Bootstrapping Video Interaction Generation with Synthetic State Transitions

Jiho Jang[1], **Jin-Young Kim**[2], **Nojun Kwak**[1] and **Kyungjune Baek**[3]
[1]Seoul National University
[2]Independent Researcher
[3]Sejong University
{geographic, nojunk}@snu.ac.kr, seago0828@gmail.com, kyungjune.baek@sejong.ac.kr

Abstract

While recent video generative models can synthesize high-fidelity videos, they struggle to portray plausible physical interactions and the resulting state transitions, a critical bottleneck for applications in robotics and VR/AR. To address this, we introduce a framework to generate a scalable synthetic dataset of controllable interactions. Our pipeline leverages a structured taxonomy and state-of-the-art image editing models to create explicit 'start' and 'end' state images, which serve as visual anchors for the interaction. To generate a seamless video utilizing these anchors, we propose State-Guided Sampling (SGS), a novel sampling technique that mitigates artifacts common in naive conditional generation. Furthermore, we develop and validate a new automated evaluation system that aligns with human judgments to ensure data quality. Experiments show that fine-tuning a base model on our dataset significantly enhances its ability to generate plausible interactions.

1 Introduction

Recent advances in video generative models [Ho *et al.*, 2022; Kong *et al.*, 2024; Wan, 2025] have enabled the synthesis of high-fidelity videos. This capability has led to growing exploration of their use in applications, e.g., world models for robotics [Agarwal *et al.*, 2025] and developing immersive content for virtual and augmented reality (VR/AR) [Wu *et al.*, 2025; Liu *et al.*, 2025b]. Despite their visual fidelity, current models often hallucinate physical laws, lacking the causal consistency required for robotic world models to reliably predict action-outcome dynamics. By anchoring generation between explicit start and end states, our framework bridges this gap, enabling the synthesis of physically valid transitions essential for embodied AI.

To address this issue, some studies have incorporated auxiliary conditions, such as segmentation maps [Akkerman *et al.*, 2025] or used LLM to improve physical fidelity [Xue *et al.*, 2025; Zhang *et al.*, 2025a]. Others have focused on fine-tuning models on newly collected, large-scale human-object interaction datasets [Liu *et al.*, 2025a].

Nevertheless, methods using auxiliary conditions lack generalizability, while data-driven approaches are constrained by the limited scope of existing datasets. Manually curating large-scale datasets is prohibitively costly, making synthetic data a viable alternative. However, generating high-quality, diverse, and physically plausible synthetic interactions remains an open challenge.

In this work, inspired by the recent success of training LLMs with synthetic data [Li *et al.*, 2024a; Zhao *et al.*, 2025], we propose a framework for constructing a dataset to improve the capacity of video generative models in synthesizing physically plausible interactions and subsequent object state transitions as shown in Figure 1. Specifically, we first define a taxonomy for generating prompts that describe plausible interactions, comprising attributes such as the interactable object and the type of state transition-oriented interaction. However, we identified a key challenge in directly applying these prompts to video generation models, as the resulting outputs often fail to accurately depict the intended interactions.

Therefore, we incorporate an intermediate image generation step to ensure fidelity. Instead of direct text-to-video synthesis, we first generate an initial image from the prompt. Following that, a state-of-the-art image editing model alters this image to reflect the object's subsequent state change according to the given interaction. The video generation is then conditioned on this pair of images, which clearly defines the start and end states of the interaction.

However, we observe that relying solely on the pair for the first-to-last frame generation often leads to undesirable artifacts, including abrupt scene changes or unnatural transformations that compromise the video's temporal consistency. To address this challenge, we propose State-Guided Sampling (SGS), a novel sampling method designed to guide the model toward a smooth and plausible state transition.

To validate our approach, we develop an automated system to evaluate interaction quality, confirming its reliability against human judgments. Our experiments show that a model fine-tuned on our curated dataset significantly enhances its capability to generate complex interactions. Our main contributions are as follows:

- We introduce a novel pipeline to construct a synthetic dataset for diverse object interactions, based on a structured taxonomy and state-of-the-art image editing models that create explicitly 'start' and 'end' state images.

