

Moreover, real-world videos typically contain complex, non-rigid objects and motions, which present significant challenges for quantitative evaluation and even human validation. The rich textures and appearances in such videos can act as confounding factors, distracting the model from focusing on the underlying physics. To mitigate these issues, we specifically focus on classical mechanics and develop a 2D simulator with objects represented by simple geometric shapes. Each video depicts the motion or collision of these 2D objects, governed entirely by one or two fundamental physical laws, given the initial frames. This simulator allows us to generate large-scale datasets to support the scaling of video generation models. Additionally, we have developed a tool to infer internal states (*e.g.*, the position and size) of each object in the generated video from pixels. This enables us to establish quantitative evaluation metrics for physical law discovery.

We begin by investigating how scaling video generation models affects ID and OOD generalization. We select three fundamental physical laws for simulation: *uniform linear motion* of a ball, *perfectly elastic collision* between two balls, and *parabolic motion* of a ball. We scale the dataset from 30K to 3 million examples and increase the video diffusion model's parameters from 22M to 310M. Consistently, we observe that the model achieves near-perfect ID generalization across all tasks. However, the OOD generalization error does not improve with increased data and model size, revealing the limitations of scaling video generation models in handling OOD data. For combinatorial generalization, we design an environment that involves multiple objects undergoing free fall and collisions to study their interactions. Every time, four objects from eight are selected to create a video. In total, 70 combinations (C_8^4) are possible. We use 60 of them for training and 10 for testing. We train models by varying the number of training data from 600K to 6M. We manually evaluate the generated test samples by labeling them as "abnormal" if the video looks physically implausible. The results demonstrate that scaling the data substantially reduces the percentage of abnormal cases, from 67% to 10%. This suggests that scaling is critical for improving combinatorial generalization.

Our empirical analysis reveals two intriguing properties of the generalization mechanism in video generation models. First, these models can be easily biased by "deceptive" examples from the training set, leading them to generalize in a "case-based" manner under certain conditions. This phenomenon, also observed in large language models (Hu et al., 2024), describes a model's tendency to reference similar training cases when solving new tasks. For instance, consider a video model trained on data of a high-speed ball moving in uniform linear motion. If data augmentation is performed by horizontally flipping the videos, thereby introducing reverse-direction motion, the model may generate a scenario where a low-speed ball reverses direction after the initial frames, even though this behavior is not physically correct. Second, we explore how different data attributes compete during the generalization process. For example, if the training data for uniform motion consists of red balls and blue squares, the model may transform a red square into a ball immediately after the conditioning frames. This behavior suggests that the model prioritizes color over shape. Our pairwise analysis reveals the following prioritization hierarchy: color > size > velocity > shape. This ranking could explain why current video generation models often struggle with maintaining object consistency.

We hope these findings provide valuable insights for future research in video generation and the development of world models.

2 DISCOVERING PHYSICS LAWS WITH VIDEO GENERATION

2.1 PROBLEM DEFINITION

In this section, we aim to establish the framework and define the concept of physical laws discovery in the context of video generation. In classical physics, laws are articulated through mathematical equations that predict future state and dynamics from initial conditions. In the realm of video-based observations, each frame represents a moment in time, and the prediction of physical laws corresponds to generating future frames conditioned on past states.

Consider a physical procedure which involves several latent variables $\boldsymbol{z} = (z_1, z_2, \ldots, z_k) \in \mathcal{Z} \subseteq \mathbb{R}^k$, each standing for a certain physical parameter such as velocity or position. By classical mechanics, these latent variables will evolve by differential equation $\dot{\boldsymbol{z}} = F(\boldsymbol{z})$. In discrete version, if time gap between two consecutive frames is δ, then we have $\boldsymbol{z}_{t+1} \approx \boldsymbol{z}_t + \delta F(\boldsymbol{z}_t)$. Denote rendering function as $R(\cdot) : \mathcal{Z} \mapsto \mathbb{R}^{3\times H\times W}$ which render the state of the world into an image of shape $H \times W$ with RGB channels. Consider a video $V = \{I_1, I_2, \ldots, I_L\}$ consisting of L frames that follows the classical mechanics dynamics. The physical coherence requires that there exists a series of latent variables which satisfy following requirement: 1) $\boldsymbol{z}_{t+1} = \boldsymbol{z}_t + \delta F(\boldsymbol{z}_t), t = 1, \ldots, L-1$. 2) $I_t = R(\boldsymbol{z}_t), \quad t = 1, \ldots, L$. We train a video generation model p parametried by θ, where $p_\theta(I_1, I_2, \ldots, I_L)$ characterizes its understanding of video frames. We can predict the subsequent frames by sampling from $p_\theta(I'_{c+1}, \ldots I'_L \mid I_1, \ldots, I_c)$ based on initial frames' condition. The variable c usually takes the value of 1 or 3 depends on tasks. Therefore, physical-coherence loss can be simply defined as $-\log p_\theta(I_{c+1}, \ldots, I_L \mid I_1, \ldots, I_c)$. It measures how likely the predicted value will cater to the real world development. The model must understand the underlying physical process to accurately forecast subsequent frames, which we can quantatively evaluate whether video generation model correctly discover and simulate the physical laws.

2.2 VIDEO GENERATION MODEL

Following Sora (Brooks et al., 2024), we adopt the Variational Auto-Encoder (VAE) and DiT architectures for video generation. The VAE compresses videos into latent representations both spatially and temporally, while the DiT models the denoising process. This approach demonstrates strong scalability and achieves promising results in generating high-quality videos.

VAE Model. We employ a (2+1)D-VAE to project videos into a latent space. Starting with the SD1.5-VAE structure, we extend it into a spatiotemporal autoencoder using 3D blocks (Yu et al., 2023b). All parameters of the (2+1)D-VAE are pretrained on high-quality image and video data to maintain strong appearance modeling while enabling motion modeling. More details are provided in Appendix A.3.1. In this paper, we *fix* the pretrained VAE encoder and use it as a video compressor. Results in Appendix A.3.2 confirm the VAE's ability to accurately encode and decode the physical event videos. This allows us to focus solely on training the diffusion model to learn the physical laws.

Diffusion model. Given the compressed latent representation from the VAE model, we flatten it into a sequence of spacetime patches, as transformer tokens. Notably, self-attention is applied to the entire spatio-temporal sequence of video tokens, without distinguishing between spatial and temporal dimensions. For positional embedding, a 3D variant of RoPE (Su et al., 2024) is adopted. As stated in Sec.2.1, our video model is conditioned on the first c frames. The c-frame video is zero-padded to the same length as the full physical video. We also introduce a binary mask "video" by setting the value of the first c frames to 1, indicating those frames are the condition inputs. The noise, condition and mask videos are concatenated along the channel dimension to form the final input to the model.

2.3 ON THE VERIFICATION OF LEARNED LAWS

Suppose we have a video generation model learned based on the above formulation. How do we determine if the underlying physical law has been discovered? A well-established law describes the behavior of the natural world, *e.g.*, how objects move and interact. Therefore, a video model incorporating true physical laws should be able to withstand experimental verification, producing reasonable predictions under any circumstances, which demonstrates the model's generalization ability. To comprehensively evaluate this, we consider the following categorization of generalization (see Figure 1) within the scope of this paper: 1) **In-distribution** (ID) generalization describes the

setting where training data and testing data are from the same distribution. In our case, both training and testing data follow the same law and are located in the same domain. 2) A human who has learned a physical law can easily extrapolate to scenarios that have never been observed before. This ability is referred to as **out-of-distribution** (OOD) generalization. Although it sounds challenging, this evaluation is necessary as it indicates whether a model can learn principled rules from data. 3) Moreover, there is a situation between ID and OOD, which has more practical value. We call this **combinatorial** generalization, representing scenarios where every "concept" or object has been observed during training, but not their every combination. It examines a model's ability to effectively combine relevant information from past experiences in novel ways. A similar concept has been explored in LLMs (Riveland & Pouget, 2024), which demonstrated that models can excel at linguistic instructing tasks by recombining previously learned components, without task-specific experience.

3 In-Distribution and Out-of-Distribution Generalization

In this section, we study how in-distribution and out-of-distribution generalization is correlated with model or data scaling. We focus on deterministic tasks governed by basic kinematic equations, as they allow clear definitions of ID/OOD and straightforward quantitative error evaluation.

