RoboScape: Physics-informed Embodied World Model

Yu Shang[1], **Xin Zhang**[2], **Yinzhou Tang**[1], **Lei Jin**[1], **Chen Gao**[1], **Wei Wu**[2]*, **Yong Li**[1]*

[1]Tsinghua University
[2]Manifold AI

Abstract

World models have become indispensable tools for embodied intelligence, serving as powerful simulators capable of generating realistic robotic videos while addressing critical data scarcity challenges. However, current embodied world models exhibit limited physical awareness, particularly in modeling 3D geometry and motion dynamics, resulting in unrealistic video generation for contact-rich robotic scenarios. In this paper, we present RoboScape, a unified physics-informed world model that jointly learns RGB video generation and physics knowledge within an integrated framework. We introduce two key physics-informed joint training tasks: temporal depth prediction that enhances 3D geometric consistency in video rendering, and keypoint dynamics learning that implicitly encodes physical properties (e.g., object shape and material characteristics) while improving complex motion modeling. Extensive experiments demonstrate that RoboScape generates videos with superior visual fidelity and physical plausibility across diverse robotic scenarios. We further validate its practical utility through downstream applications including robotic policy training with generated data and policy evaluation. Our work provides new insights for building efficient physics-informed world models to advance embodied intelligence research. Our code and demos are available at: `https://github.com/tsinghua-fib-lab/RoboScape`.

1 Introduction

The advancement of large language and vision models [1, 2] has demonstrated the critical role of high-quality, large-scale training data for their superior performance. However, the robotic learning is significantly hindered by the prohibitive cost of collecting real-world data [3, 4, 5, 6], which often relies on human teleoperation to acquire high-quality demonstrations. This limitation poses a great challenge for scaling robotic learning and deploying agents in complex, real-world environments.

World models [7, 8, 9, 10, 11], which simulate environmental dynamics by predicting future states based on current observations and given actions, offer a promising solution to this data scarcity problem. Such models hold significant promise for advancing embodied intelligence by generating realistic robotic data [12] and enabling scalable simulation environments [13]. However, current embodied world models [13, 14, 15] predominantly focus on video generation, with training objectives centered on optimizing the RGB pixels. While capable of producing visually plausible 2D images, they often fail to maintain crucial physical properties, such as motion plausibility and spatial consistency [16]. Particularly, in robotic manipulation tasks involving deformable objects (e.g., cloth), the generated videos frequently contain artifacts such as unrealistic object morphing or discontinuous motion. These limitations become particularly detrimental in interaction-rich robotic scenarios, where even minor physical inconsistencies can dramatically compromise the effectiveness of learned policies.

*Corresponding author, correspondence to liyong07@tsinghua.edu.cn.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

The root cause lies in existing models' overreliance on visual token fitting without awareness of physical knowledge [17, 18, 19]. To address this, we propose a physics-informed world model that jointly learns depth information and temporal keypoint consistency to implicitly encode physical constraints. Existing efforts of integrating physical knowledge into video generation fall into three categories: physics-prior regularization, physics simulator-based knowledge distillation, and material field modeling. Current regularization-based methods enforce constraints such as local rigidity [20] or rotational similarity [21] on Gaussian splatting (GS) features or 3D point clouds. However, these methods are limited to narrow domains like human motion [22] or rigid-body dynamics [20], hindering generalization to diverse robotic scenarios. Another line of work employs physics simulators to extract motion signals or semantic maps as conditions to guide video generation models [23, 24, 25, 26]. Although this approach yields reliable physical priors, the resulting cascaded pipeline introduces excessive computational complexity, hindering their practical deployment. There have been some recent works trying to enhance the physical simulation via material field modeling [27, 28]. However, such methods are confined to object-level modeling and are hard to apply to scene-level generation.

To overcome these limitations, we propose RoboScape, a physics-informed world model based on a multi-task learning auto-regressive framework to generate visually realistic and physics-adherent robotic videos. Specifically, our approach incorporates physics knowledge through two auxiliary physics-informed supervision tasks within the world model itself to alleviate heavy external model cascading. First, to empower the model with 3D spatial physical understanding, we augment the RGB prediction backbone with a temporal depth prediction branch and inject the learned depth features into the RGB prediction to enhance spatial awareness. Such synergistic learning of temporal depth maps enables the model to implicitly acquire 3D scene reconstruction priors rather than merely fitting 2D RGB images. Second, we introduce an adaptive keypoint dynamics learning task to address unrealistic object deformation and implausible motion issues. To achieve this, we first perform dynamic keypoint sampling to automatically identify regions with significant motion (typically involving robots and interacting objects), then encourage temporal token consistency for these keypoints across frames. Through this, the model effectively captures the deformation properties and motion behaviors of objects, implicitly encoding material properties (e.g., rigidity and softness) through self-supervised keypoint consistency, eliminating the need for explicit material modeling. Although some recent world models [29, 30] also explore joint RGB-depth prediction, their learning remains constrained at the image level, failing to capture the fine-grained motion dynamics and object deformation details that are crucial for robotic manipulation scenarios. Furthermore, these approaches exhibit a performance trade-off, where gains in 3D perception come at the cost of reduced RGB prediction fidelity. Differently, our model captures global spatial knowledge through learning temporal depth dynamics, while modeling local object deformation and motion characteristics via learning temporal keypoint tracking.

We conduct comprehensive experiments to evaluate our world model from three aspects: video generation quality, robotic policy learning using synthetic data, and robotic policy evaluation. RoboScape achieves state-of-the-art performance in both RGB and depth prediction accuracy, achieving a superior balance between these metrics compared to existing world model baselines. Additionally, we validated that synthetic data from our world model consistently improves the performance of robotic policy models within a simulated robotic environment including Diffusion Policy [31] and pi0 [32], confirming the model's practical utility for robotic learning. Finally, our model can also serve as a reliable policy evaluator, with assessment results showing strong correlation with ground-truth simulator outcomes, confirming our model's capability to accurately model the physical world.

In summary, the main contributions of the paper are as follows:

- We propose RoboScape, a physics-informed embodied world model that unifies RGB video generation, temporal depth prediction, and adaptive keypoint tracking in a joint learning framework, achieving both high visual fidelity and physical plausibility.
- We design an automated robotic data processing pipeline with physical prior information labels. Trained on the carefully curated large-scale, high-quality dataset, our model achieves SOTA performance on visual quality, geometric accuracy, and action controllability.
- We demonstrate the practical utility of RoboScape on downstream applications including robotic policy training and evaluation. Extensive experimental results demonstrate its effectiveness in accurately modeling embodied environments, validating its potential for advancing real-world robotic deployment.

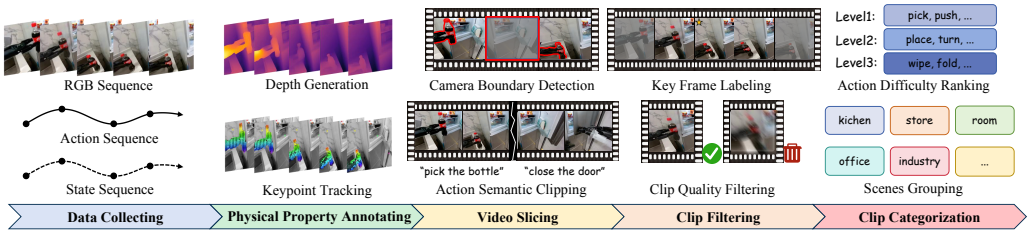


Figure 1: Illustration of the proposed robotic data processing pipeline with physical priors annotation.