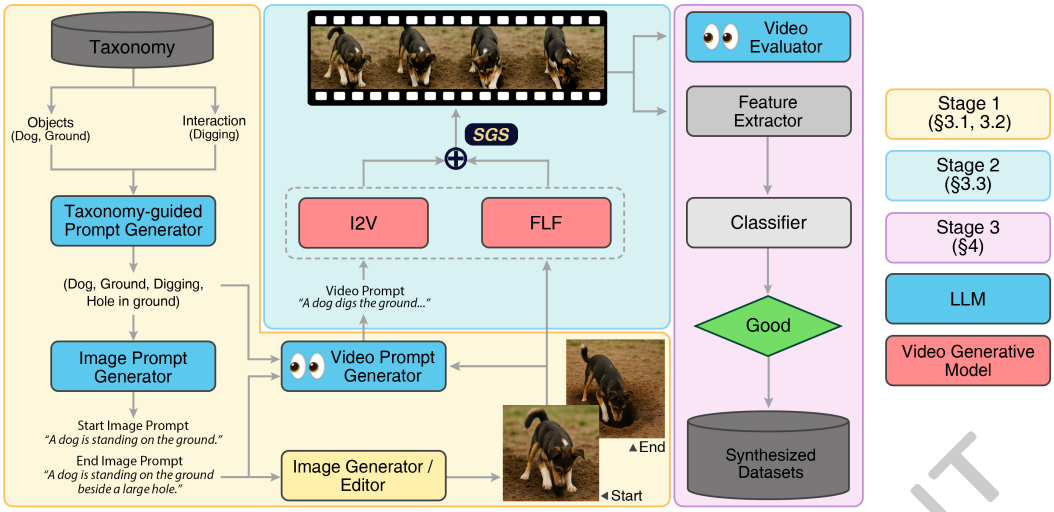

Figure 1: **Overview of Interaction-Centric Video Dataset Generation Pipeline.**

- We propose State-Guided Sampling (SGS), a novel sampling strategy that mitigates visual artifacts and guides video models to generate seamless state transitions.
- We develop a model-based evaluation system to assess interaction quality in generated videos, and validate its alignment with human judgments to ensure data quality and facilitate scalable dataset curation.
- We release the dataset, data generation pipeline, and evaluation tools to the public at this github page.

2 Related Works

2.1 Video Generative Model

Early video diffusion models evolved from U-Net architectures by incorporating separate temporal modules [Ho *et al.*, 2022; Blattmann *et al.*, 2023; Guo *et al.*, 2023; Xing *et al.*, 2023; Hong *et al.*, 2022]. The subsequent emergence of the Diffusion Transformer (DiT) architecture [Peebles and Xie, 2023] unified spatial and temporal modeling into a single backbone [Ma *et al.*, 2024], demonstrating superior scalability in modeling complex dynamics. While recent DiT-based models [Liu *et al.*, 2024b; Zheng *et al.*, 2024; Yang *et al.*, 2024; Kong *et al.*, 2024; Wan, 2025] excel in visual fidelity, they often exhibit limitations in accurately capturing physical laws and interactions, particularly under out-of-distribution (OOD) prompts [Xue *et al.*, 2025].

2.2 Generating Dynamic Interactions in Videos

Prior works on interaction generation face trade-offs between controllability and generality. For example, InterDyn [Akkerman *et al.*, 2025] uses segmentation maps for fine-grained control, which limits its applicability. Similarly, HOIGen [Liu *et al.*, 2025a] is constrained to the human-object interaction domain. Another line of research focuses on injecting knowledge into the models. Approaches like PhyT2V [Xue *et al.*, 2025] utilize MLLMs to iteratively refine prompts, but this does not enhance the internal capabilities of the generative model itself. DiffPhy [Zhang *et al.*, 2025a] enhances training prompts with physically-grounded descriptions from an LLM, while VideoREPA [Zhang *et al.*, 2025b] injects physical knowledge via representation matching from a video encoder. However, these methods are fundamentally oriented towards acquiring knowledge from real-world data, thus limiting their capacity to generate novel and creative interaction scenarios.

In contrast, our work decouples the generation process from the dependency on existing video datasets. We achieve this by introducing a new form of general-purpose controllability, which utilizes 'start' and 'end' images as result-driven visual anchors to precisely define an interaction's outcome. This novel control mechanism is the key that enables our scalable, synthetic dataset generation framework, capable of generating high-quality, novel, and creative interactions.