3.1 Fundamental Physical Scenarios

Specifically, we consider three physical scenarios illustrated in Fig. 2. 1) **Uniform Linear Motion:** A colored ball moves horizontally with a constant velocity. This is used to illustrate *the law of Intertia*. 2) **Perfectly Elastic Collision:** Two balls with different sizes and speeds move horizontally toward each other and collide. The underlying physical law is *the conservation of energy and momentum*. 3) **Parabolic Motion:** A ball with a initial horizontal velocity falls due to gravity. This represents *Newton's second law of motion*. Each motion is determined by its initial frames.

Training data generation. We use Box2D to simulate kinematic states for various scenarios and render them as videos, with each scenario having 2-4 degrees of freedom (DoF), such as the balls' initial velocity and mass. An in-distribution range is defined for each DoF. We generate training datasets of 30K, 300K, and 3M videos by uniformly sampling a high-dimensional grid within these ranges. All balls have the same density, so their mass is inferred from their size. Gravitational acceleration is constant in parabolic motion for consistency. Initial ball positions are randomly initialized within the visible range. Further details are provided in Appendix A.4.

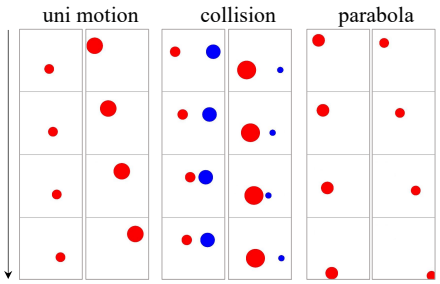

Figure 2: Downsampled video visualization. The arrow indicates the progression of time.

Test data generation. We evaluate the trained model using both ID and OOD data. For ID evaluation, we sample from the same grid used during training, ensuring that no specific data point is part of the training set. OOD evaluation videos are generated with initial radius and velocity values outside the training range. There are various types of OOD setting, *e.g.* velocity/radius-only or both OOD. Details are provided in Appendix A.4.

Table 1: Details of DiT model sizes.

Model	Layers	Hidden size	Heads	#Param
DiT-S	12	384	6	22.5M
DiT-B	12	768	12	89.5M
DiT-L	24	1024	16	310.0M
DiT-XL	28	1152	16	456.0M

Models. For each scenario, we train models of varying sizes from scratch, as shown in Table 1. This ensures that the outcomes are not influenced by uncontrollable pretrain data. The first three frames are provided as conditioning, which is sufficient to infer the velocity of the balls and predict the subsequent frames. Diffusion model is trained for 100K steps using 32 Nvidia A100 GPUs with a batch size of 256, which was sufficient for convergence, as a model trained for 300K steps achieves a similar performance. We keep the pretrained VAE fixed. Each video consists of 32 frames with a resolution of 128x128. We also experimented with a 256x256 resolution, which yielded a similar generalization error but significantly slowed down the training process.

Evaluation metrics. We observed that the learned models are able to generate balls with consistent shapes. To obtain the center positions of the i-th ball in the generated videos, x_t^i, we use a heuristic

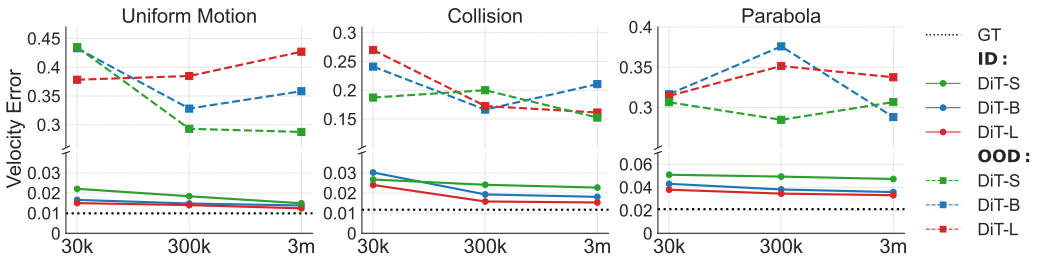

Figure 3: The error in the velocity of balls between the ground truth state in the simulator and the values parsed from the generated video by the diffusion model, given the first 3 frames.

algorithm based on the mean of colored pixels, distinguishing the balls by color. To ensure the correctness of x_t^i, we exclude frames with part of a ball out of view, yielding valid frames T. For collision scenarios, only the frames after the collision are considered. We then compute the velocity of each ball, v_t^i, at each moment by differentiating their positions. The *error* for a video is defined as: $e = \frac{1}{N|T|}\sum_{i=1}^{N}\sum_{t\in T}\left|v_t^i - \hat{v}_t^i\right|$, where v_t^i is the computed velocity at time t, $\hat{v}_t^i$ is the ground-truth velocity in the simulator, N is the number of balls, and $|T|$ is the number of valid frames.

Baseline. We calculate the error between the ground truth velocity and the parsed values from the ground truth video, referred to as *Groundtruth* (GT). This represents the system error—caused by parsing video into velocity—and defines the minimum error a model can achieve.

3.2 Main Result of Scaling Data and Model

For **in-distribution** (ID) generalization in Figure 3, increasing the model size (DiT-S to DiT-L) or the data amount (30K to 3M) consistently decreases the *velocity error* across all three tasks, strongly evidencing the importance of scaling for ID generalization. Take the uniform motion task as an example: the DiT-S model has a velocity error of 0.022 with 30K data, while DiT-L achieves an error of 0.012 with 3M data, very close to the error of 0.010 obtained with ground truth video.

However, the results differ significantly for **out-of-distribution** (OOD) predictions. First, OOD velocity errors are an order of magnitude higher than ID errors in all settings. For example, the OOD error for the DiT-L model on uniform motion with 3M data is 0.427, while the ID error is just 0.012. Second, scaling up the training data and model size has little or negative impact on reducing this prediction error. The variation in velocity error is higly random as data or model size changes, *e.g.*, the error for DiT-B on uniform motion is 0.433, 0.328 and 0.358, with data amounts of 30K, 300K and 3M. We also trained DiT-XL on the uniform motion 3M dataset but observed no improvement in OOD generalization. As a result, we did not pursue training of DiT-XL on other scenarios or datasets constrained by resources. These findings suggest the inability of scaling to perform reasoning in OOD scenarios. The sharp difference between ID and OOD settings further motivates us to study the generalization mechanism of video generation in Section 5.2.

4 Combinatorial Generalization

In Section 3, video generation models failed to reason in OOD scenarios. This is understandable—deriving precise physical laws from data is difficult for both humans and models. For example, it took scientists centuries to formulate Newton's three laws of motion. However, even a child can intuitively predict outcomes in everyday situations by combining elements from past experiences. This ability to combine known information to predict new scenarios is called *combinatorial generalization*. In this section, we evaluate the combinatorial abilities of diffusion-based video models.

4.1 Combinatorial Physical Scenarios

We selected the PHYRE simulator (Bakhtin et al., 2019) as our testbed—a 2D environment involves multiple objects to free fall then collide with each other, forming complex physical interactions. It features diverse object types, including balls, jars, bars, and walls, which can be either fixed or dynamic. This enables complex interactions such as collisions, parabolic trajectories, rotations, and

friction to occur simultaneously within a video. Despite this complexity, the underlying physical laws are deterministic, allowing the model to learn the laws and predict unseen scenarios.

Training Data. There are eight types of objects considered, including two dynamic gray balls, a group of fixed black balls, a fixed black bar, a dynamic bar, a group of dynamic standing bars, a dynamic jar, and a dynamic standing stick. Each task contains one red ball and four randomly chonsen objects from the eight types, resulting in $C_8^4 = 70$ unique templates. See Figure 4 for examples.

Each template was initialized with random sizes and positions for four objects, generating 100K videos to cover a range of possible scenarios. To explore the model's combinatorial ability and scaling effects, we structured the training data at three levels: a minimal set of 6 templates (0.6M videos) that includes all types of two-object interactions among the eight object types, and larger sets with 30 and 60 templates (3M/6M videos), with the 60-template set nearly covering the entire template space. The minimal training set places the highest demand on the model's ability for compositional generalization.

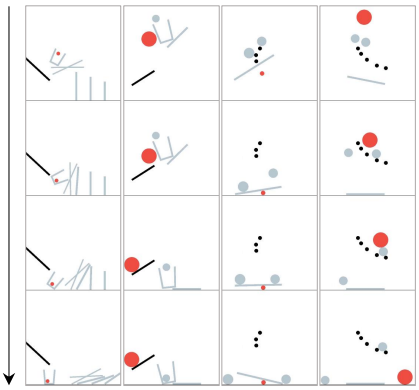

Figure 4: Downsampled videos. The black objects are fixed and others are dynamic.

Test Data. For each training template, we reserve a small set of videos to create the *in-template* evaluation set. Additionally, 10 unused templates are reserved for the *out-of-template* evaluation set to assess the model's ability to generalize to new combinations not seen during training.