2 Methodology

2.1 Problem Formulation

In this work, we focus on robot manipulation scenarios and learn an embodied world model f_θ as a dynamics function that predicts the next visual observation $\mathbf{o}_{t+1}$ given past observations $\mathbf{o}_{1:t}$ and robotic actions $\mathbf{a}_{1:t}$:

$$\mathbf{o}_{t+1} \sim f_\theta(\mathbf{o}_{t+1}|\mathbf{o}_{1:t}, \mathbf{a}_{1:t}), \tag{1}$$

where $\mathbf{o} \in \mathbb{R}^{H\times W\times 3}$ is a video frame and $\mathbf{a} \in \mathbb{R}^k$ is a k-degree continuous action control vector.

2.2 Robotic Data Processing Pipeline with Physical Priors Annotation

Learning a physics-informed embodied world model requires high-quality dataset covering high-resolution RGB and depth sequences, action sequences that control the robot, and state sequences that the robot executes. In this section, we present our data processing pipeline to construct a multi-modal embodied dataset with physical priors based on AGIBOT-World dataset [6], as shown in Fig. 1.

Physical Property Annotating based on Depth Generation and Keypoint Tracking. Integrating explicit physics constraints remains a significant challenge for current video generation-based world models. To address this, our approach concentrates on two crucial, visually-expressible physics constraints highly relevant to robotic manipulation: temporal depth consistency and keypoint motion trajectories. These features can be efficiently extracted using off-the-shelf pretrained models, enabling enhanced generalization while maintaining practical feasibility. Specifically, we utilize Video Depth Anything [33] to generate the depth map sequence of the video. Furthermore, we apply SpatialTracker [34] as the keypoint tracking model to sample the keypoint and track their trajectories.

Video Slicing based on Camera Boundary Detection and Action Semantic. The original videos have different attributes, such as lengths and resolution, with camera jumps or editing traces, and a video may contain multiple action semantics. Thus, we slice the video into clips with normalized attributes, consistent motion, no camera jumps, and single action semantics. Specifically, we use TransNetV2 [35] to perform camera boundary detection and use Intern-VL [36] to generate the action semantic of a specific clip.

Clip Filtering based on Key Frame and Clip Quality. The generated clips are highly heterogeneous in terms of quality, semantics, and presentation form. To ensure the validity and adaptability of the training data, we introduce a clip filtering mechanism including: (1) using FlowNet [37] to filter out clips with indistinct motion and disordered movement patterns, and (2) using Intern-VL [36] to label the key frame of the clip and filter out the frames without explicit relationship to the key frame.

Clip Categorization based on Action Difficulty and Scenes. In this stage, we categorize and reorganize the dataset based on action difficulty and clip scenes to support the curriculum learning strategy [38], which trains the world model from easier to harder tasks.

2.3 RoboScape: A Physics-informed Embodied World Model

RoboScape is designed to achieve frame-level action-controllable robot video generation, enabling interactive future frame prediction. At its core, we adopt an auto-regressive Transformer-based framework that iteratively predicts the next frame based on historical frames and the current robot

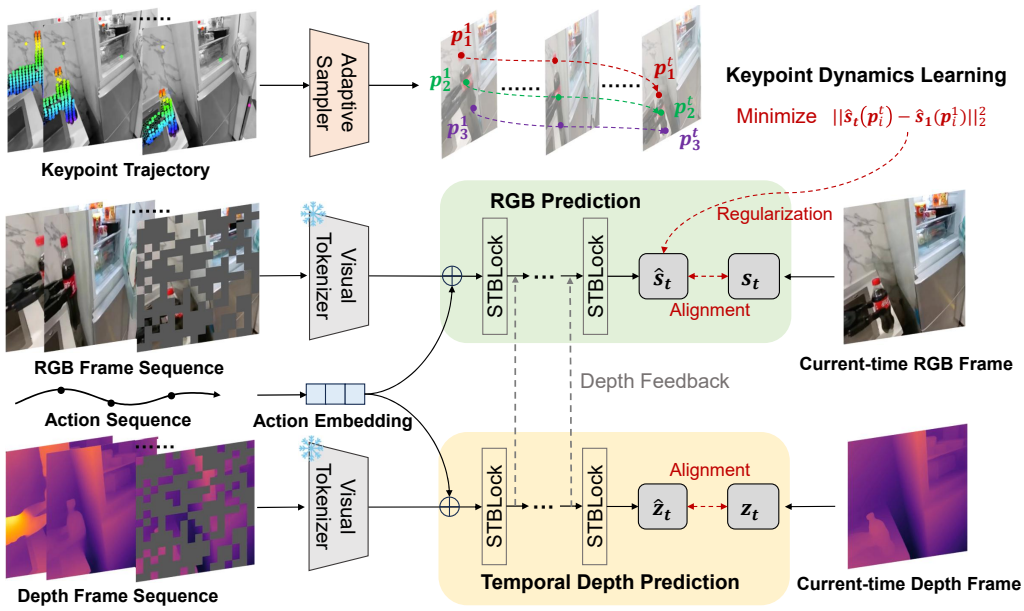


Figure 2: Overview of the physics-informed world model, where physical knowledge is integrated through joint learning of temporal depth estimation and adaptively sampled keypoint dynamics.

action. To enhance the physical plausibility of generated videos, we introduce two physics-informed auxiliary training tasks in addition to the normal RGB image prediction: (1) temporal depth prediction, which encourages global geometric consistency across frames, and (2) adaptively sampled keypoint dynamics learning, which captures the motion and deformation details of local dynamic objects. The whole pipeline is illustrated in Figure 2. Joint training with these physics-aware regularizers provides an efficient approach to embed physical priors into world models, significantly reducing the reality gap between generated videos and real-world dynamics.

Video Tokenization. To enable efficient video generation, we leverage MAGVIT-2 [39] to compress raw RGB frames $\mathbf{o}_{1:T} \in \mathbb{R}^{T\times H\times W\times 3}$ into discrete latent tokens $\mathbf{s}_{1:T} \in \mathbb{R}^{T\times H'\times W'\times D}$, where $H' = H/\alpha$ and $W' = W/\alpha$ denote the reduced spatial dimensions (α being the downsampling factor), and D represents the latent channel dimension. Similarly, we tokenize temporal depth maps $\mathbf{d}_{1:T} \in \mathbb{R}^{T\times H\times W\times 1}$ into latent depth tokens $\mathbf{z}_{1:T} \in \mathbb{R}^{T\times H'\times W'\times D}$.

Geometry Consistency Enhancement via Temporal Depth Prediction. While RGB-based video generation has achieved remarkable progress, it often suffers from inconsistent 3D geometry due to the lack of explicit spatial constraints. Considering that inter-frame depth variations encode crucial 3D structure information, we propose to jointly learn temporal RGB and depth information, leveraging depth features as geometric constraints to ensure spatially coherent video generation. For uniform modeling, we also utilize MAGVIT-2 as the visual encoder for depth maps. We convert depth maps to a three-channel RGB format to ensure compatibility with MAGVIT-2. For joint prediction of both RGB and depth images, we propose a dual-branch co-autoregressive Transformer (DCT). Each branch consists of a stack of 32 spatial-temporal Transformer (ST-Transformer) blocks, which implement a causal attention mechanism in the temporal attention layers for generation causality, and bidirectional attention in the spatial attention layers to enable full context modeling.