2.3 Synthetic Dataset Generation

While the performance of large-scale generative models relies heavily on the amount of high-quality data, the growing challenges of collecting and curating real-world datasets have made the use of synthetic data an increasingly essential research direction. Early works leveraged synthetic data to induce novel capabilities; LLaVa [Liu *et al.*, 2023] used image metadata to create visual instruction-tuning data, and InstructPix2Pix [Brooks *et al.*, 2023] built datasets by refining ambiguous prompts. The approach employing LLM has been further advanced through self-improvement methods that enhance performance via self-correct [Liu *et al.*, 2024a; Welleck *et al.*, 2022] and filtering [Li *et al.*, 2024b]. Building on these ideas, subsequent works like GLAN [Li *et al.*, 2024a] and Absolute-Zero [Zhao *et al.*, 2025] have even eliminated the need for initial seed data. Inspired by this line of research, we propose a framework that generates and filters a synthetic dataset specialized for interaction and state change. Our framework uniquely leverages image generation and editing models, in conjunction with our State-Guided Sampling method, to create high-quality, targeted data.

(a) A person wearing gloves picks up the spray paint can and carefully sprays a vibrant, glossy coat over the daffodil's petals and stem, transforming its appearance in the evening garden.

(b) A person picks up the yellow sponge and thoroughly scrubs the kitchen counter, removing the layer of grime to reveal a spotless, gleaming surface.

Figure 2: **Qualitative Evaluation.** The red arrow indicates abrupt frame changes, while the plus (+) denotes our fine-tuned model.

3 Synthetic Interaction Video Generation

To generate high-fidelity videos of complex interactions, we propose a multi-stage data generation pipeline. Our approach is designed to overcome the limitations of direct text-to-video synthesis by ensuring both semantic accuracy and temporal consistency. The pipeline consists of three main stages: (1) structured prompt generation based on a custom taxonomy focusing on state transitions, (2) synthesis of 'start' and 'end' state image pairs using state-of-the-art image editing, and (3) temporally coherent video generation guided by the proposed State-Guided Sampling (SGS) technique. Figure 1 overviews our proposed dataset generation pipeline, where each stage is designed to address key challenges in generating plausible object interactions, as detailed in the following subsections.

3.1 Taxonomy-Guided Prompt Generation

To generate diverse scenarios and avoid the bias of naive LLM querying towards common interactions (e.g., *holding a cup*), we introduce a structured prompt generation process. We utilize a custom taxonomy inspired by ImageNet [Deng *et al.*, 2009], covering approximately 1,300 objects and 500 interaction types organized by physical outcomes (e.g., *Deformation*). Our process employs two complementary sampling strategies. *Object-centric sampling* randomly pairs objects and uses an LLM to validate plausibility and generate state changes, facilitating the discovery of novel scenarios. Conversely, *interaction-centric sampling* generates diverse object combinations for specific interactions using localized contexts to prevent repetition. This dual strategy yields struc-

tured tuples of (`object1`, `object2`, `interaction`, `state change`). Taxonomy details are provided in the Appendix.

3.2 Interaction-State Image Pair Synthesis

While the structured tuple provides a clear semantic description, direct text-to-video synthesis often fails to faithfully render the specified interaction and its precise state transition. To address this, we adopt an image-to-video framework that relies on explicit visual conditions. The generated tuple in Section 3.1 is used to synthesize a pair of images: a 'start' image depicting the scene prior to the interaction, and an 'end' image reflecting the object's state change.

To synthesize a sample, we first use the tuple to compose a detailed prompt for a text-to-image model to generate the initial frame. Subsequently, the `state change` component of the tuple guides a state-of-the-art image editing model to modify the initial image into the final frame. This resulting image pair provides strong visual anchors that explicitly define the start and end points of the interaction for the video generation model. We use GPT-4o [Hurst *et al.*, 2024] for prompt generation, image creation, and editing.

3.3 State-Guided Sampling

Given the 'start' (first-frame) and 'end' (last-frame) images, the next stage is to synthesize a video that plausibly and seamlessly connects these two states while representing the prompt. The primary challenge is balancing global guidance toward the end state with local, frame-to-frame temporal coherence. An I2V model (start-frame conditioned) provides local coherence but lacks global direction, whereas an FLF model (start-and-end-frame conditioned), which is equivalent to a video infilling model [Höppe *et al.*, 2022], has strong global guidance but can produce artifacts. To resolve this trade-off, we introduce State-Guided Sampling (SGS), a novel sampling technique that dynamically combines the velocity fields of an I2V model (v_I) and an FLF model (v_F). SGS performs vector field interpolation within the flow-matching framework [Lipman *et al.*, 2022]. Analogous to Classifier-Free Guidance [Ho and Salimans, 2022], blending the locally-coherent field v_I and the globally-constrained field v_F steers the generation trajectory within the valid data manifold, satisfying both initial and terminal conditions. Our final velocity field, v_{sgs}, is defined as a dynamic, frame-wise weighted sum of the two as follows:

