Worldscape-MoE: A Unified Mixture-of-Experts Architecture for Scalable Multi-Control Video Generation World Modeling

Jianjie Fang[1,*], Yongyan Xu[3,*], Ziyou Wang[1,*], Chen Gao[1,*,†]
Yuchao Huang[4], Zhaolu Wang[5], Rongze Tang[5], Mingyuan Jia[1]
Baining Zhao[1], Weichen Zhang[1], Xin Zhang[2], Haisheng Su[2]
Yu Shang[1], Wei Wu[2], Xinlei Chen[1,†], Yong Li[1,†]

[1]Tsinghua University [2]Manifold AI [3]Peking University
[4]Northeastern University [5]University of Science and Technology of China

chgao96@gmail.com, chen.xinlei@sz.tsinghua.edu.cn, liyong07@tsinghua.edu.cn
[*]Equal contribution. [†]Corresponding authors.

Abstract

With the rapid progress of interactive video generation, video generation world models have gradually emerged as one of the mainstream paradigms in world model research and are increasingly regarded as a promising path toward efficient intelligent agents. However, existing video generation world models are typically developed under different forms of control supervision and mainly focus either on interactive world modeling or embodied world modeling, leaving the compatibility of heterogeneous control signals largely unexplored. In this work, we introduce **Worldscape-MoE**, the first training framework that enables unified learning of world models under heterogeneous supervisory controls by incorporating a Mixture-of-Experts (MoE) design into Diffusion Transformers (DiT). We further propose **Worldscape-MoE Tuning**, a continually extensible heterogeneous training strategy for world models, which supports diverse control signals, including robotic arms, hand joints, and camera poses, within a single world model and allows the model to be progressively expanded to new control settings. By enabling joint learning from more diverse data sources, this training strategy alleviates the scaling bottleneck of current world models. Experiments show that MoE-based heterogeneous supervision brings consistent mutual gains across control types, achieving state-of-the-art performance on WorldArena, strong locomotion and hand-motion capabilities, robust out-of-distribution generalization, and loco-manipulation generation.

1 Introduction

The rapid progress of open-source, high-quality video generation models has greatly advanced the development of video generation world models [41, 45, 49]. Under various forms of controllable supervision, world models are able to model physical scenes and generate high-quality, scalable embodied data, thereby further promoting the development of intelligent agents [8, 39]. At the same time, they also serve as high-fidelity simulators, acting as effective engines for accelerating progress in gaming and interactive environments [4, 40]. These action-conditioned simulations exhibit strong interactivity, establishing a close connection between virtual generation and the real world, and making world models a promising path toward Artificial General Intelligence.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