At timestep t, the model processes historical latent tokens through parallel branches $\mathcal{F}_{\text{RGB}}$ and $\mathcal{F}_{\text{Depth}}$, conditioned on learned action embeddings $\mathbf{c}_{1:t-1} \in \mathbb{R}^{(t-1)\times 1\times 1\times D} : \mathcal{E}_a(\mathbf{a}_{1:t-1})$ and position embeddings $\mathbf{e}_{1:t-1} \in \mathbb{R}^{(t-1)\times H'\times W'\times D}$, where $\mathcal{E}_a$ denotes the robot action encoder. The auto-regressive prediction of each branch is formulated as:

$$\begin{aligned} \hat{\mathbf{s}}_t &= \mathcal{F}_{\text{RGB}}(\mathbf{s}_{1:t-1} \oplus \mathbf{c}_{1:t-1} \oplus \mathbf{e}_{1:t-1}), \\ \hat{\mathbf{z}}_t &= \mathcal{F}_{\text{Depth}}(\mathbf{z}_{1:t-1} \oplus \mathbf{c}_{1:t-1} \oplus \mathbf{e}_{1:t-1}), \end{aligned} \tag{2}$$

where $\oplus$ denotes element-wise addition with broadcasting. Empirically, we find that simple additive fusion provides effective action control while maintaining model efficiency.

To inject depth predictions as physical priors into the RGB branch and enhance spatial structure fidelity of rendered videos, we introduce cross-branch interaction pathways. Specifically, at each ST-Transformer block l, we project the depth branch's intermediate features $\mathbf{h}^l_{\text{depth}}$ and fuse them additively with the corresponding RGB features:

$$\mathbf{h}^l_{\text{RGB}} = \mathbf{h}^l_{\text{RGB}} + \mathcal{W}^l(\mathbf{h}^l_{\text{depth}}), \tag{3}$$

where $\mathcal{W}^l$ is a learnable linear projection layer. This hierarchical feature fusion enables the RGB branch to maintain precise geometric structure while generating photorealistic video frames. Both RGB and depth branches are optimized using the cross-entropy loss of tokens:

$$\mathcal{L}_{\text{RGB}} = -\sum_{t=1}^{T} \mathbf{s}_t \log p(\hat{\mathbf{s}}_t), \quad \mathcal{L}_{\text{Depth}} = -\sum_{t=1}^{T} \mathbf{z}_t \log p(\hat{\mathbf{z}}_t). \tag{4}$$

Implicit Material Understanding via Keypoint Dynamics Learning. Modeling physically plausible object deformations and motions in robot manipulation scenarios remains challenging for RGB-based world models, as material properties (e.g., rigidity, elasticity) cannot be effectively learned through RGB pixel fitting alone. While physics engines provide accurate simulations, their computational expense and scene-specific constraints limit practical applicability. To tackle this, we propose a keypoint-induced material learning approach, with the insight that physical material understanding can emerge from self-supervised tracking of contact-driven keypoint dynamics. For example, when a robot places an apple into a plastic bag, accurately capturing the motion of keypoints on the deforming bag implicitly captures the material properties. This method can be integrated naturally with video generation frameworks while maintaining strong generalization capabilities.

Specifically, for each video $\mathcal{V}$, we utilize SpatialTracker [34] to densely sample N_0 keypoints in the initial frame and track their temporal coordinate trajectories across T frames, yielding $\mathcal{T}_{dense} = \{(\mathbf{p}_i^1, ..., \mathbf{p}_i^T)\}_{i=1}^{N_0}$, where the element $\mathbf{p}_i^t \in \mathbb{R}^2$ represents its coordinates in the tokenized feature map of frame t. Rather than relying on costly segmentation masks to identify contact regions and guide keypoint sampling, we observe that the most informative keypoints are empirically characterized by large motion magnitudes. Thus, we adaptively select the top-K most active keypoints based on their motion magnitudes $\mathcal{M}_i = \sum_{t=1}^{T-1} ||\mathbf{p}_i^{t+1} - \mathbf{p}_i^t||_2, \forall i \in 1, ..., N_0$, producing the sampled trajectory set $\mathcal{T}_{sample} = \{(\mathbf{p}_i^1, ..., \mathbf{p}_i^T)\}_{i=1}^{K}$.

To enhance the keypoint dynamic learning, we encourage temporal consistency between the visual tokens of sampled keypoints by aligning all frames to the initial frame ($t = 1$) through the following loss:

$$\mathcal{L}_{\text{Keypoint}} = \frac{1}{(T-1)K} \sum_{i=1}^{K} \sum_{t=2}^{T} \|\hat{\mathbf{s}}_t(\mathbf{p}_i^t) - \hat{\mathbf{s}}_1(\mathbf{p}_i^1)\|_2^2, \tag{5}$$

where $\hat{\mathbf{s}}_t(\mathbf{p}_i^t) \in \mathbb{R}^D$ denotes the i-th keypoint-located predicted token at frame t.

Furthermore, we observe that these dynamically active keypoint regions often exhibit higher token errors due to their complex motion patterns. To address this, we propose a keypoint-guided attention mechanism that adaptively enhances token learning in regions intersected by keypoint trajectories. Specifically, we compute a spatiotemporal attention map $\mathbf{A} \in \mathbb{R}^{T \times H' \times W'}$, with each element defined as:

$$\mathbf{A}_{t,x,y} = \begin{cases} \gamma & \text{if } (t, x, y) \in \mathcal{T}_{sample}, \\ 1 & \text{otherwise}, \end{cases} \tag{6}$$

where γ is a hyperparameter controlling the importance weight. The attention-augmented training objective is formulated as:

$$\mathcal{L}_{\text{Attention}} = -\sum_{t=1}^{T} \mathbf{A}_t \odot \mathbf{s}_t \log p(\hat{\mathbf{s}}_t). \tag{7}$$

Physics-informed Joint Training Objectives. By integrating the above designs, we train a unified physics-aware world model through multi-task learning, with the final objective formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{RGB}} + \lambda_1 \mathcal{L}_{\text{Depth}} + \lambda_2 \mathcal{L}_{\text{Keypoint}} + \lambda_3 \mathcal{L}_{\text{Attention}}, \tag{8}$$

where $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}^+$ are are tunable coefficients balancing the loss terms.