$$v_{\text{sgs}}(z_t, t, c') = (\mathbf{1} - W) \odot v_F(z_t, t, c') + W \odot v_I(z_t, t, c), \quad (1)$$

where $\odot$ denotes the frame-wise product, W is the frame-wise weight. c represents the condition for the prompt and start image, and c' further includes the last image. We find that naively linearly interpolating the velocity fields creates a 'ghosting effect' as shown in Appendix, a translucent overlay of the start and end states where the FLF model's rigid guidance conflicts with local temporal consistency. To alleviate this, we design a dynamic frame-wise weighting scheme where the weight W_f for each frame index f to smoothly transit the model's reliance from global guidance to local coherence, calculated using a normalized exponential curve:

$$W_f = \alpha + (\beta - \alpha) \cdot \frac{e^{k \cdot \frac{f}{F-1}} - 1}{e^k - 1} \quad (2)$$

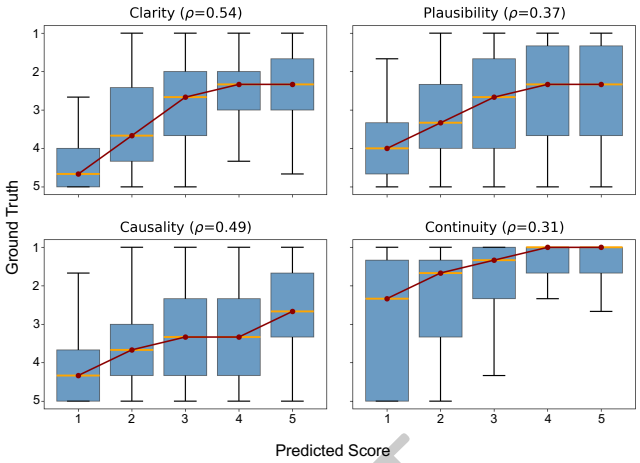

Figure 3: **VLM-Human Rating Correlation.**

where F is the total number of frames. The interpolation begins with a starting weight α to ensure an initial blend of both models and concludes with $\beta = 1.0$, allowing the I2V model to dominate the final frames for a coherent finish. The parameter k controls the curve's steepness, determining how long the FLF model's stronger state guidance is maintained. We set $\alpha = 0.5$ and $k = 5.0$ in our experiments. In essence, SGS resolves the conflict between global and local objectives by utilizing the FLF model's trajectory guidance and gradually shifting to the I2V model's strength in visual consistency, ultimately producing a plausible and seamless state transition.

4 Interaction Quality Assessment

Holistically evaluating whether a generated video contains plausible interactions and state transitions is crucial, yet a standardized metric is currently lacking. Existing metrics are often limited to specific domains and fail to capture semantic plausibility. Therefore, we identify the need for a comprehensive framework capable of assessing the diverse aspects of quality of generated dynamic events. This section introduces our proposed hybrid evaluation framework, which combines the semantic understanding of a large Vision-Language Model (VLM) with specialized, auxiliary features to ensure robust and reliable assessment.

4.1 Criteria for Interaction Quality

To overcome the limitations of prior metrics, we first define a framework that assesses video quality across four key criteria. The first, *Interaction Presence & Clarity*, evaluates whether the specified interaction occurs and is unambiguously depicted in the video. The second criterion, *Interaction-State Causality*, assesses if the object's state transition is a direct and causal consequence of the specified interaction. The third, *Physical Plausibility*, determines if the object's motion and the interaction's outcome adhere to physical principles. The last criterion, *Temporal Continuity*, checks for a smooth and consistent flow, free of visual artifacts such as dissolves, scene cuts, or distortions.

	F1 (Good)	F1 (Bad)	Macro F1	AUC
VideoPhy2	0.00	0.80	0.40	0.57
Base	0.53	0.80	0.66	0.71
(a) + Prefilter	0.59	0.77	0.68	0.72
(b) + PC score	0.60	0.78	0.69	0.72
(c) + Surprise	0.61	0.79	0.70	0.73
(d) + PP	0.61	0.79	0.70	0.76
(e) **+ QC**	**0.63**	**0.80**	**0.71**	**0.77**

Table 1: **Ablation Study on SVM Inputs.** PP and QC denote the features from Plausibility Probe and Quality Classifier, respectively.