Models. The first frame is used as the conditioning for video generation since the inital objects are static. We found that smaller models like DiT-S struggled with complex videos, so we primarily used DiT-B and DiT-XL. All models were trained for long 1000K gradient steps on 64 Nvidia A100 GPUs with a batch size of 256, ensuring near convergence. To better capture the complexity of physical events, we increased the resolution to 256x256 with 32 frames.

Evaluation Metrics. We use several metrics to assess the fidelity of generated videos compared to the ground truth. Frechet Video Distance (FVD) (Unterthiner et al., 2018) calculates feature distances between generated and real videos using features from Inflated-3D ConvNets (I3D) pretrained on Kinetics-400 (Carreira & Zisserman, 2017). SSIM and PSNR (Wang et al., 2004) are pixel-level metrics: SSIM evaluates brightness, contrast, and structural similarity, while PSNR measures the ratio between peak signal and mean squared error, both averaged across frames. LPIPS (Zhang et al., 2018) gauges perceptual similarity between image patches. We include human evaluations, reporting the *abnormal ratio* of generated videos that violate physical laws assessed by humans.

4.2 Main Results

Table 2: Combinatorial generalization results.
The results are presented in the format of {*in-template result*} / {*out-of-template result*}.

Model	#Templates	FVD (↓)	SSIM (↑)	PSNR (↑)	LPIPS (↓)	Abnormal (↓)
DiT-XL	6	18.2 / 22.1	**0.973** / 0.943	**32.8** / 25.5	**0.028** / 0.082	3% / 67%
DiT-XL	30	19.5 / 20.7	0.971 / 0.948	32.0 / 26.7	0.031 / 0.068	4% / 18%
DiT-XL	60	**17.6** / **18.7**	0.972 / **0.951**	32.4 / **27.3**	0.030 / **0.062**	**2%** / **10%**
DiT-B	60	18.4 / 21.4	0.967 / 0.949	30.9 / 27.0	0.035 / 0.066	3% / 24%

It requires higher resolution, much more training iterations, and larger model sizes to perform well on this task due to increased complexity. Consequently, we are unable to conduct a comprehensive sweep of all data and model size combinations as in Section 3. Therefore, we start with the largest model, DiT-XL, to study data scaling behavior for combinatorial generalization. As shown in Table 2, when the number of templates increases from 6 to 60, all metrics improve on the out-of-template testing sets. Notably, the abnormal rate for human evaluation significantly reduces from 67% to 10%. Conversely, the model trained with 6 templates achieves the best SSIM, PSNR, and LPIPS scores on the in-template testing set. This can be explained by the fact that each training example in the 6-template set is exposed ten times more frequently than those in the 60-template set, allowing it to

better fit the in-template tasks associated with template 6. Furthermore, we conducted an additional experiment using a DiT-B model on the full 60 templates to verify the importance of model scaling. As expected, the abnormal rate increases to 24%. These results suggest that both model capacity and coverage of the combination space are crucial for combinatorial generalization. This insight implies that scaling laws for video generation should focus on increasing combination diversity, rather than merely scaling up data volume. The video generation visualizations of our models can be found in Figure 17 and Figure 18.

5 Deeper Analysis

In this section, we aim to investigate the generalization mechanism of a video generation model, through systemic experimental designs. Based on the findings, we try to identify certain patterns in combinatorial generalization that might be helpful in harnessing or prompting the models.

5.1 Understanding Generalization from Interpolation and Extrapolation

The generalization ability of a model roots from its interpolation and extrapolation capability (Xu et al., 2020; Balestriero et al., 2021). In this section, we design experiments to explore the limits of these abilities for a video generation model. We design datasets which delibrately leave out some latent values, i.e. velocity. After training, we test model's prediction on both seen and unseen scenarios. We mainly focus on uniform motion and collision processes.

Uniform Motion. We create a series of training sets, where a certain range of velocity is absent. Each set contains 200K videos to ensure fairness. As shown in Figure 5 (1)-(2), with a large gap in the training set, the model tends to generate videos where the velocity is either high or low to resemble training data when initial frames show middle-range velocities. We find video generation model's OOD accuracy is closely related to the size of gap, as seen in Figure 5 (3), when the gap is reduced, the model correctly interpolates for most of OOD data. Moreover, as shown in Figure 5 (4) and (5), when a subset of the missing range is reintroduced (without increasing data amount), the model exhibits stronger interpolation abilities.

Collision involves multiple variables, which is more challenging since the model has to learn a two-dimensional non-linear function. Specifically, we exclude one or more square regions from the training set of initial velocities for two balls and then assess the velocity prediction error after the collision. For each velocity point, we sample a grid of radius parameters to generate multiple video cases and compute the average error. As shown in Figure 6 (1)-(2), an interesting phenomenon happens. The video generation model's extrapolation error demonstate an intringuing discrepancy among OOD points: For the OOD velocity combinations that lie within the convex hull of the training set, *i.e.*, the internal red squares in the yellow region, the model generalizes well. However, the model experiences large errors when the latent values lies in exterior space of training set's convex hull.

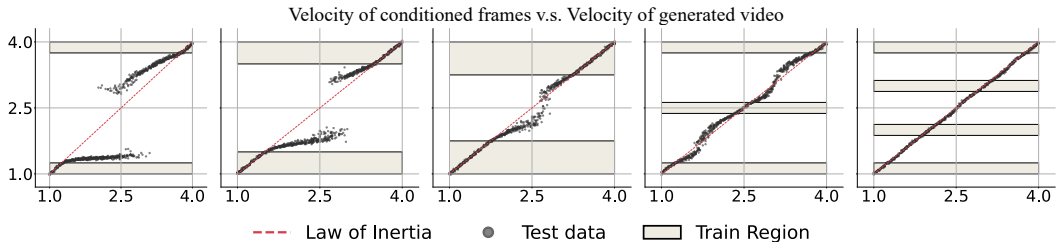

Figure 5: Uniform motion video generation. Models are trained on datasets with a missing middle velocity range. For example, in the first figure, training velocities cover $[1.0, 1.25]$ and $[3.75, 4.0]$, excluding the middle range. When evaluated with velocity condition from the missing range $[1.25, 3.75]$, the generated velocity tends to shift away from the initial condition, breaking the Law of Inertia.

5.2 Memorization or Generalization

Previous work (Hu et al., 2024) indicates that LLMs rely on memorization, reproducing training cases during inference instead of learning the underlying rules for tasks like addition arithmetic. In

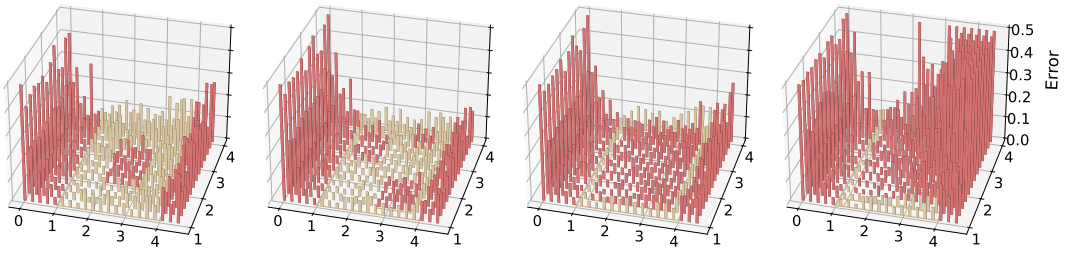

Figure 6: Collision video generation. Models are trained on the yellow region and evaluated on data points in both the yellow (ID) and red (OOD) regions. When the OOD range is surrounded by the training region, the OOD generalization error remains relatively small and comparable to the ID error.

this section, we investigate whether video generation models display similar behavior, memorizing data rather than understanding physical laws, which limits their generalization to unseen data.

We train our model on uniform motion videos with velocities $v \in [2.5, 4.0]$, using the first three frames as input conditions. Two training sets are used: *Set-1* only contains balls moving from left to right, while *Set-2* includes movement in both direction, by using horizontal flipping at training time. At evaluation, we focus on low-speed balls ($v \in [1.0, 2.5]$), which were not present in the training data. As shown in Figure 7, the *Set-1* model generates videos with only positive velocities, biased toward the high-speed range. In contrast, the *Set-2* model occasionally produces videos with negative velocities, as highlighted by the green circle. For instance, a low-speed ball moving from left to right may suddenly reverse direction after its condition frames. This could occur since the model identifies reversed training videos as the closest match for low-speed balls. This distinction between the two models suggests that the video generation model is influenced by "deceptive" examples in the training data. Rather than abstracting universal rules, the model appears to rely on memorization, and case-based imitation for OOD generalization.