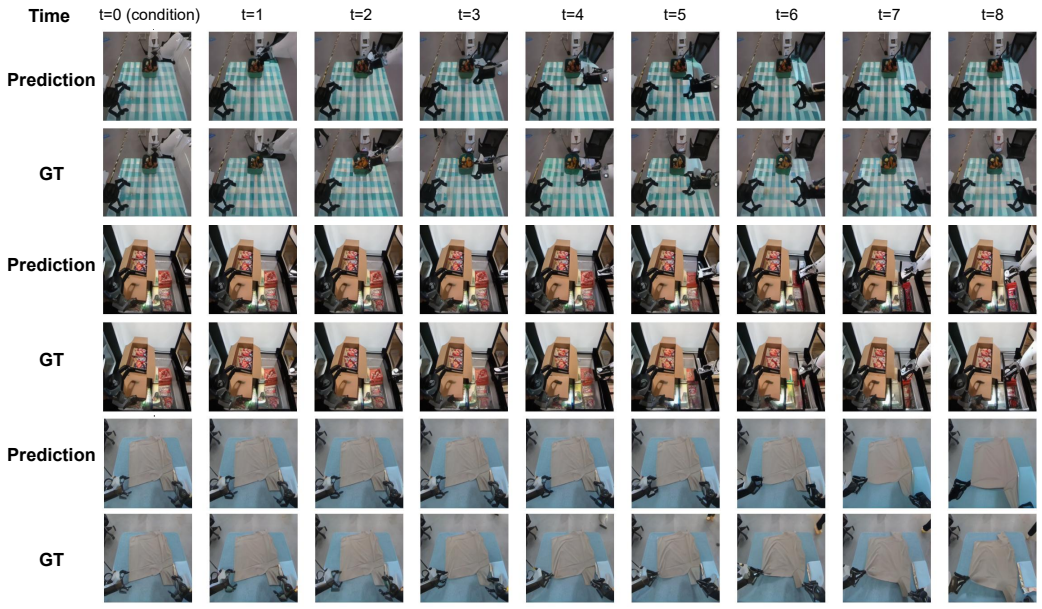


Figure 3: Qualitative results visualization of our model (only the subsequent 8 frames are shown). More results can be found in the appendix.

3 Experiments

In this section, we begin by detailing our experimental protocol (Section 3.1), including the dataset statistics, baseline information, and the implementation of our model. We then evaluate our model from three aspects: video quality evaluation (Section 3.2), robot policy learning with synthetic data (Section 3.3), and robotic policy evaluation (Section 3.4).

3.1 Experimental Settings

Dataset Statistics. In our experiment, we use 50,000 videos extracted from the AgiBotWorld-Beta dataset [40], covering 147 tasks and 72 skills. We concatenate the end position, end orientation, and effector position of the embodiment as the action sequence. Our dataset comprises approximately 6.5M training clips and 1.2K test clips.

Baselines. We compare our model with four advanced baselines, including both embodied world models (IRASim [14] and iVideoGPT [13]) and general world models (Genie [41] and CogVideoX [42]). Due to unavailable training codes in some recent works [29, 30], these methods are excluded from direct comparison. Details of baselines are presented in the appendix.

Implementation Details. We preprocess videos by extracting 16-frame clips sampled at 2Hz, yielding approximately 6.5 million training clips. The model is trained for 5 epochs using the following hyperparameters: $\lambda_1 = 1$, $\lambda_2 = 0.01$, $\lambda_3 = 1$, and $\gamma = 5$. Training completes in approximately 24 hours on a cluster of 32 NVIDIA A800-SXM4-80GB GPUs. During inference, we use the first frame as a conditional input to autoregressively predict the subsequent 15 frames.

3.2 Video Quality Evaluation

We evaluate video generation quality through three key dimensions: appearance fidelity, geometric consistency and action controllability. The details of the six used metrics are as follows:

- **PSNR**: It measures pixel-level reconstruction accuracy between generated and ground-truth frames.
- **LPIPS**: It assesses perceptual quality using visual feature similarity.
- **AbsRel**: It computes relative depth estimation errors.
- δ_1/δ_2: They evaluate depth prediction accuracy at different precision levels.

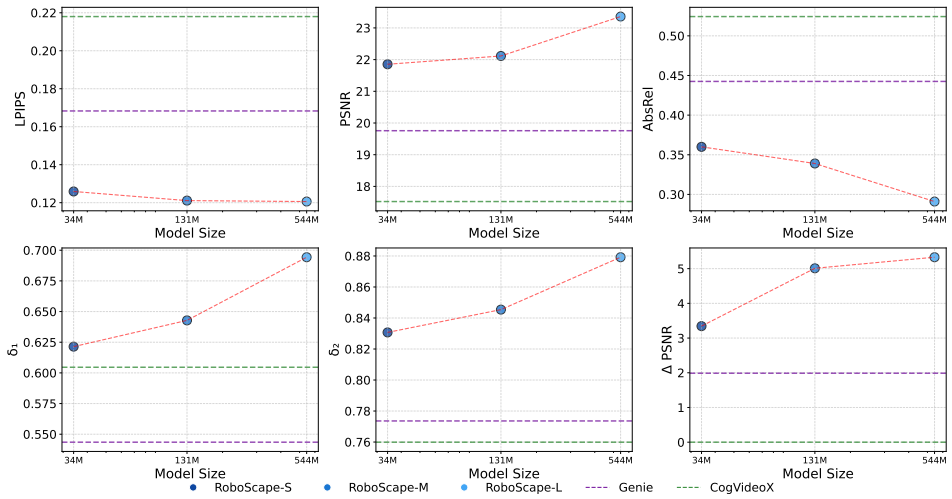


Figure 4: Model scaling law of RoboScape.

Table 1: Quantitative comparison of our model and baselines with 5 independent runs.

Method	Appearance Fidelity		Geometric Consistency			Action Controllability
	LPIPS (↓)	PSNR (↑)	AbsRel (↓)	δ_1 (↑)	δ_2 (↑)	ΔPSNR (↑)
IRASim	0.663±0.005	11.382±0.122	0.641±0.030	0.508±0.014	0.701±0.026	0.160±0.122
iVideoGPT	0.501±0.003	16.234±0.038	0.749±0.018	0.364±0.012	0.586±0.010	0.148±0.067
Genie	0.165±0.002	19.875±0.064	0.446±0.014	0.545±0.011	0.778±0.016	1.868±0.090
CogVideoX	0.202±0.008	17.957±0.153	0.508±0.026	0.594±0.017	0.739±0.016	—
RoboScape	**0.128±0.002**	**21.730±0.120**	**0.378±0.026**	**0.608±0.011**	**0.814±0.013**	**3.442±0.139**

- **ΔPSNR**: It quantifies output sensitivity to action condition, with higher values indicating better action control ability.

We present some generation results in Figure 3, where we predict future frames conditioned on an initial frame and robot action commands (we visualize 8 frames while the model supports long-horizon rollouts). The visualizations demonstrate that our model effectively simulates realistic robot manipulation scenarios, with generated sequences showing strong similarity to ground truth observations. Notably, our approach successfully handles deformable object interactions, as evidenced by the cloth-dragging sequence where the generated deformations accurately follow physical laws and capture material properties.

As shown in Table 1, we conduct comprehensive comparisons with four advanced baselines: two embodied world models (IRASim and iVideoGPT) and two general world models (Genie and CogVideoX). Our model consistently outperforms all baselines across six evaluation metrics, demonstrating its superior capability in video prediction for robotic scenarios. Detailed analysis reveals that while CogVideoX can generate high-quality videos, its inability to follow action commands leads to substantial deviations in future frames. The two embodied world models are not good at motion learning when conducting long-term generation, thus receiving poor metrics. Our model's novel integration of keypoint dynamics learning effectively addresses these limitations, simultaneously achieving high-fidelity visual generation and superior action controllability.