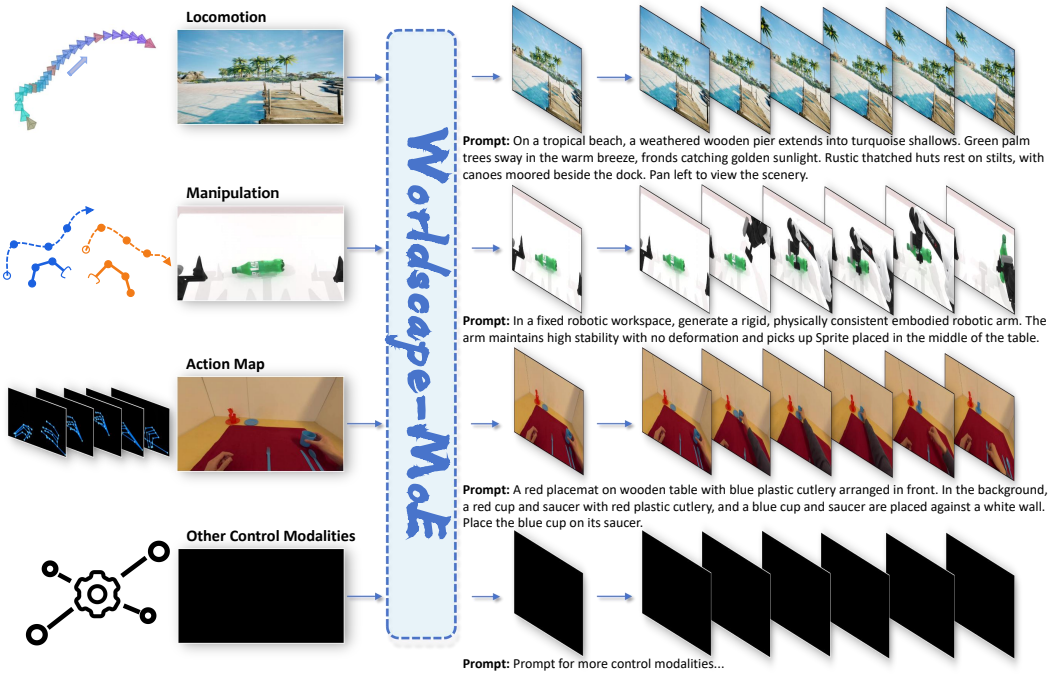


Figure 1: **Worldscape-MoE Overview.** Worldscape-MoE supports three mainstream control modalities: **Locomotion** for trajectory-conditioned world navigation, **Manipulation** for robot-action-conditioned embodied tasks, and **Action Map** for hand-joint-conditioned egocentric interaction generation. The framework can also be extended to additional control injection settings.

Video generation world models guided by conditional control can dynamically simulate environments based on current observations and given actions [11, 38]. For example, prior works such as Genie3 [4], the Matrix-Game series [16, 44, 50], and HYWorld 1.5 [36] have effectively explored environment simulation under camera control, demonstrating strong intelligence and maintaining high visual quality under real-time interaction. Other works, such as GigaWorld-0 [39] and Cosmos-Predict 2.5 [3], have advanced embodied world models (EWMs) [33, 23], where the model receives robotic-arm inputs for embodied manipulation. In addition, methods such as Hand2World [43] and Generated Reality [47] inject hand-joint signals to enable first-person hand control and interactive simulation. Although these different control signals all target first-person video generation, they partition egocentric world-modeling data into separate subsets according to control type. Since each subset is limited in scale, it is difficult for world models trained under a single control setting to achieve effective scaling [8].

In contrast, both humans and intelligent agents naturally integrate multiple forms of control input when perceiving, reasoning, and producing representations [9, 13]. The world representations learned under different supervisory controls should therefore be consistent, and core capabilities such as physical regularity modeling and logical consistency should ideally be shared across control settings. Existing approaches, however, restrict world model development to isolated single-control scenarios and lead to fragmented usage of valuable world-modeling data. Sparse Mixture-of-Experts (MoE) architectures [20], which process data by activating only a subset of parameters, have been widely adopted in prior vision-language modeling research and provide a natural foundation for handling heterogeneous supervision [7, 24].

Motivated by this observation, we propose **Worldscape-MoE**, as illustrated in Fig. 1, the first world model framework that incorporates an MoE design into DiT to unify heterogeneous supervisory controls. By combining heterogeneous data representations with a learnable router, Worldscape-MoE enables the first unified world modeling framework that is compatible with different forms of control supervision. This allows the model to jointly leverage ego-centric world-modeling data under different control types, substantially increasing the amount of data available for training. As a result, the model can learn physical dynamics and world representations from broader and more diverse sources, thereby improving the world-simulation capability of video generation.

Moreover, through extensive experimentation, we develop an effective and scalable MoE training strategy for world models, termed **Worldscape-MoE Tuning**. Under this strategy, all training samples are routed through both a shared expert and a control-specific expert, so that the shared expert continually absorbs world knowledge across all supported control settings. This training paradigm yields a shared-expert backbone with strong physical and logical consistency, robust generalization, and strong support for heterogeneous controls. It also enables continual expansion to new control settings without disrupting previously learned control capabilities.

In summary, our main contributions are as follows:

- We propose **Worldscape-MoE**, the first world model framework that supports heterogeneous action-control injection within a unified architecture, enabling training on all available egocentric world-modeling data and alleviating the scaling bottleneck caused by single-modality supervision.
- We present **Worldscape-MoE Tuning**, a scalable training strategy that supports continual integration of additional control modalities while allowing the shared expert to continually absorb world knowledge across all controls.
- Extensive experiments show clear positive transfer across modalities in Worldscape-MoE, with each single-modality inference setting outperforming existing baselines. In particular, Worldscape-MoE surpasses Ctrl-World by 2.86% on WorldArena and further achieves stronger physical consistency, higher visual quality, robust OOD generalization, and loco-manipulation generation.