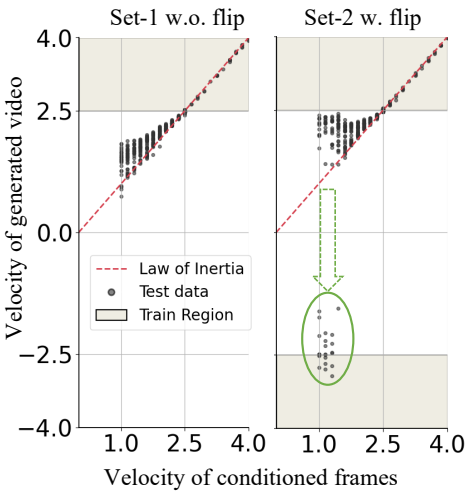

Figure 7: The example of uniform motion illustrating memorization.

5.3 How Does Diffusion Model Retrieve Data?

We aim to investigate the ways a video model performs *case matching*—identifying close training examples for a given test input. We use *uniform linear motion* for this study. Specifically, we compare four attributes, *i.e.*, color, shape, size, and velocity, in a pairwise manner. Through comparisons, we seek to determine the model's preference for relying on specific attributes in case matching.

Every attribute has two disjoint sets of values. For each pair of attributes, there are four types of combinations. We use two combinations for training and the remaining two for testing. For example, we compare color and shape in Figure 8 (1). Videos of red balls and blue squares with the same range of size and velocity are used for training. At test time, a blue ball changes shape into a square immediately after the condition frames, while a red square transforms into a ball. We observed no exceptions on 1,400 test cases, showing that the model prioritizes color over shape for case matching. A similar trend is observed in the comparisons of size vs. shape and velocity vs. shape, as illustrated in Figure 8 (2)-(3), indicating that shape is the least prioritized attribute. This suggests that diffusion-based video models inherently favor other attributes over shape, which may explain why current open-set video generation models usually struggle with shape preservation.

The other three pairs are presented in Figure 9. For velocity *vs.* size, the combinatorial generalization performance is surprisingly good. The model effectively maintains the initial size and velocity for most test cases beyond the training distribution. However, a slight preference for size over velocity is noted, particularly with extreme radius and velocity values (top left and bottom right in Figure 9 (1)).

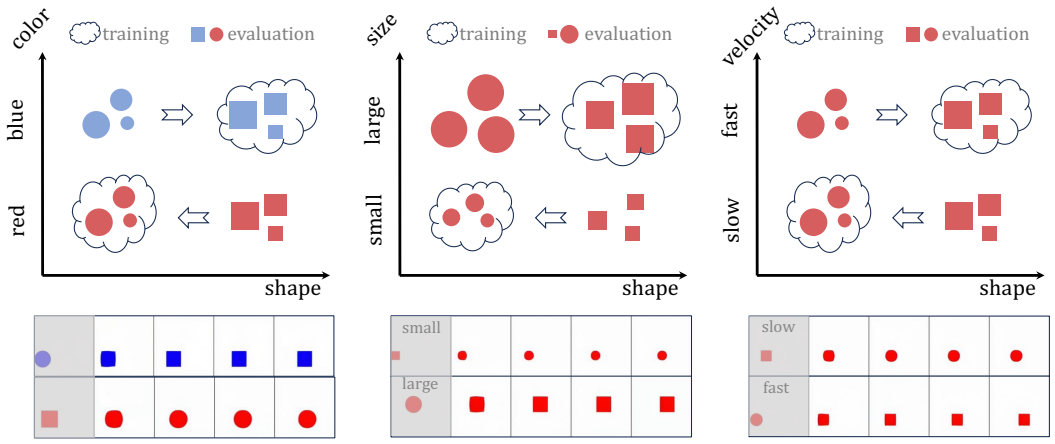

Figure 8: Uniform motion. (1) Color *v.s.* shape, (2) Size *v.s.* shape, (3) Velocity *v.s.* shape. The arrow $\Rightarrow$ signifies that the generated videos shift from their specified conditions to resemble similar training cases. For example, in the first figure, the model is trained on videos of blue balls and red squares. When conditioned with a blue ball, as shown in the bottom, it transforms into a blue square, i.e., mimicking the training case by color.

In Figure 9 (2), color can be combined with size most of the time. Conversely, for color *vs.* velocity in Figure 9 (3), high-speed blue balls and low-speed red balls are used for training. At test time, low-speed blue balls appear much faster than their conditioned velocity. No ball in the testing set changes its color, indicating that color is prioritized over velocity. Based on the above analysis, we conclude that the prioritization order is as follows: color > size > velocity > shape.

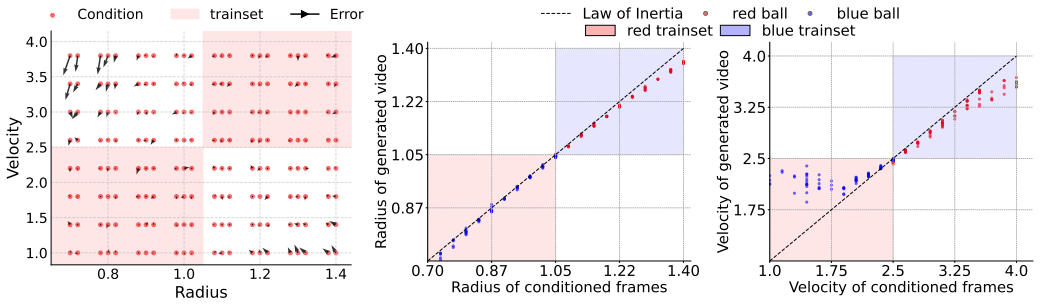

Figure 9: Uniform motion. (1) Velocity *v.s.* size: The arrow $\rightarrow$ indicates the direction of generated videos shifting from their initial conditions. (2) Color *v.s.* size: Models are trained with small red balls and large blue balls, and evaluated on reversed color-size pair conditions. All generated videos retain the initial color but show slight size shifts from the original. (3) Color *v.s.* velocity: Models are trained with low-speed red balls and high-speed blue balls, and evaluated on reversed color-velocity pair conditions. All generated videos retain the initial color but show large velocity shifts from the original.

5.4 How Does Complex Combinatorial Generalization Happen?

In Section 4, we show that scaling the data coverage can boost combinatorial generalization. But what kind of data can actually enable conceptually-combinable video generation? In this section, we idenfify three foundamental combinatorial patterns through experimental design.

Attribute composition. As shown in Figure 9 (1)-(2), certain attribute pairs—such as velocity and size, or color and size—exhibit some degree of combinatorial generalization.

Spatial composition. As given by Figure 11 (left side) in the appendix, the training data contains two distinct types of physical events. One type involves a blue square moving horizontally with a constant velocity while a red ball remains stationary. In contrast, the other type depicts a red ball moving toward and then bouncing off a wall while the blue square remains stationary. At test time, when the red ball and the blue square are moving simultaneously, the learned model is able to generate the scenario where the red ball bounces off the wall while the blue square continues its uniform motion.

Temporal combination. As illustrated on the right side of Figure 11, when the training data includes distinct physical events—half featuring two balls colliding without bouncing and the other half showing a red ball bouncing off a wall—the model learns to combine these events temporally. Consequently, during evaluation, when the balls collide near the wall, the model accurately predicts the collision and then determines that the blue ball will rebound off the wall with unchanged velocity.

With these attribute, spatial, and temporal combinatorial patterns, the video generation model can identify basic physical events in the training set and combine them across attributes, time, and space to generate videos featuring complex chains of physical events.

5.5 Is Video Sufficient for Complete Physics Modeling?

For a video generation model to function as a world model, the visual representation must provide sufficient information for complete physics modeling. In our experiments, we found that visual ambiguity leads to significant inaccuracies in fine-grained physics modeling. For example, in Figure 10, it is difficult to determine if a ball can pass through a gap based on vision alone when the size difference is at the pixel level, leading to visually plausible but incorrect results. Similarly, visual ambiguity in a ball's horizontal position relative to a block can result in different outcomes. These findings suggest that relying solely on visual representations, may be inadequate for accurate physics modeling.

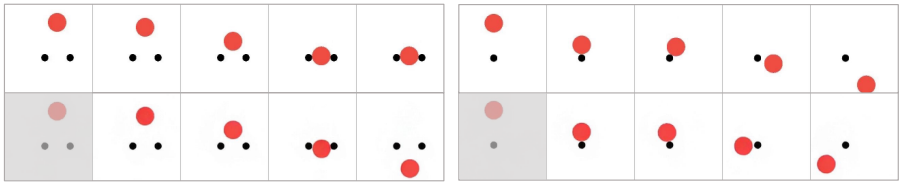

Figure 10: First row: Ground truth; second row: generated video. Ambiguities in visual representation result in inaccuracies in fine-grained physics modeling.