We further conduct ablation studies to demonstrate the complementary benefits of our two core components: temporal depth learning and keypoint dynamics learning. The results are shown in Table 2. The quantitative results reveal that both components contribute significantly to overall performance; removing either one leads to measurable degradation across different metrics. The depth learning primarily preserves geometric consistency of moving objects, and the keypoint learning proves essential for maintaining both visual fidelity and action controllability. We provide a case study in Figure 6. It can be seen that the missing of temporal depth learning will lead to geometric distortions in moving objects, while the absence of key-point dynamics learning results in unreal motion patterns. These findings collectively validate the necessity of our key designs.

Table 2: Ablation study of our key designs of physics prior injection with 5 independent runs.

Method	LPIPS (↓)	PSNR (↑)	AbsRel (↓)	δ_1 (↑)	δ_2 (↑)	ΔPSNR (↑)
whole model	0.128±0.002	21.730±0.120	0.378±0.026	0.608±0.011	0.814±0.013	3.442±0.139
w/o depth	0.126±0.001	21.885±0.046	0.408±0.010	0.560±0.012	0.789±0.022	3.514±0.023
w/o keypoint	0.128±0.001	21.634±0.043	0.346±0.012	0.637±0.012	0.848±0.012	2.953±0.036
w/o depth & keypoint	0.130±0.001	21.477±0.029	0.371±0.012	0.598±0.018	0.800±0.009	1.945±0.054

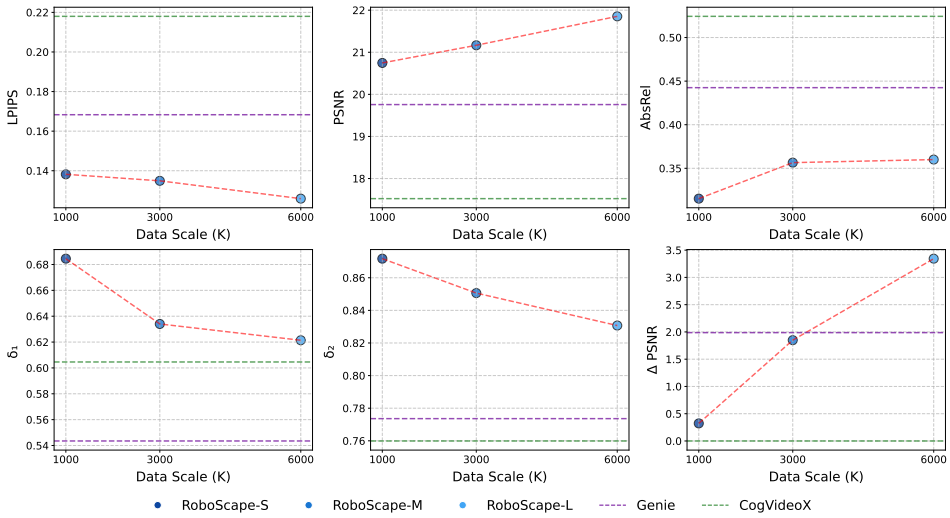


Figure 5: Data scaling law of RoboScape-S.

We also investigate the scaling behavior of RoboScape in terms of both model and data scales. As shown in Figure 4, we evaluate three model variants—RoboScape-S (34M), RoboScape-M (131M), and RoboScape-L (544M)—and observe a clear scaling law: all six evaluation metrics improve significantly as model capacity increases. In addition, we study the impact of data scale by training RoboScape-S on 1,000K, 3,000K, and 6,000K clips (Figure 5). While increasing data size consistently enhances visual quality and action controllability, geometric accuracy exhibits marginal improvement or even slight degradation. We find that this is because smaller datasets encourage overfitting to the final frame of conditional inputs, artificially inflating geometric metrics without generating meaningful temporal dynamics. Despite this, the overall trend confirms that more training data leads to better model performance.

3.3 Robotic Policy Learning with Synthetic Data

We validate our world model's utility by generating synthetic robotic video data for downstream policy learning based on Diffusion Policy (DP) [31] and $\pi 0$ [32]. The actions and initial observations for our synthetic data are directly drawn from the raw RoboMimic and LIBERO datasets. Through controlled experiments with progressively adding synthetic data, we systematically measure the impact of generated data on policy learning performance. The results are shown in Table 3. "Real data" mentioned consists of original videos with their corresponding action annotations from the raw dataset. For "Synthetic Data", we use our world model to generate the videos based on actions and initial observations.

In the experiments on the Robomimic Lift task [43], DP trained for 10k steps with only generated data achieved nearly the same performance as DP trained with real data. Notably, the policy success rate exhibited consistent improvement with increasing synthetic training data, highlighting the effectiveness of our model. We further validated our approach using the π_0 [32] model on the challenging LIBERO [44] task suite. These tasks present three key challenges beyond the Robomimic Lift environment: (1) complex multi-object manipulation requirements, (2) cluttered scene configurations, and (3) extended action sequence horizons. Therefore, we employ a small amount of real data (200 trajectories) as a training warm-up. Remarkably, when training π_0 policies with increasing generated data, the model performance achieves gradual improvement. These results

Table 3: Results of policy learning with DP on Robomimic task and $\pi 0$ on LIBERO tasks.

DP on Robomimic tasks	
# Synthetic Data	Success Rate
50	40%
100	77%
150	84%
200	91%
Real (200)	92%

$\pi 0$ on LIBERO tasks					
#Synthetic Data	Spatial	Object	Goal	10	Average
200	77.6%	81.8%	71.0%	36.0%	66.6%
400	79.4%	85.2%	74.6%	46.2%	71.4%
600	81.6%	86.0%	78.0%	51.8%	74.4%
800	84.6%	89.0%	82.8%	60.0%	79.1%
Real (200)	77.2%	79.8%	68.8%	34.8%	65.2%

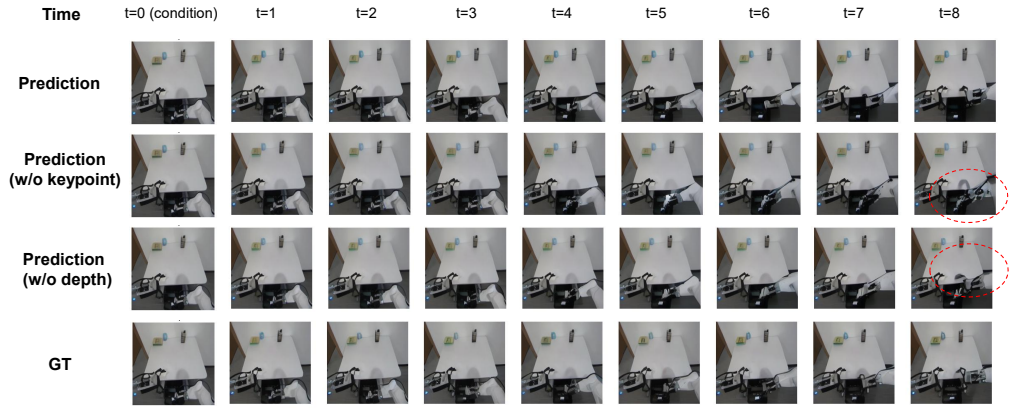


Figure 6: Effect of the physics knowledge learning. Omission of temporal depth learning leads to geometric distortions in moving objects, while the absence of key-point dynamics learning results in unreal motion patterns.

demonstrate our model's capability to generate physically plausible trajectories even for demanding, long-horizon manipulation scenarios.