4.2 VLM-Assisted Evaluation and Limitations

We initially used a VLM (Gemini-2.5-Pro [Comanici *et al.*, 2025]) to score videos from 1 to 5 across our four criteria, validating it against 1,146 human-annotated videos. Pearson correlation was strong in semantic like *Clarity* ($\rho = 0.54$), and *Causality* ($\rho = 0.49$). but weaker for Plausibility ($\rho = 0.37$) and *Continuity* ($\rho = 0.31$). It shows that using VLM only for evaluation is not sufficient to match human judgment.

4.3 Enhanced Evaluation with Auxiliary Features

To enhance the reliability of our framework, we adopt a hybrid approach by integrating auxiliary features for compensating weaker criteria. Specifically, to augment *Physical Plausibility* and *Temporal Continuity*, we use two external features: *Physical Commonsense* score [Bansal *et al.*, 2025] from a specialized VLM and *Surprising* score [Garrido *et al.*, 2025] from a pre-trained V-JEPA 2 model. Furthermore, we incorporate predictions from our own trained Plausibility Probe (PP) and Quality Classifier (QC) as additional features, which are lightweight attention modules trained on V-JEPA2 to predict the human-annotated physical plausibility scores and quality labels, respectively. Lastly, we utilize a dedicated temporal artifact detector, which is detailed in the following subsection. We experimentally verify the effectiveness of each auxiliary feature in Section 5.2.

4.4 Temporal Artifact Detection for Continuity

A persistent challenge that degrades *Temporal Continuity* is the occurrence of abrupt scene transitions or dissolve artifacts in generated videos. While we initially investigated using a powerful VLM for this task, we found it was unable to reliably detect such sudden and unnatural scene changes. To this end, we employ a frozen V-JEPA 2 [Assran *et al.*, 2025] model as a feature extractor and train a transformer-based attention classifier for a binary classification task.

For the detector, we constructed a synthetic dataset by applying one of three distinct augmentations to video clips from the UCF-101 dataset [Soomro *et al.*, 2012], which suffices for learning low-level structural discontinuities independent of high-resolution semantics. Each process simulates a different type of temporal artifact described as follows.

Cross-Fade Transition is an alpha blending transition applied between two clips, where the start time and duration of the fade are randomized.

Hard Cut simulates an abrupt scene change by concatenating segments from two distinct videos at a randomized temporal midpoint without any blending.

Intra-Scene Displacement creates continuity errors that mimic a camera jump. A video clip is cut at a random point, the entire second segment is spatially translated, and the two segments are then rejoined with a brief cross-fade. A detailed performance evaluation is given in Appendix.

5 Experiments

To evaluate our dataset and methodology, we select the state-of-the-art open-source model, Wan 2.1 [Wan, 2025], as our base model for fine-tuning. Its strong performance in Image-to-Video (I2V) generation makes it a suitable foundation for validating our proposed contributions. For performance comparison, we benchmark against two strong baselines: HunyuanVideo [Kong *et al.*, 2024] and PhyI2V[1] adapted from PhyT2V [Xue *et al.*, 2025].

5.1 Dataset Construction Details

Our synthetic dataset is built through an iterative generation and curation process as an offline data engine, where the one-time generation cost is effectively amortized by the downstream model efficiency. We first generate an initial pool of 1,146 videos from 191 prompts using six different sampling methods (I2V, FLF, and four SGS with varying settings of α). This initial set undergoes detailed human annotation by the domain experts, who provide (1) a binary 'Good'/'Bad' quality label, (2) 1-5 scores for our four proposed criteria (*Clarity, Causality, Plausibility, Continuity*), and (3) a relative ranking of the generated videos per prompt. The final 'Good'/'Bad' labels are determined by majority vote, and the criterion scores are averaged. This high-quality, human-filtered data is then used to fine-tune the base model in a lightweight manner [Hu *et al.*, 2022].

To construct the complete dataset, we employ both original and fine-tuned models to generate an additional 4,971 videos from 1,018 new prompts. This expanded set is then curated using our automated evaluation system detailed in Section 4. The final training dataset combines the initially human-filtered samples from the bootstrapping phase with the videos that successfully passed our automated filtering process. The final training set consists of 1,525 videos from 681 prompts, a volume we believe is sufficient for effective alignment [Zhou *et al.*, 2023]. Each video has 81 frames and is 5 seconds long. A held-out set of 105 prompts is used for validation.