6 Related Works

Video generation. Open-set video generation is mostly based on diffusion models (Rombach et al., 2022; Ho et al., 2022b;a; He et al., 2024) or auto-regressive models (Yu et al., 2023a;b; Kondratyuk et al., 2023). These models often require a pretrained image or video VAE (Kingma, 2013; Van Den Oord et al., 2017) for data compression to improve computational efficiency. Some approaches leverage pretrained Text-to-Image (T2I) models for zero-shot (Khachatryan et al., 2023; Zhang et al., 2023b) or few-shot (Wu et al., 2023) video generation. Additionally, Image-to-Video(I2V) generation (Zeng et al., 2024; Girdhar et al., 2023; Zhang et al., 2023a) shows that video quality improves substantially when conditioned on an image. The Diffusion Transformer (DiT) (Peebles & Xie, 2023) demonstrates better scaling behavior than U-Net (Ronneberger et al., 2015) for T2I generation. Sora (Brooks et al., 2024) leverages the DiT architecture to directly operate on spacetime patches of video and image latent codes. Our model follows Sora's architecture and conceptually aligns with I2V generation, relying on image(s) for conditioning instead of text prompts.

World model. World models (Ha & Schmidhuber, 2018) aim to learn models that can accurately predict how an environment evolves after some actions are taken. Previously, they often operated in an abstracted space and were used in reinforcement learning (Sutton, 2018) to enable planning (Silver et al., 2016; Schrittwieser et al., 2020) or facilitate policy learning through virtual interactions (Hafner et al., 2019; 2020). With the advancement of generative models, world models can now directly work with visual observations by employing a general framework of conditioned video generation. For example, in autonomous driving (Hu et al., 2023; Jia et al., 2023; Gao et al., 2024; Zheng et al., 2024), the condition is the driver's operations, while in robot world models (Yang et al., 2023; Black et al., 2023; 1xw, 2024), the condition is often the control signals. Genie (Bruce et al., 2024) instead recovers the conditions from video games in an unsupervised learning manner. In our physical law discovery setting, it does not require a per-step action/condition since the physical event is determined by the underlying laws once an initial state is specified. See more related works in Appendix A.1.

7 CONCLUSION

Video generation is believed as a promising way towards scalable world models. However, its capability to learn physical laws from visual observations has not yet been verified. We conducted the first systematic study in this area by examining its generalization performance across three typical scenarios: in-distribution, out-of-distribution (OOD), and combinatorial generalization. The findings indicate that scaling alone cannot address the OOD problem, although it does enhance performance in other scenarios. Our in-depth analysis suggests that video model generalization relies more on referencing similar training examples rather than learning universal rules. We observed a prioritization order of color > size > velocity > shape in this "case-based" behavior. In conclusion, our study suggests that naively scaling is insufficient for video generation models to discover fundamental physical laws.

REFERENCES

1x world model. 2024. URL https://www.1x.tech/discover/1x-world-model. 10

Kelsey R Allen, Kevin A Smith, and Joshua B Tenenbaum. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, 117(47):29302–29310, 2020. 17

Tayfun Ates, M Samil Atesoglu, Cagatay Yigit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. Craft: A benchmark for causal reasoning about forces and interactions. *arXiv preprint arXiv:2012.04293*, 2020. 17

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32, 2019. 5, 17

Randall Balestriero, Jerome Pesenti, and Yann LeCun. Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485*, 2021. 7

Romain Beaumont and Christoph Schuhmann. Laion-aesthetics v1. https://github.com/LAION-AI/laion-datasets/blob/main/laion-aesthetic.md, 2022. 18

Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3479–3487, 2015. 17

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023. 10

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

Katherine L Bouman, Bei Xiao, Peter Battaglia, and William T Freeman. Estimating the material properties of fabric from video. In *Proceedings of the IEEE international conference on computer vision*, pp. 1984–1991, 2013. 17

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. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators. 1, 3, 10

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 10

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 18

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017. 6

Brian M de Silva, David M Higdon, Steven L Brunton, and J Nathan Kutz. Discovery of physics from data: Universal laws and discrepancies. *Frontiers in artificial intelligence*, 3:25, 2020. 17

Yilun Du and Leslie Kaelbling. Compositional generative modeling: A single model is not all you need. *arXiv preprint arXiv:2402.01103*, 2024. 2

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alexandros G. Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. *ArXiv*, abs/2304.14108, 2023. URL https://api.semanticscholar.org/CorpusID:258352812. 18

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024. 10

Rohit Girdhar, Laura Gustafson, Aaron Adcock, and Laurens van der Maaten. Forward prediction for physical reasoning. *arXiv preprint arXiv:2006.10734*, 2020. 17

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 10

Oliver Groth, Fabian B Fuchs, Ingmar Posner, and Andrea Vedaldi. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *Proceedings of the european conference on computer vision (eccv)*, pp. 702–717, 2018. 17

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018. 10

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019. 10

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020. 10

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 19

Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *arXiv preprint arXiv:2406.11138*, 2024. 10

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

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a. 10

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022b. 10

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 1, 10

Yi Hu, Xiaojuan Tang, Haotong Yang, and Muhan Zhang. Case-based or rule-based: How do transformers do the math? *ICML*, 2024. 2, 7

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2016. URL https://api.semanticscholar.org/CorpusID:6200260. 18

Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023. 10

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023. 10

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 10

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 10

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5404–5411, 2024. 18

Andrew Melnik, Robin Schiewer, Moritz Lange, Andrei Muresanu, Mozhgan Saeidi, Animesh Garg, and Helge Ritter. Benchmarks for physical reasoning ai. *arXiv preprint arXiv:2312.10728*, 2023. 17

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. 10

Reidar Riveland and Alexandre Pouget. Natural language instructions induce compositional generalization in networks of neurons. *Nature Neuroscience*, 27(5):988–999, 2024. 4, 19

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 10

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015. 10

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 18

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020. 10

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016. 10

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 17

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3

Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018. 10

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 10

Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Swap attention in spatiotemporal diffusions for text-to-video generation. 2023. URL https://api.semanticscholar.org/CorpusID:258762479. 18

Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, and Lu Fang. Panda: A gigapixel-level human-centric video dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3265–3275, 2020. doi: 10.1109/CVPR42600.2020.00333. 18

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

Erik Weitnauer, Robert L Goldstone, and Helge Ritter. Perception and simulation during concept learning. *Psychological Review*, 2023. 17

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633, 2023. 10

Jiajun Wu, Joseph J Lim, Hongyi Zhang, Joshua B Tenenbaum, and William T Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, volume 2, pp. 7, 2016. 17

Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848*, 2020. 7

Cheng Xue, Vimukthini Pinto, Chathura Gamage, Ekaterina Nikonova, Peng Zhang, and Jochen Renz. Phy-q as a measure for physical reasoning intelligence. *Nature Machine Intelligence*, 5(1): 83–93, 2023. 17

Tianfan Xue, Baian Chen, Jiajun Wu, D. Wei, and William T. Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, pp. 1–20, 2017. URL https://api.semanticscholar.org/CorpusID:40412298. 18

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 1, 10

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 17

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023a. 10

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023b. 3, 10, 18

Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8850–8860, 2024. 10

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. 6

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023a. 10

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023b. 10

Wenzhao Zheng, Ruiqi Song, Xianda Guo, and Long Chen. Genad: Generative end-to-end autonomous driving. *arXiv preprint arXiv:2402.11502*, 2024. 10

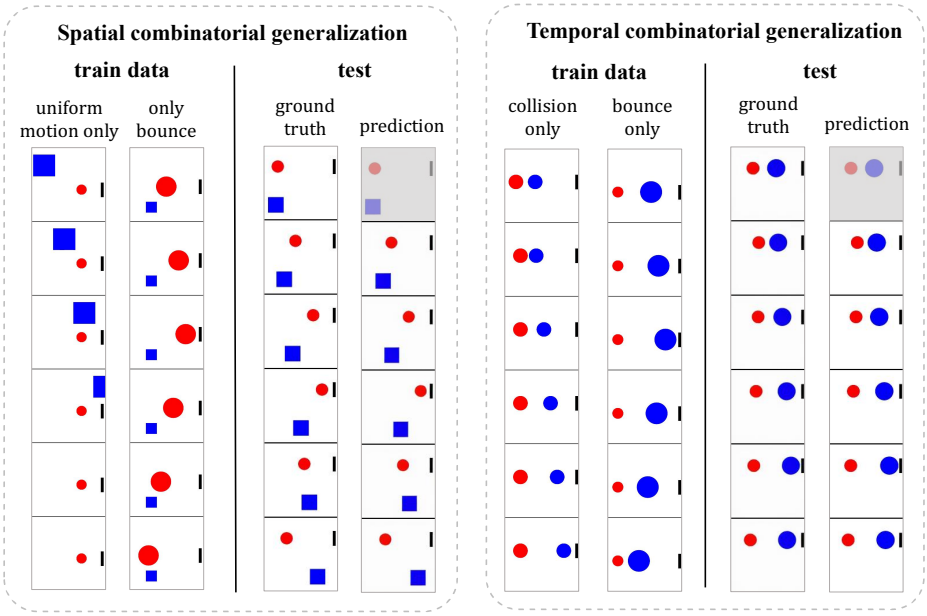

Figure 11: Spatial and temporal combinatorial generalization. The two subsets of the training set contain disjoint physical events. However, the trained model can combine these two types of events across spatial and temporal dimensions.