2 Methodology

In this section, we present the design of the Worldscape-MoE framework in detail. In 2.1, we formally define the video generation world modeling problem. In 2.2, we describe the processing and curation of heterogeneous control data. In 2.3, we introduce the model design, including control representation injection and expert routing. Finally, in 2.4, we present the detailed training strategy and implementation details.

2.1 Problem Formulation

In this work, we study video generation world models under heterogeneous action supervision and seek to learn a scalable unified model that can accommodate diverse mainstream control signals. Formally, given the current observation frame O_t, a textual world description C_t, and a continuous action condition A_t, the model predicts the subsequent world observation:

$$W_{t+1} = f_\theta(O_t, C_t, A_t), \tag{1}$$

where A_t may take different forms of control input, including camera trajectory A_{traj}, hand-joint representation A_{hand}, dual-arm robot control vector A_{act}, and other extensible action modalities A_{other}, i.e.,

$$A_t \in \{A_{\text{traj}},\ A_{\text{hand}},\ A_{\text{act}},\ A_{\text{other}}\}\,. \tag{2}$$

The predicted W_{t+1} is a future video segment composed of consecutive frames, and its last frame O_{t+1} is fed back as the initial observation for the next prediction step, enabling long-horizon autoregressive world generation. Specifically, A_{traj} is a temporal camera-motion sequence, $A_{\text{traj}} \in \mathbb{R}^{T_v \times d_{\text{traj}}}$, which can be further converted into dense camera-control features. A_{hand} is a dense action-map sequence, $A_{\text{hand}} \in \mathbb{R}^{T_v \times H \times W \times C_{\text{hand}}}$, while A_{act} is a low-dimensional manipulation sequence, $A_{\text{act}} \in \mathbb{R}^{T_a \times d_{\text{act}}}$. In our implementation, $T_a = 17$ and $d_{\text{act}} = 14$, giving a 17×14 dual-arm action tensor per prediction window.

2.2 Heterogeneous Control Data Processing Pipeline

Generating worlds under heterogeneous control requires carefully curated and diverse training data, including high-quality control annotations, informative captions, and temporally coherent frame sequences. Our training corpus consists of three categories of data, corresponding to three different interaction modalities, with a total of 0.8M clips. More detailed descriptions of data processing are provided in C.

Camera-control data. For camera-trajectory-conditioned world generation, we curate 200K clips in total from three sources. **(I) Real-world data:** we use 60K clips from RealEstate10K [53] and 20K clips from DL3dv-10k [25] to cover diverse real-world scenes. **(II) Simulation data:** following the camera file design and data collection protocol of iWorld-Bench [10], we perform stylized data collection in Unreal Engine using 81 camera motion patterns across more than 100 diverse environments, resulting in 80K training clips. **(III) Diverse world data:** we further select 40K clips from Sekai [22] and SpatialVid [42], and re-annotate their trajectories using VIPE [18]. All collected data are then passed through a unified annotation pipeline. For textual descriptions, we adopt the caption annotation strategy introduced in Matrix-Game 3.0 [44].

Robotic embodied data. For robot-action-conditioned embodied world modeling, we process 350K clips in total. The data construction pipeline consists of two stages. **(I) Simulation data generation:** we build a large-scale simulation pipeline for robotic manipulation scenarios on the RoboTwin [28, 5] platform and generate data using RTX 4090 GPUs. This part covers 50 robotic tasks and diverse object-interaction scenarios. We segment the generated sequences along the temporal dimension into fixed-length samples, using a sliding window of 17 frames per training clip. **(II) Data generalization:** to improve the diversity and generalization of embodied tasks, we further develop an embodied data augmentation pipeline based on Seedance 2.0 [32] and Qwen3.6-Plus [30], which varies materials, colors, and manipulated objects. For captions, we directly use the embodied textual descriptions provided by the RoboTwin platform. This process yields 350K clips in total.

Egocentric hand data. For hand-joint-conditioned world generation, we process 250K clips in total. To construct the hand-control modality, we sample 150K and 100K clips from two large-scale egocentric video datasets, EgoDex [17] and Ego4D [15], respectively. Since the two datasets differ in annotation protocols and hand supervision formats, we apply HaMeR [29] to all selected frames to perform monocular 3D hand reconstruction in a unified manner. This yields 21 MANO joints for each hand, which are then projected into 2D pixel coordinates using the camera intrinsics and extrinsics. Finally, we rasterize them into hand-joint action maps that are spatially aligned with the video frames at the pixel level, serving as the supervision signal for the hand-control expert. For text conditioning, we use the captions provided by the original datasets.