5.2 Reliability of Evaluation Framework

To validate the reliability of the proposed evaluation framework, we use the 1,146 human-annotated videos from our bootstrapping set as the ground truth for this analysis. Our final goal for the evaluation pipeline is to train an SVM classifier [Hearst *et al.*, 1998] on the proposed features to accurately predict the human 'Good'/'Bad' labels.

We conducted an ablation study to quantify each component's contribution. While VLM scores provide a strong semantic baseline (Figure 3), they require augmentation for physical and temporal assessment. To identify the most effective feature set, we conducted a comprehensive ablation

[1]PhyI2V is PhyT2V with CogVideoX-5B-I2V.

(a) Watch as the person deftly folds the paper in half.

(b) A person in a sunny kitchen pours orange juice into the sink.

Figure 4: **Effectiveness of State-Guided Sampling.** The proposed SGS effectively resolves the unnatural scene transitions.

Model		VLM-Assisted Score (↑)					Temporal Artifact (↓)	VideoPhy2 (↑)	
		Clarity	Causality	Plausibility	Continuity	Average		SA	PC
PhyI2V[1]		2.52	2.34	2.44	3.40	2.68	0.44	0.26	0.58
HunyuanVideo		1.71	1.50	**3.06**	**4.16**	2.61	0.09	0.25	**0.77**
Wan 2.1	FLF	3.10	3.06	2.87	3.20	3.06	0.68	0.31	0.56
	I2V	2.90	2.86	2.98	3.87	3.15	0.13	0.30	0.56
	SGS	3.00	2.94	2.91	3.96	3.20	0.12	0.26	0.53
Wan 2.1 (Fine-tuned)	I2V	3.20	3.05	2.92	3.96	3.28	**0.07**	**0.34**	0.57
	SGS	**3.23**	**3.21**	3.03	3.90	**3.34**	0.11	0.32	0.53

Table 2: **Quantitative Evaluation.** SA and PC denote Semantic Adherence and Physical Commonsense, respectively. The **bold** represents the best, and the underline does the second best.

study by training an SVM classifier with cross-validation. A baseline using only a pre-existing Physical VLM [Bansal *et al.*, 2025] fails to identify any 'Good' videos, yielding an F1-score (pos) of 0.00. In contrast, using our four VLM-based scores as base features (Base) provides a much stronger starting point, achieving a Macro F1 of 0.66. To prevent temporal artifacts from confounding our analysis, the validation protocol pre-filters videos using the temporal artifact detector. As shown in Table 1(a), this step isolates the effectiveness of other features and improves the positive-class F1 score.

On the cleaned dataset, we then incrementally added our auxiliary features to complement physical plausibility. The addition of the VideoPhy2 *PC* score (Table 1(b)) and the V-JEPA 2 *surprise* score (Table 1(c)) steadily increased performance. We observed further improvement by incorporating features from our Attentive Plausibility Probe (Table 1(d)), a shallow attention module added to V-JEPA 2 that is trained to predict the fine-grained human-annotated *Plausibility* scores. The best performance was achieved with our final feature set (Table 1(e)), which additionally incorporates the output from an Attentive Quality Classifier. This classifier is trained on V-JEPA 2 features to directly predict the final 'Good'/'Bad' human labels. To prevent label leakage when using the predictions from our Attentive Modules (Table 1(d,e)) as features, we employ a k-fold cross-validation strategy (we use $k = 5$). These out-of-fold predictions are then used as a "clean" feature for training our final SVM model, ensuring a fair and rigorous evaluation. This result validates that our multi-faceted feature design is highly effective at capturing the complex nuances of human judgment. The detailed classification report for this final classifier is presented in Appendix.

5.3 Main Results

Quantitative Results. We present our quantitative results in Table 2, evaluating our models against baselines on the held-out validation set. The evaluation uses our proposed Gemini-based scores, the temporal artifact rate, and metrics from VideoPhy2. As shown, our fine-tuned models achieve the highest average Gemini score and the best *SA* score, while also exhibiting a significantly lower temporal artifact rate. While Hunyuan attains high scores in *Plausibility* and *Continuity*, we attribute this to its tendency to generate static or moderate-action videos, which results in critically low scores for *Clarity* and *Causality*. To further validate this hypothesis, we evaluated the I2V-based models on the VBench benchmark [Huang *et al.*, 2024] in Table 3. This analysis confirmed our observation: Hunyuan recorded the lowest *Dynamic Degree* (0.20). In contrast, our fine-tuned I2V not only improves the *Dynamic Degree* from 0.64 to 0.73 but does so without compromising other criteria. Furthermore, in direct human