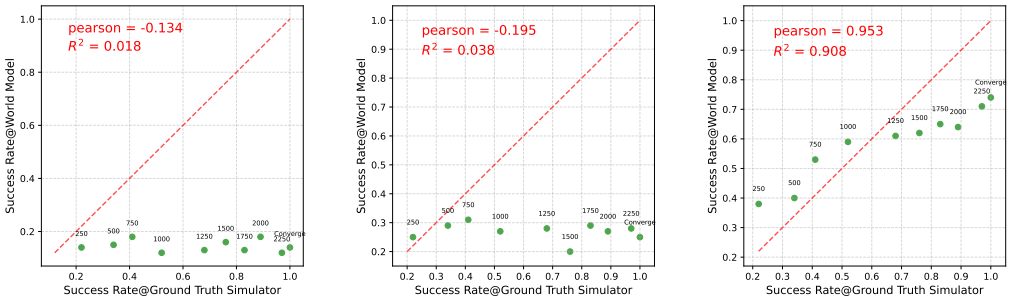
3.4 Robotic Policy Evaluation

In this section, we investigate whether our world model can act as a policy evaluator for different robotic policies. In policy evaluation, the world model acts as an environment that receives policy-generated action sequences and predicts subsequent observations in a rollout manner. Policy quality is then assessed by checking success rates in the predicted videos. Here we compare IRASim, iVideoGPT, and our model as the policy evaluator and use Diffusion Policy [31] as the policy model. Specifically, we train the policy on the Robomimic Lift task [43] using 200 trajectories and save the policy every 250 epochs until it is fully converged. Then we post-train the world model and evaluate the policy in both the ground-truth simulator and the world model by 100 runs. The success signal of each run can be directly given by the simulator, while it requires manual judgment when the policy interacts with the world model. The generated videos from all models were presented to participants in a randomized, blind order, with no model identifiers displayed. Participants were instructed to judge whether the task depicted in each video was successfully completed. This setup ensures objectivity and impartiality. Afterwards, we calculate the Pearson correlation and R^2 between different world models and the ground-truth simulator. The results in Figure 7 show that the Pearson correlation of our model is **0.953**, while the correlation of other models is rather low, indicating that our world model can be utilized as a better policy evaluator.

4 Related Work

4.1 World Model

World models learn representations of environmental states through neural networks, enabling the prediction of future states based on current observations and actions [45]. Recent advances in world models primarily leverage video generation techniques, with applications spanning three key domains including autonomous driving [46, 47, 48, 49, 50, 51, 52], embodied intelligence [53, 14, 15], and



(a) IRASim (b) iVideoGPT (c) RoboScape

Figure 7: Correlation between the success rate of different world models and the ground-truth simulator. Each point represents a policy, and the trained epochs are shown above the point.

gaming [54, 55, 56, 48]. The dominant modeling approaches fall into two main categories: diffusion models and autoregressive models. Diffusion models, such as DiT [57], generate sequences through a gradual denoising process and are well-suited for producing diverse and short-term consistent visual content. Autoregressive models, such as Genie[41], reconstruct sequences via masking mechanisms and demonstrate superior efficiency and controllability. Compared to diffusion-based approaches, autoregressive methods offer advantages in inference speed and training stability. Our work utilizes masked autoregressive models with physical information injection, aiming to build efficient and interactive world models for embodied intelligence.

4.2 Physics-aware Generative Model

Recent advances in video generation have increasingly focused on improving the modeling of physical properties [17]. Current methods in this area can be roughly divided into explicit and implicit physical modeling. Explicit methods incorporate physical information by learning explicit textures and material representations [28, 30]. In contrast, implicit methods mainly embed physical knowledge into models via training loss terms [20], or by using generative models to jointly generate RGB videos and other physical representations [29, 58]. These approaches aim to enhance physical understanding through data-driven approaches rather than predefined physical rules. Currently, there's much room for existing embodied world models to enhance the integration of physical knowledge into video generation. To advance this field, we introduce a physics-informed embodied world model that jointly learns RGB video generation, temporal depth prediction, and keypoint dynamics within a unified framework, achieving both high visual fidelity and physical plausibility.

5 Conclusion and Future Work

In this work, we propose RoboScape, a physics-informed embodied world model that efficiently integrates physical knowledge into video generation through a physics-inspired multi-task joint training framework, eliminating the need for cascaded external models such as physics engines. By incorporating temporal depth prediction, our model learns the 3D geometric structure of scenes, while dynamic keypoint learning enables implicit modeling of object deformation and motion patterns. Extensive evaluations demonstrate that our approach outperforms baseline methods in video generation quality, synthetic data utility for downstream robotic manipulation policy training, and effectiveness as a policy evaluator. In the future, we plan to combine the generative world model with real-world robots to test performance further.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Key Research and Development Program of China (No.2024YFC3307603).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[3] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[4] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[5] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[6] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

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

[8] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023.

[9] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024.

[10] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

[11] Yu Shang, Yinzhou Tang, Xin Zhang, Shengyuan Wang, Yuwei Yan, Honglin Zhang, Zhiheng Zheng, Jie Zhao, Jie Feng, Chen Gao, et al. A survey of embodied world models.

[12] Lirui Wang, Kevin Zhao, Chaoqi Liu, and Xinlei Chen. Learning re al-world action-video dynamics with heterogeneous masked autoregression. *arXiv preprint arXiv:2502.04296*, 2025.

[13] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024.

[14] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2406.14540*, 2024.

[15] 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.

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

[17] Minghui Lin, Xiang Wang, Yishan Wang, Shu Wang, Fengqi Dai, Pengxiang Ding, Cunxiang Wang, Zhengrong Zuo, Nong Sang, Siteng Huang, et al. Exploring the evolution of physics cognition in video generation: A survey. *arXiv preprint arXiv:2503.21765*, 2025.

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

[19] 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.

[20] Yunuo Chen, Junli Cao, Anil Kag, Vidit Goel, Sergei Korolev, Chenfanfu Jiang, Sergey Tulyakov, and Jian Ren. Towards physical understanding in video generation: A 3d point regularization approach. *arXiv preprint arXiv:2502.03639*, 2025.

[21] Gaurav Rai and Ojaswa Sharma. Enhancing sketch animation: Text-to-video diffusion models with temporal consistency and rigidity constraints. *arXiv preprint arXiv:2411.19381*, 2024.

[22] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809. IEEE, 2024.

[23] Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang. Physanimator: Physics-guided generative cartoon animation. *arXiv preprint arXiv:2501.16550*, 2025.

[24] Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1430–1440, 2024.

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

[26] Xiyang Tan, Ying Jiang, Xuan Li, Zeshun Zong, Tianyi Xie, Yin Yang, and Chenfanfu Jiang. Physmotion: Physics-grounded dynamics from a single image. *arXiv preprint arXiv:2411.17189*, 2024.

[27] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pages 388–406. Springer, 2024.

[28] Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3d: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024.

[29] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025.

[30] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: Learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

[31] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[32] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[33] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv preprint arXiv:2501.12375*, 2025.

[34] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024.

[35] Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11218–11221, 2024.

[36] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[37] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[38] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021.

[39] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[40] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkang Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi Zhao, and Jianchao Zhu. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[41] 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.

[42] 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.

[43] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.

[44] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

[45] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[46] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024.

[47] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pages 87–104. Springer, 2024.

[48] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. *arXiv preprint arXiv:2411.00769*, 2024.