2.3 Architecture of Worldscape-MoE

As shown in Fig. 2, Worldscape-MoE mainly consists of heterogeneous data representation, expert routing, a stack of DiT layers, and MoE blocks.

Heterogeneous Data Representation. During training, we jointly optimize the model on mixed-control data and design dedicated injection pathways according to the structural properties of different control signals. Video frames and reference images are first encoded by a frozen VAE into latent representations, which serve as the basic visual inputs to the diffusion model. Text instructions are encoded by a frozen text encoder to extract semantic features and are injected into the backbone through the original cross-attention pathway. **Locomotion** mainly characterizes camera trajectories or local viewpoint changes. Its control signals are transformed into dense camera-control features and mapped by a lightweight trainable Control Adapter into a feature space aligned with the video latents, after which they are fused with patch-level video tokens. **Hand Motion** provides dense spatiotemporal constraints in the form of action maps. These signals are therefore first encoded by the frozen VAE into latent control features and then fused into the conditional latent branch as additional visual conditions for the diffusion backbone. **Manipulation**, in contrast, is represented as a low-dimensional action sequence that specifies the intended operation. We use a trainable Action MLP to encode it into compact action embeddings, which are further integrated into the timestep embedding and its projected modulation vectors, allowing the manipulation signal to affect the generated dynamics through the temporal conditioning pathway of DiT blocks. These heterogeneous control signals can be unified as modality-dependent conditioning functions within a shared diffusion backbone:

$$\hat{\epsilon} = f_\theta\big(z_v \oplus \phi_{\text{vis}}(A),\ z_c,\ e_t + \phi_{\text{tmp}}(A)\big), \tag{3}$$

where z_v denotes the VAE latent of video frames and reference images, z_c denotes the text embedding, e_t denotes the timestep embedding, $\phi_{\text{vis}}(A)$ denotes the control-dependent visual injection function, and $\phi_{\text{tmp}}(A)$ denotes the control-dependent temporal modulation function. For **Locomotion**, $\phi_{\text{vis}}(A)$ is implemented by the Control Adapter and fused at the patch-token level; for **Hand Motion**, it corresponds to VAE-encoded action-map latents injected into the conditional visual branch; and