A Appendix

A.1 More related works

Physical reasoning. It refers to the ability to understand and predict the way objects will interact under certain conditions according to physical laws. Melnik et al. (2023) categorize physical reasoning tasks into *passive* and *interactive* tasks. Passive tasks often require the AI to predict certain properties of objects, *e.g.*, materials (Bell et al., 2015; Bouman et al., 2013), physical parameters (Wu et al., 2016), stability of a physical system (Groth et al., 2018), or involve question-answering for the agent to recognize conceptual differences (Weitnauer et al., 2023; Yi et al., 2019), or describe a physical scenario (Ates et al., 2020). For interactive tasks, AI is required to control some objects in the environment to complete certain tasks based on physical commonsense, *e.g.*, solving classical mechanics puzzles (Bakhtin et al., 2019), flying a bird to reach a target position (Xue et al., 2023), and using a tool (Allen et al., 2020). Two works closely related to ours are Girdhar et al. (2020) and de Silva et al. (2020). Girdhar et al. (2020) introduce a forward prediction model to aid physical reasoning but do not address what is learned in the prediction model. de Silva et al. (2020) attempt to discover universal physical laws from data with abstracted internal states and human expertise introduced in the dynamic model design. In contrast, we focus on recovering physical laws from raw observation without any human priors, akin to a newborn baby.

A.2 Diffusion preliminaries

Let $p(x)$ be the real data distribution. Diffusion models (Ho et al., 2020; Song et al., 2020) learn the data distribution by denoising samples from a noise distribution step-by-step. In this paper, we use the Gaussian diffusion models, where the video V is progressively corrupted by gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$ during the forward process, denoted by

$$V_t = \alpha_t V + \beta_t \boldsymbol{\epsilon}, \tag{1}$$

where α_t, β_t are the time-dependent noise scheduler. We use the original DDPM formulation (Ho et al., 2020), where $\alpha_t = \sqrt{\gamma_t}, \beta_t = \sqrt{1-\gamma_t}$, γ_t is a monotonically decreasing scheduler from 1 to 0. The diffusion models are trained to reverse the forward corruptions, denoted by

$$\mathbb{E}_{V\sim p(x), t\sim\mathcal{U}(0,\boldsymbol{I}), \boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I})}\left[\|\mathbf{y} - p_\theta(V_t, \mathbf{c}, t)\|^2\right], \tag{2}$$

where the target $\mathbf{y}$ can be the corrupted noise $\boldsymbol{\epsilon}$, the original video V, or the velocity between data and noise. Following Salimans & Ho (2022), we use the velocity prediction to train the video diffusion models, denoted by

$$\mathbf{y} = \sqrt{1-\gamma_t}\boldsymbol{\epsilon} - \sqrt{\gamma_t}V. \tag{3}$$

Also, to eliminate the training and inference gap and ensure a zero signal-to-noise ratio at the final timestep, following Lin et al. (2024), we set γ_t to 1 when $t = 1$.

A.3 VAE

A.3.1 VAE Architecture and Pretrain

We commence with the structure of the SD1.5-VAE, retaining the majority of the original 2D convolution, group normalization, and attention mechanisms on the spatial dimensions. To inflate this structure into a spatial-temporal auto-encoder, we convert the final few 2D downsample blocks of the encoder and the initial few 2D upsample blocks of the decoder into 3D ones, and employ multiple extra 1D layers to enhance temporal modeling. For all the downsample blocks where temporal downsampling is required, we replace all the original 2D downsample layers with re-initialized causal 3D downsample layers by adding causal paddings to the head of the frame sequence (Yu et al., 2023b) and introduce an additional causal 1D convolution layer after the original *ResNetBlock*. As for the decoder part, all the 2D *Nearest-Interpolation* operations are substituted by a 2D convolution layer and a channel-to-space transformation, which are specifically initialized to behave precisely the same as *Nearest-Interpolation* operations before the first training step. For the discriminator part, we inherit the structure of the original 2D PatchGAN (Isola et al., 2016) discriminator used by SD1.5-VAE, and design a 3D PatchGAN discriminator based on the 2D version. Different from the generator module, we train the group of discriminators from scratch for the consideration of stability. Subsequently, all parameters of the (2+1)D-VAE are jointly trained with high-quality image and video data to preserve the capability of appearance modeling and to enable motion modeling. For the image dataset, we filter data samples from LAION-Aesthetics (Beaumont & Schuhmann, 2022), COYO (Byeon et al., 2022) and DataComp (Gadre et al., 2023) with high aesthetics and clarity to form a high-quality subset. As for the video dataset, we collect a high-quality subset from Vimeo-90K (Xue et al., 2017), Panda-70M (Wang et al., 2020) and HDVG Wang et al. (2023). In the training process, we train the entire structure for 1M steps and only the random resized crop and random horizontal flip are applied in the data augmentation process.

A.3.2 VAE Reconstruction

In this paper, we *fix* the pretrained VAE encoder and use it as a video compressor. To verify the VAE's ability to accurately encode and decode the physical event videos, we evaluate its reconstruction performance. Specifically, we use the VAE to encode and decode (i.e., reconstruct) the ground truth videos and calculate the reconstruction error, e_{recon}. We then compare this error to the ground truth error, e_{gt}, as shown in Table 3. The results show that e_{recon} is very close to e_{gt}, and both are an order of magnitude lower than the OOD error, as illustrated in Figure 3. It confirms the pretrained VAE's ability to accurately encode and decode the physical event videos used in this paper.

Table 3: Comparison of errors for ground truth videos and VAE reconstruction videos.

Scenario	Ground Truth Error	VAE Reconstruction Error
Uniform Motion	0.0099	0.0105
Collision	0.0117	0.0131
Parabola	0.0210	0.0212

A.4 Fundamental Physical Scenarios Data

For the Box2D simulator, we initialize the world as a 10×10 grid, with a timestep of 0.1 seconds, resulting in a total time span of 3.2 seconds (32 frames). For all scenarios, we set the radius $r \in [0.7, 1.5]$ and velocity $v \in [1, 4]$ as in-distribution (in-dist) ranges. Out-of-distribution (OOD) ranges are defined as $r \in [0.3, 0.6] \cup [1.5, 2.0]$ and $v \in [0, 0.8] \cup [4.5, 6.0]$.

Collision Scenario: The four degrees of freedom (DoFs) are the masses of the two balls and their initial velocities, fully determining the collision outcomes. We generate 3k, 30k, and 3M training samples by sampling grid points from the 4-dimensional in-dist joint space of radii and velocities. For in-dist evaluation, we randomly sample about 2k points from the grid, ensuring they are not part of the training set. For OOD evaluation, we sample from the OOD ranges, generating approximately 4.8k samples across six OOD levels: (1) only r_1 OOD, (2) only v_1 OOD, (3) both r_1 and r_2 OOD, (4) both v_1 and v_2 OOD, (5) r_1 and v_1 OOD, and (6) r_1, v_1, r_2, and v_2 OOD. Additionally, for collisions, we ensure that all collisions occur after the 4th frame in each video, allowing the initial velocities of both balls to be inferred from the conditioned frames.

Uniform and Parabolic Motion: The two DoFs are the ball's mass and initial velocity. We generate 3k, 30k, and 3M training samples by sampling from the 2-dimensional in-dist joint space of radius and velocity. For in-dist evaluation, we sample approximately 1.05k for uniform motion and 1.1k for parabolic motion. For OOD evaluation, we generate about 2.4k (uniform motion) and 2.5k (parabolic motion) samples across three OOD levels: (1) only r OOD, (2) only v OOD, and (3) both r and v OOD.