	Vbench			Human Comparison	
	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Rank(↓)	WR
PhyI2V	0.99	0.57	0.57	2.90	8%
Hunyuan	0.99	0.20	0.63	3.43	2%
Wan2.1 I2V	0.98	0.64	0.63	1.98	29%
+Fine-tune	0.98	**0.73**	0.63	**1.68**	**53%**

Table 3: **VBench (Motion) and Human Evaluation.**

	Clar.	Caus.	Plau.	Cont.	Rank(↓)	Good	Temporal Artifact(↓)
FLF	2.96	2.64	2.82	2.41	4.08	0.15	0.66
I2V	2.64	2.36	3.01	4.22	3.71	0.22	0.13
$\alpha = 0.3$	3.04	2.72	3.05	4.21	3.05	0.30	0.26
$\alpha = 0.4$	3.13	2.81	**3.11**	4.30	**2.98**	0.35	0.20
$\alpha = 0.5$	**3.14**	**2.86**	3.07	**4.56**	3.05	**0.38**	**0.11**
$\alpha = 0.6$	2.95	2.65	3.08	4.55	3.38	0.31	**0.11**

Table 4: **Ablation Study on the Initial I2V Weight, α, for SGS.**

comparisons, our model demonstrated the best rank and the highest win rate (WR) against all baselines. Those results show the effectiveness of our synthetic data.

In addition, to validate whether our model overfits to our data synthesis pipeline, we evaluated it on PhyGenBench [Meng *et al.*, 2024]. As PhyGenBench is a text-to-video benchmark, we adapted it to our image-to-video setting. For each video prompt, we first employed an LLM to generate a corresponding image prompt describing the initial frame. We then utilized Flux-dev [Labs, 2024], an open-source text-to-image model, to synthesize the start image. By using a different image generator from the one in our data creation pipeline, we rigorously tested whether our model generalizes to a novel visual distribution. As reported in Table 5, the fine-tuned model (+Fine-tuning) shows marginal improvement over the Wan2.1 I2V in automated evaluation. However, it achieves a significant gain in human evaluation, attaining the highest win rate (42%) and best rank (1.90). We conjecture this discrepancy stems from limitations of the PhyGenBench automated evaluation system in the I2V context.

Firstly, when the initial image already contains complex phenomena described in the prompt (e.g., reflections in a mirror, shadows from a light source), the video model is rewarded for generating a static video, penalizing plausible motion. Secondly, the system's reliance on retrieved key-frames hinders distinguishing subtle yet critical differences between videos generated from the same start image. Notably, PhyI2V excelled in the Mechanics and Material categories. Its success in Material likely stems from the LLM's world knowledge (e.g., vinegar is poured into a glass of litmus solution). In contrast, its high Mechanics score appears to be an evaluation hacking artifact; its VLM-based refinement process may overfit to the VLM-based evaluator, a bias suggested by near-zero scores of other models' plausible videos. Nevertheless, our model's superior human-evaluation performance demonstrates that our synthetic dataset enables robust generalization to OOD scenarios and complex physical phenomena. Qualitative results for PhyGenBench are provided in Appendix.

	PhyGenBench					Human Comparison	
	Mechanics	Optics	Thermal	Material	Average	Rank (↓)	WR
PhyI2V	**0.51**	0.62	0.54	**0.49**	**0.55**	2.79	19%
Hunyuan	0.43	0.57	0.42	0.28	0.44	2.93	12%
Wan2.1 I2V	0.49	0.59	0.54	0.39	0.51	2.38	27%
+Fine-tune	0.47	**0.63**	**0.57**	0.41	0.52	**1.90**	**42%**

Table 5: **Evaluation on the PhyGenBench. (I2V)**

Qualitative Results. Figure 2 shows a qualitative comparison between baselines and our fine-tuned models. PhyI2V and Hunyuan fail to generate proper interaction and the target state change. The FLF model shows some evidence of interaction and state transition, but suffers from severe temporal artifacts; in example (a), an abrupt change occurs between subsequent frames (indicated by a red arrow) while in, (b), the video culminates in a final frame with a suddenly different appearance. The I2V model also struggles, and while it attempts the interaction, it fails to depict the target state change (e.g., not removing the grime in (b)) and introduces critical object consistency artifacts, such as a second spray can in (a) or an extra hand appearing abruptly in (b). In contrast, our zero-shot SGS produces a plausible interaction and its result. This performance is further enhanced with our fine-tuned models, as both I2V$^+$ and SGS$^+$ generate significantly clearer interactions and state transitions. Consistent with the findings in Figure 2, Figure 4 illustrates the effectiveness of SGS over FLF and I2V. While FLF introduces visual artifacts (e.g., background shifts in the red box) and I2V fails to generate any interaction, SGS successfully creates the interactions.