[49] Jingcheng Ni, Yuxin Guo, Yichen Liu, Rui Chen, Lewei Lu, and Zehuan Wu. Maskgwm: A generalizable driving world model with video mask reconstruction. *arXiv preprint arXiv:2502.11663*, 2025.

[50] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024.

[51] 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.

[52] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

[53] Chaoqi Liu Xinlei Chen Lirui Wang, Kevin Zhao. Learning robotic video dynamics with heterogeneous masked autoregression. In *Arxiv*, 2025.

[54] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025.

[55] Haoran He, Yang Zhang, Liang Lin, Zhongwen Xu, and Ling Pan. Pre-trained video generative models as world simulators. *arXiv preprint arXiv:2502.07825*, 2025.

[56] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

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

[58] Dingkang Liang, Dingyuan Zhang, Xin Zhou, Sifan Tu, Tianrui Feng, Xiaofan Li, Yumeng Zhang, Mingyang Du, Xiao Tan, and Xiang Bai. Seeing the future, perceiving the future: A unified driving world model for future generation and perception. *arXiv preprint arXiv:2503.13587*, 2025.

[59] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper proposes a physics-informed embodied world model, which is reflected in the methodology and experiment validation part.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This work does not invlove theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide implementation details in the experiment part.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide an anonymous code repository in the abstract.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See the "implementation details" section in the main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide independent repeated experimental results in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See the "implementation details" section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our video generation experiment, in every respect, follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have added citations and footnotes to note the source of used codes and datasets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLM usage.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Broader Impacts

Our physics-informed world model offers positive impacts by advancing robotic learning: it generates high-fidelity synthetic data with inherent physical plausibility, which drastically reduces reliance on costly real-world data collection and improves sim-to-real transfer for applications in healthcare, disaster response, and industry. While this computational efficiency lowers research barriers, the high fidelity necessitates ethical consideration. Specifically, the generated content could be misused to create deepfakes or misrepresent safe procedures. To mitigate these risks, future work should focus on training fake detection models to identify synthetic content and incorporating anomaly detection mechanisms to flag or prevent the generation of videos depicting unsafe or non-standard robot behaviors, ensuring responsible deployment in safety-critical domains.

B Limitations

While our physics-informed world model achieves significant gains in physical consistency, we acknowledge the following shortcomings and inherent assumptions: (1) Our current physical constraints focus primarily on geometric (depth) and kinematic (keypoint dynamics) properties. This leaves out broader critical information such as dynamic force interactions. Furthermore, our model relies on the empirical assumption that keypoint dynamics implicitly capture material properties, which lacks a theoretical guarantee. (2) Our model's inference is currently limited to a 48-frame rollout. This computational constraint makes it challenging to simulate and evaluate very long-horizon tasks (e.g., complete cloth folding), where we have observed that cumulative error can lead to physically implausible outcomes. (3) Generalization to more diverse embodiments, such as quadrupeds or humanoids, is currently limited by the lack of readily available, high-quality, multi-domain datasets. Our current validation is primarily confined to single- and dual-arm manipulation settings. Future work will be dedicated to overcoming these limitations by incorporating richer physical laws, exploring more efficient hybrid architectures, and expanding our training to more diverse embodiment data.

C Baseline Details

We provide details of the compared baselines as follows:

- **IRASim**: A DiT-based robotic video generation model, capable of generating videos conditioned on robot actions and trajectories.
- **iVideoGPT**: An auto-regressive interactive world model that takes the current video frame observation and action as input to predict the next frame while simultaneously estimating the reward signal for robotic operations.
- **Genie**: A foundation world model trained through unsupervised learning on massive video data. We implement it with a reproduced open-source repository [*].
- **CogVideoX**: An advanced DiT-based text-to-video generation framework, with superior performance in prompt-driven video generation.

D Supplemented Evaluation toward Physics Correctness

In this section, we provide a comparison regarding the physical correctness of our model and baselines. While existing physics benchmarks for video generation are primarily designed for general text-to-video models and aren't directly applicable to our embodied world model (which is driven by robotic actions), we've identified that key metrics from these benchmarks can be utilized for our evaluation. Specifically, we utilize four metrics from the Physics-IQ benchmark [59] that are particularly effective in assessing the physical realism of motion plausibility and object deformation, detailed as follows:

- **Spatial IoU**: Measures "Whether the location where an action happens is correct".
- **Spatiotemporal IoU**: Measures "Where does action happen and whether it occurs at the right time".

[*] https://github.com/1x-technologies/1xgpt

Table 4: Comparison of world models on physics correctness.

Model	Spatial IoU (↑)	Spatiotemporal IoU (↑)	Weighted Spatial IoU (↑)	MSE (↓)
IRASim	0.2431	0.1255	0.2081	0.0655
Genie	0.6429	0.3420	0.6082	0.0397
CogVideoX	0.6523	0.2058	0.4675	0.0943
RoboScape	**0.7573**	**0.4454**	**0.7023**	**0.0184**

Table 5: Policy success rate using generated data from different world models on LIBERO benchmark.

Model	LIBERO-Spatial	LIBERO-Object	LIBERO-Goal	LIBERO-10	Average
Only real data	77.2%	79.8%	68.8%	34.8%	65.2%
IRASim	72.8%	77.4%	75.2%	34.0%	64.8%
CogVideoX	81.0%	79.6%	74.4%	44.2%	69.8%
RoboScape	**84.6%**	**89.0%**	**82.8%**	**60.0%**	**79.1%**

- **Weighted Spatial IoU**: Measures "Where and how much action happens".
- **MSE**: Measures "How objects look and interact" at pixel-level.

  The results shown in Table 4 demonstrate our model's superior performance in terms of physical correctness.

E Supplemented Results of Policy Learning

We compare training the pi0 model using data synthesized by baseline world models and our model. For fair comparisons, both CogVideoX and IraSim were fine-tuned on the LIBERO dataset, with each model generating 800 synthetic data points. Table 5 presents the experimental results across various task subsets in the LIBERO environment, indicating the superiority of our model in generating synthetic data for VLA training compared to baselines.

F More Visualization Results

F.1 Video Generation Results

We provide more visualization results of generated videos using our model, as illustrated in Figure 8.

F.2 Robotic Policy Learning

We provide some visualization results of generated data on Robomimic and LIBERO using our model, which are shown in Figure 9 and Figure 10.

F.3 Robotic Policy Evaluation (add visualization results of our model and baselines

In this part, we provide visualization results of RoboScape and other baselines in policy evaluation. The failure cases are presented in Figure 11 while the successful cases are shown in Figure 12.