For all scenarios, we filter out videos where the ball exits the field of view prematurely by inspecting the simulation state, as these videos do not provide sufficient meaningful information for the model.

A.5 More Experiments and Discussions

A.5.1 Can Language and Numerics Aid in Learning Physical Laws?

As discussed in Section 3, video generation based solely on image frames fails to learn physical laws, showing significant prediction errors in OOD scenarios, despite containing all the necessary information. In reinforcement learning, numerical values (*e.g.,* states) are often used as conditions for world models (Hafner et al., 2023), and language representations have shown generalization capabilities in LLMs (Riveland & Pouget, 2024). This raises the question: can additional multimodal inputs, such as numerics and text, improve video prediction and capture physical laws? We experimented with collision scenarios and DiT-B models, adding two variants: one conditioned on vision and numerics, and the other on vision and text. For numeric conditioning, we map the state vectors to embeddings and add the layer-wise features to video tokens. For text, we converted initial physical states into natural language descriptions, obtained text embeddings using a T5 encoder, and then add a cross-attention layer to aggregate textual representations for video tokens.

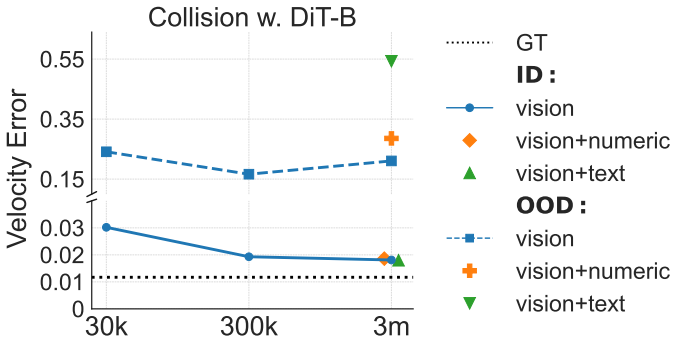

Figure 12: Comparison of different modal conditions for video generation.

As shown in Figure 12, for in-distribution generalization, adding numeric and text conditions resulted in prediction errors comparable to using vision alone. This suggests that visual frames already contain sufficient information for accurate predictions, and the additional numeric or text data do not provide further benefits. However, in OOD scenarios, the vision-plus-numerics condition exhibited slightly higher errors, while the vision-plus-language condition showed significantly higher errors. We hypothesize that the embeddings of language tokens and numerics may cause the model to overfit to specific patterns present in the training data, thereby impairing its ability to generalize to unseen OOD scenarios. Additionally, language tokens are discrete and exhibit greater variability compared to continuous numeric embeddings, making the model more susceptible to overfitting when

using language inputs, accounting for the higher OOD errors in the vision-plus-language condition compared to the vision-plus-numerics condition.

A.5.2 Principle Behind Data Retrieval in the Diffusion Model

As discussed in Section 5.3, the diffusion model appears to rely more on memorization and case-based imitation, rather than abstracting universal rules for OOD generalization. The model exhibits a preference for specific attributes during case matching, with a prioritization order of color > size > velocity > shape. In this section, we aim to explore the reasoning behind this prioritization.

Since the diffusion model is trained by minimizing the loss associated with predicting VAE latent, we hypothesize that the prioritization may be related to the distance in VAE latent space (though we use pixel space here for clearer illustration) between the test conditions and the training set. Intuitively, when comparing color and shape as in Figure 9 (1), a shape change from a ball to a rectangle results in minor pixel variation, primarily at the corners. In contrast, a color change from blue to red causes a more significant pixel difference. Thus, the model tends to preserve color while allowing shape to vary. From the perspective of pixel variation, the prioritization of color > size > velocity > shape can be explained by the extent of pixel change associated with each attribute. Changes in color typically result in large pixel variations because it affects nearly every pixel across its surface. In contrast, changes in size modify the number of pixels but do not drastically alter the individual pixels' values. Velocity affects pixel positions over time, leading to moderate variation as the object shifts, while shape changes often involve only localized pixel adjustments, such as at edges or corners. Therefore, the model prioritizes color because it causes the most significant pixel changes, while shape changes are less impactful in terms of pixel variation.

To further validate this hypothesis, we designed a variant experiment comparing color and shape, as shown in Figure 13. In this case, we use a blue ball and a red ring. For the ring to transform into the ball without changing color, it would need to remove the ring's external color, turning it into blank space, and then fill the internal blank space with the ball's color, resulting in significant pixel variation. Interestingly, in this scenario, unlike the previous experiments shown in Figure 9 (1), the prioritization of color > shape does not hold. The red ring can transform into either a red ball or a blue ring, as demonstrated by the examples. This observation suggests that the model's prioritization may indeed depend on the complexity of the pixel transformations required for each attribute change. Future work could explore more precise measurements of these variations in pixel or VAE latent space to better understand the model's training data retrieval process.

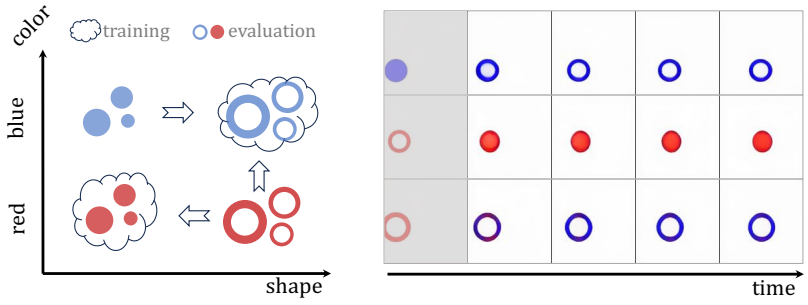

Figure 13: Uniform motion. Color vs. shape. The shapes are a ball and a ring. Transforming from a ring to a ball leads to a large pixel variation.

A.5.3 Failure Cases in Combinatorial Generalization

In Figure 11, we present successful cases of spatial and temporal combinatorial generalization, where the trained model can combine these two types of events across spatial and temporal dimensions to generate videos under unseen conditions. However, the model does not always succeed in performing such compositions, and here we illustrate some failure cases. As shown in Figure 14, when the training set lacks a red ball in a bounce event, the model *sometimes* struggles. In the model's generated video, the red ball sometimes disappears after a collision. This likely stems from the model's data

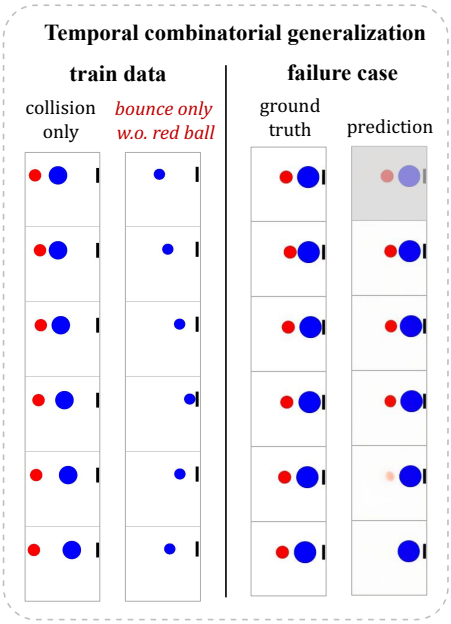

Figure 14: Failure cases in combinatorial generalization. Note that the bounce cases in the training set do not include the red ball.

retrieval mechanism: since the training set does not include a red ball in a collision scenario, when collision happens, the model retrieves similar training cases without the red ball, causing it to vanish post-collision. In summary, while combinatorial generalization allows the diffusion model to generate novel videos by composing spatial and temporal segments from the training set, its reliance on data retrieval limits its effectiveness. As a result, the model may produce unrealistic outcomes by retrieving and combining segments without understanding the underlying rules.

A.6 VISUALIZATION

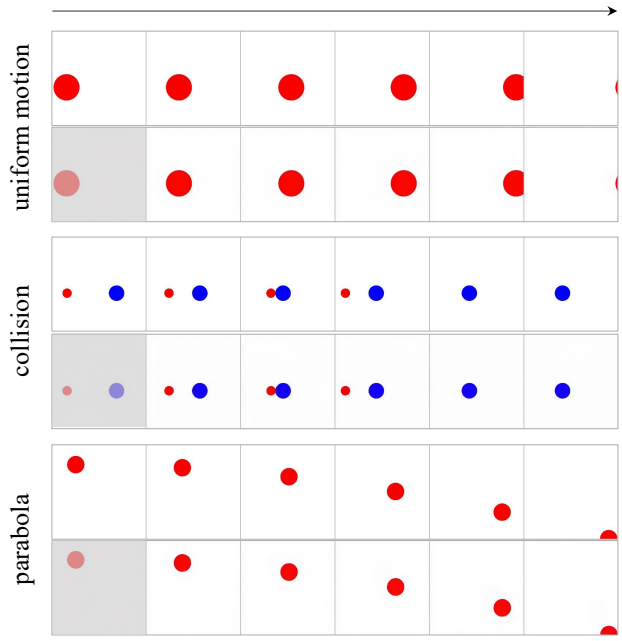

Figure 15: The visualization of *in-distribution evaluation* cases with very *small* prediction errors.