5.4 Ablation Study on Sampling Methods

To validate SGS, we conduct human evaluation across six sampling configurations (Table 4). The baselines confirm a distinct trade-off. FLF achieves high semantic adherence but suffers from temporal artifacts (0.66), whereas I2V preserves continuity but often fails to depict the target state change. SGS effectively resolves this conflict. We find that performance is sensitive to the weighting parameter α; values ≤ 0.4 fail to sufficiently mitigate artifacts, while $\alpha = 0.6$ dilutes the global guidance from the FLF model. Consequently, $\alpha = 0.5$ yields the optimal balance, achieving the highest 'Good' ratio (0.38) and superior scores across all criteria.

6 Conclusion

We proposed a novel, modular framework using a taxonomy-guided pipeline, visual anchors, and State-Guided Sampling to generate controllable object interaction videos. Our experiments show that fine-tuning on this synthetic data significantly enhances a model's ability to create complex interactions. While the current implementation relies on a proprietary model, the pipeline's modularity enables integration with open-source alternatives, offering a path toward transparent, reproducible data generation for various applications.

Contribution Statement

Jiho Jang and Jin-Young Kim contributed equally to this work. Kyungjune Baek served as the corresponding author.

Acknowledgements

This work was supported by the Korean Government through the grants from IITP (RS-2021-II211343 and RS-2025-25442338), KOCCA (RS-2024-00398320), and NRF (RS-2026-25470591).

References

[Agarwal *et al.*, 2025] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[Akkerman *et al.*, 2025] Rick Akkerman, Haiwen Feng, Michael J Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[Assran *et al.*, 2025] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[Bansal *et al.*, 2025] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

[Blattmann *et al.*, 2023] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[Brooks *et al.*, 2023] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

[Comanici *et al.*, 2025] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.

[Garrido *et al.*, 2025] Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

[Guo *et al.*, 2023] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[Hearst *et al.*, 1998] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4), 1998.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

[Hong *et al.*, 2022] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[Höppe *et al.*, 2022] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2), 2022.

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

[Hurst *et al.*, 2024] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

[Labs, 2024] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[Li *et al.*, 2024a] Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, et al. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*, 2024.

[Li *et al.*, 2024b] Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. Superfiltering: Weak-to-strong data filtering for

fast instruction-tuning. *arXiv preprint arXiv:2402.00530*, 2024.

[Lipman *et al.*, 2022] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023.

[Liu *et al.*, 2024a] Dancheng Liu, Amir Nassereldine, Ziming Yang, Chenhui Xu, Yuting Hu, Jiajie Li, Utkarsh Kumar, Changjae Lee, Ruiyang Qin, Yiyu Shi, et al. Large language models have intrinsic self-correction ability. *arXiv preprint arXiv:2406.15673*, 2024.

[Liu *et al.*, 2024b] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[Liu *et al.*, 2025a] Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. Hoigen-1m: A large-scale dataset for human-object interaction video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[Liu *et al.*, 2025b] Tianqi Liu, Zihao Huang, Zhaoxi Chen, Guangcong Wang, Shoukang Hu, Liao Shen, Huiqiang Sun, Zhiguo Cao, Wei Li, and Ziwei Liu. Free4d: Tuning-free 4d scene generation with spatial-temporal consistency. *arXiv preprint arXiv:2503.20785*, 2025.

[Ma *et al.*, 2024] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

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

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

[Soomro *et al.*, 2012] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[Wan, 2025] Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[Welleck *et al.*, 2022] Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*, 2022.

[Wu *et al.*, 2025] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[Xing *et al.*, 2023] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. 2023.

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

[Yang *et al.*, 2024] 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.

[Zhang *et al.*, 2025a] Ke Zhang, Cihan Xiao, Yiqun Mei, Jiacong Xu, and Vishal M Patel. Think before you diffuse: Llms-guided physics-aware video generation. *arXiv preprint arXiv:2505.21653*, 2025.

[Zhang *et al.*, 2025b] 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*, 2025.

[Zhao *et al.*, 2025] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025.

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

[Zhou *et al.*, 2023] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2023.