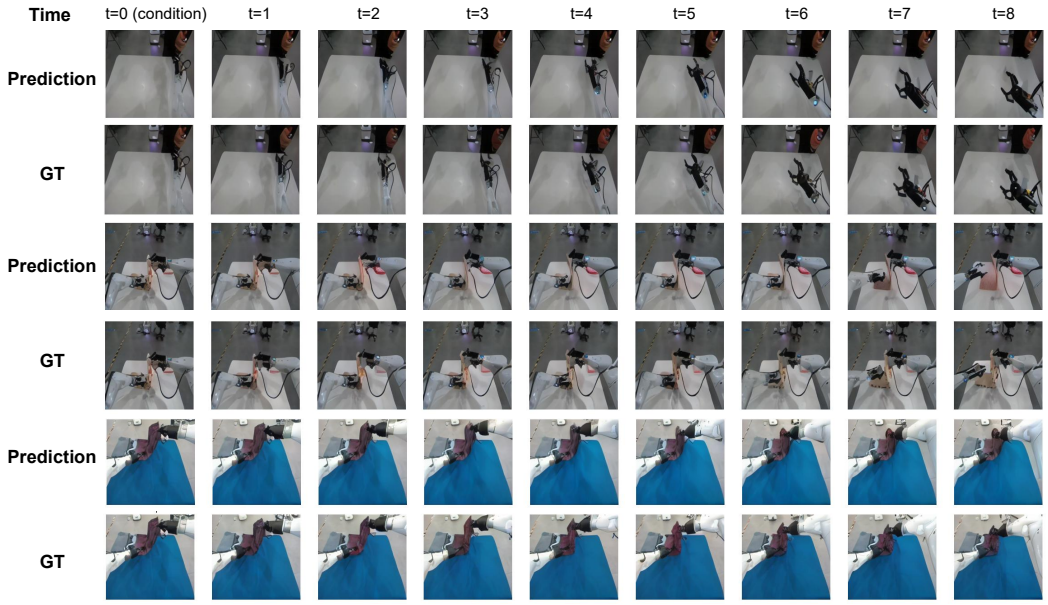


Figure 8: Supplemented visualization results from our model (only the subsequent 8 frames are shown).

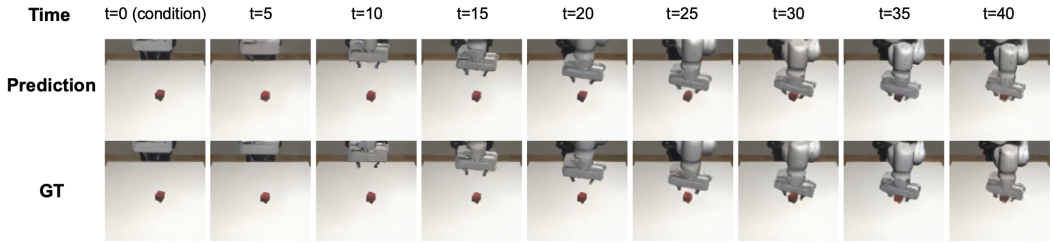


Figure 9: Supplemented visualization results on Robomimic (displaying every 5th frame; 8 frames shown from t=0 to t=40).

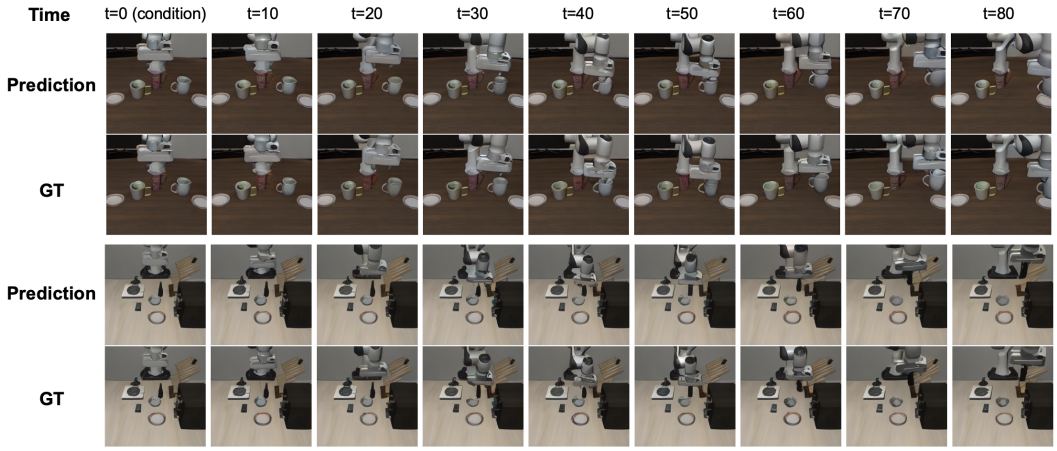


Figure 10: Supplemented visualization results on LIBERO (displaying every 10th frame; 8 frames shown from t=0 to t=80).

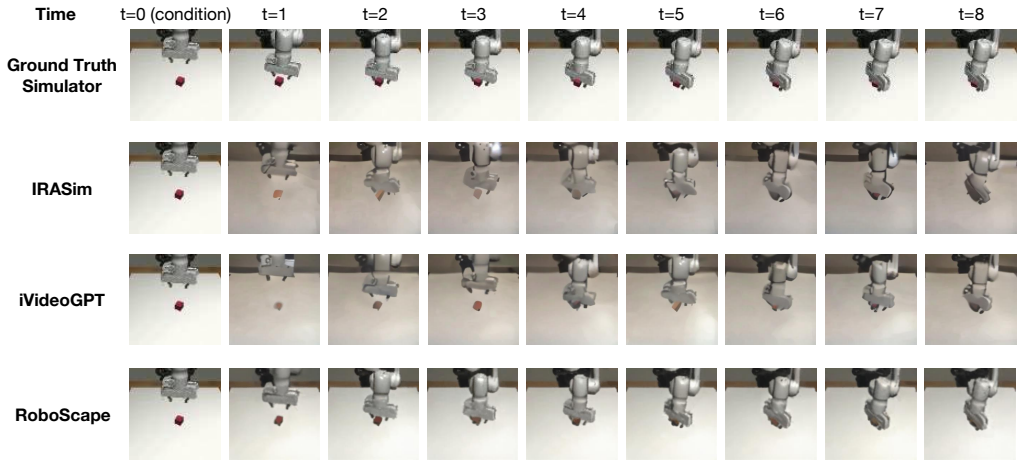


Figure 11: Supplemented visualization results of failure cases in policy evaluation.

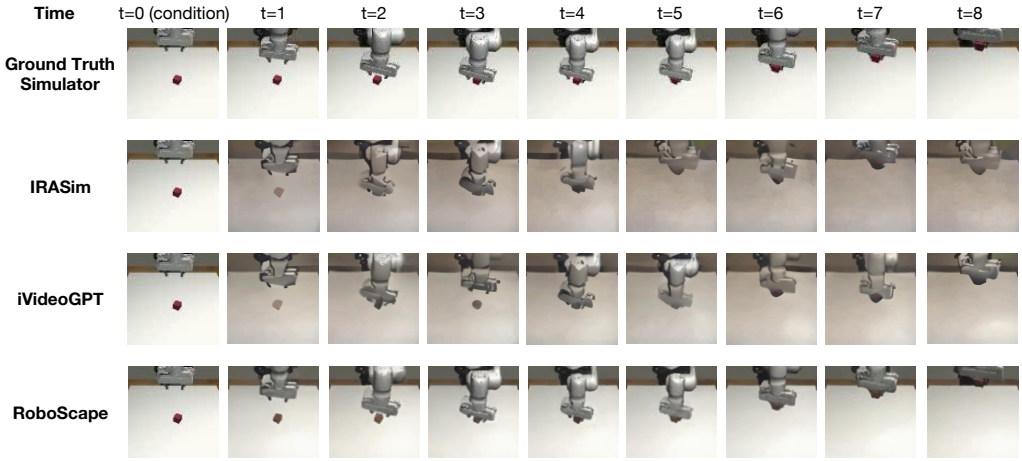


Figure 12: Supplemented visualization results of successful cases in policy evaluation.