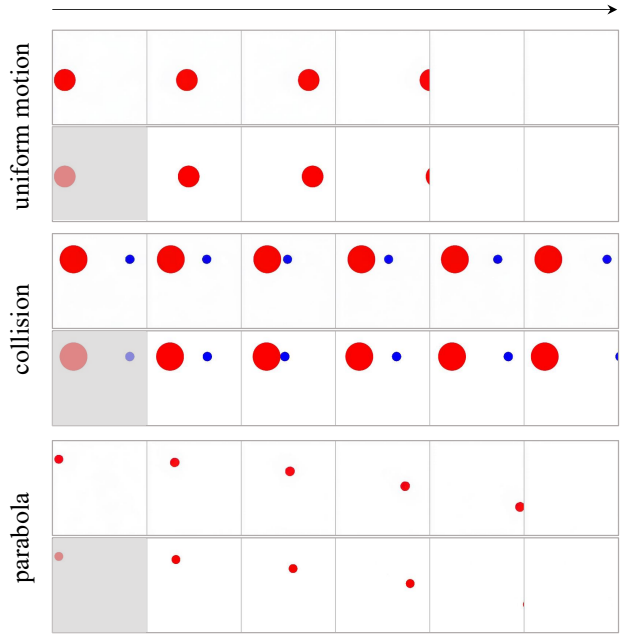

Figure 16: The visualization of *out-of-distribution evaluation* cases with *large* prediction errors.

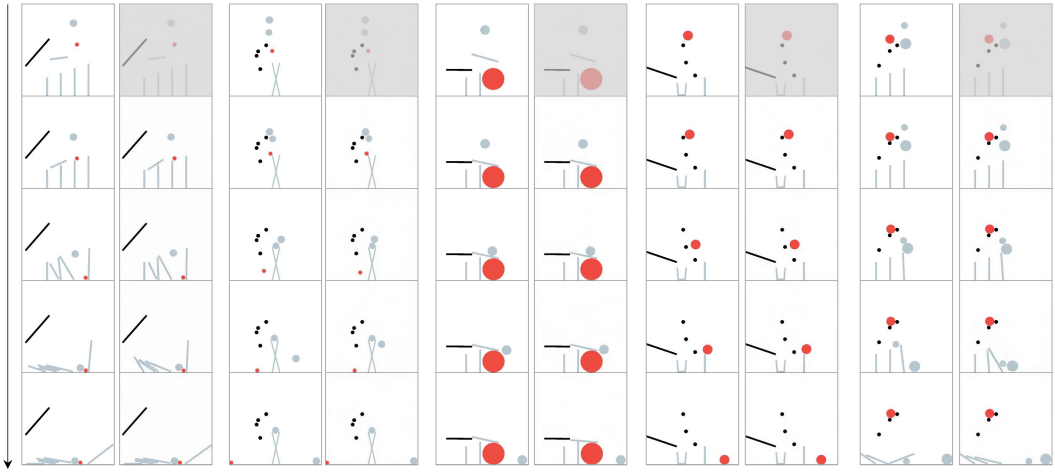

Figure 17: The visualization of *out-of-template evaluation* cases that appear *plausible and adhere to physical laws*, generated by DiT-XL trained on 6M data (60 templates). Zoom in for details. Notably, the first four cases generated by the model are nearly identical to the ground truth. In some cases, such as the rightmost example, the generated video seems physically plausible but differs from the ground truth due to visual ambiguity, as discussed in Section 5.5.

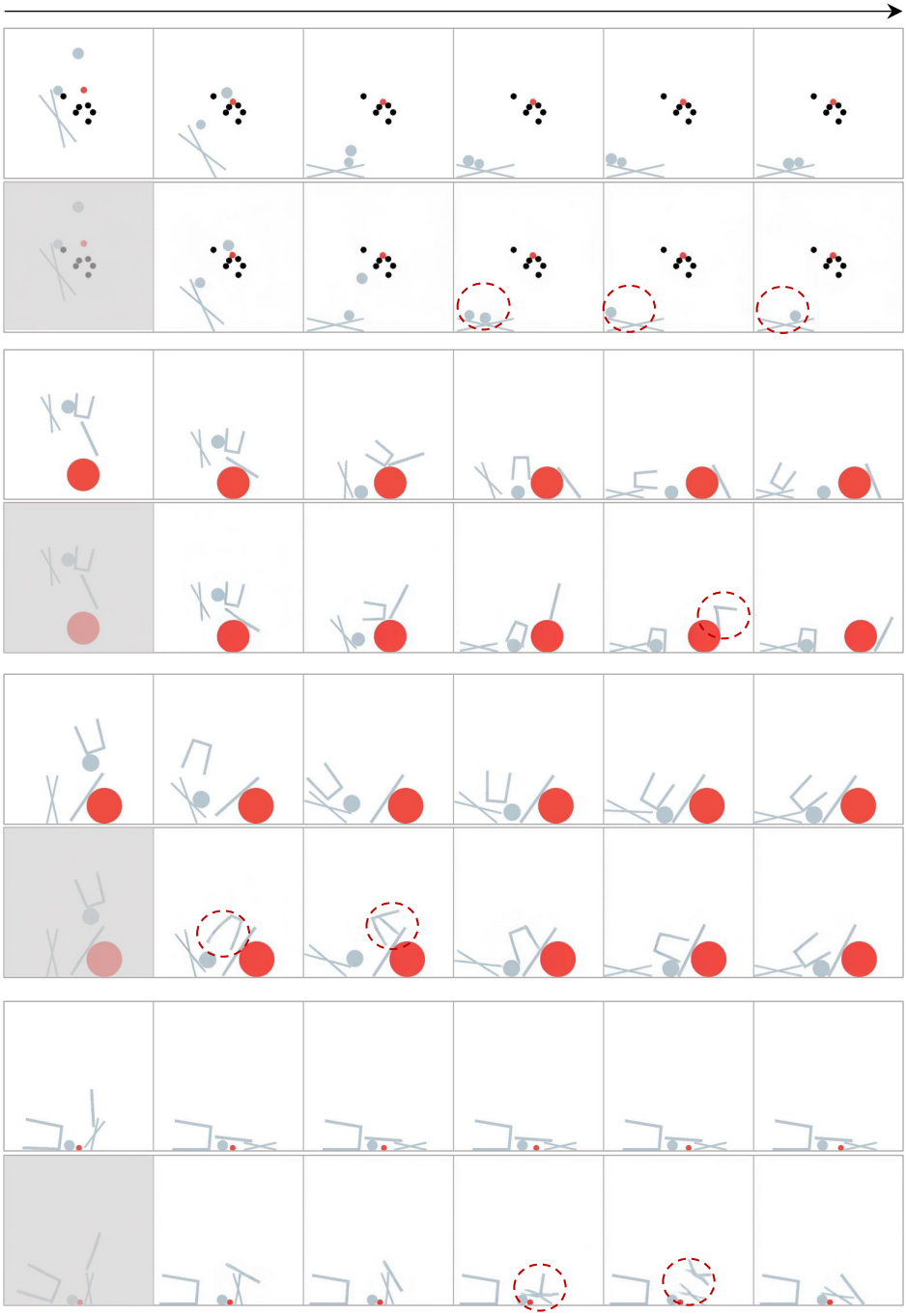

Figure 18: The visualization of *out-of-template evaluation* cases that appear *abnormal and violate physical laws*, generated by DiT-XL trained on 6M data (60 templates). Abnormalities are highlighted with red dotted circles. **Case 1**: A grey ball suddenly disappears. **Case 2**: The rigid-body bar breaks in several intermediate frames during contact with the ball, then recovers after contact. **Case 3**: The rigid-body jar fails to maintain its shape when interacting with the bar in several intermediate frames. **Case 4**: The rigid-body bar breaks in several intermediate frames during contact with the standing sticker. Most of the abnormal cases we observed involve object disappearance or shape inconsistencies, which can be explained by the case matching preference discussed in Section 5.3.