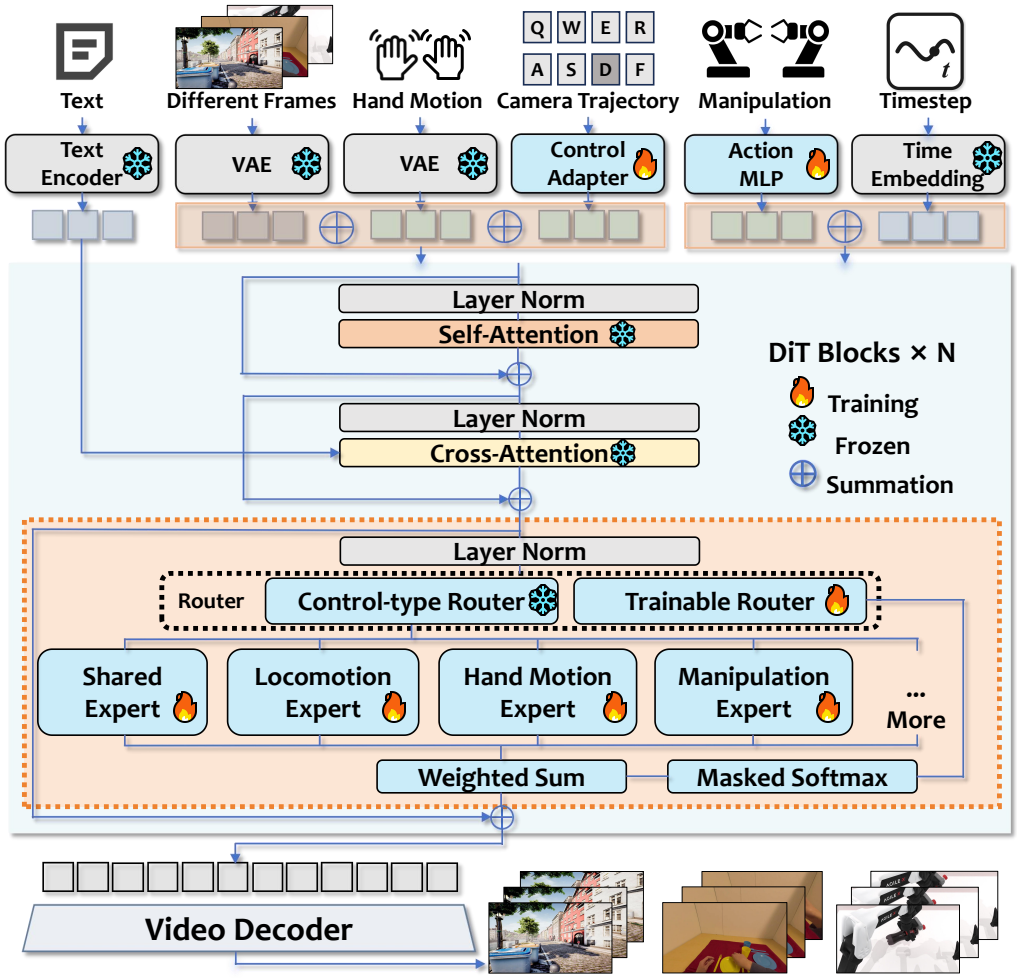


Figure 2: **Worldscape-MoE Architecture.** Given the current world observation and different forms of supervisory control, our framework generates world dynamics under heterogeneous control signals. It supports both egocentric world exploration and embodied task execution.

for **Manipulation**, $\phi_{\text{tmp}}(A)$ is implemented by the Action MLP and injected through the timestep modulation pathway. Through this unified formulation, heterogeneous controls are incorporated into different pathways of the same DiT backbone, enabling effective multi-control modeling while preserving the priors of the pretrained video generator.

Expert Routing and MoE Forward. We replace the feed-forward network in each DiT block with a control-aware MoE-FFN, enabling shared generative computation and per-modality specialization under heterogeneous supervision. For a model supporting M control modalities, the expert set at block l is $\mathcal{E}^{(l)} = \{\mathcal{E}_0^{(l)}, \mathcal{E}_1^{(l)}, \ldots, \mathcal{E}_M^{(l)}\}$, where $\mathcal{E}_0^{(l)}$ is shared across all samples and $\mathcal{E}_r^{(l)}$ is specialized for the r-th control modality. All experts are initialized from the pretrained FFN or the current shared expert (Sec. 2.4) rather than from random weights, preserving the video-generation prior.

For sample i, let $H_i^{(l)} = [h_{i,1}^{(l)}, \ldots, h_{i,N}^{(l)}] \in \mathbb{R}^{N \times d}$ denote the hidden tokens at block l. We encode the available controls by a binary vector $\mathbf{u}_i = [u_{i,1}, \ldots, u_{i,M}] \in \{0,1\}^M$ and construct an eligibility mask

$$\mathbf{m}_i = [1, u_{i,1}, \ldots, u_{i,M}] \in \{0,1\}^{M+1}, \qquad \mathcal{A}_i = \{k \in \{0, \ldots, M\} \mid m_{i,k} = 1\}, \tag{4}$$

where the shared expert is always eligible and each modality-specific expert is activated only when its corresponding control signal is present.

The router then predicts sample-wise logits from mean-pooled hidden states and computes routing weights by a masked softmax over the eligible experts:

$$\mathbf{z}_i^{(l)} = \frac{1}{N}\sum_{n=1}^{N} h_{i,n}^{(l)}, \qquad \mathbf{a}_i^{(l)} = \mathbf{W}_{\text{route}}^{(l)}\,\mathbf{z}_i^{(l)}, \tag{5}$$

$$\alpha_{i,k}^{(l)} = \frac{m_{i,k}\exp(a_{i,k}^{(l)}/\tau)}{\sum_{j=0}^{M} m_{i,j}\exp(a_{i,j}^{(l)}/\tau)}, \qquad k \in \{0,\dots,M\}, \tag{6}$$

where $\mathbf{W}_{\text{route}}^{(l)} \in \mathbb{R}^{(M+1)\times d}$ is a learnable routing matrix and τ is a temperature hyperparameter. By construction, ineligible experts receive zero probability and do not participate in the forward computation or the sample-wise gradient update.

The MoE-FFN output is then aggregated over the active expert set $\mathcal{A}_i$:

$$\text{MoE}^{(l)}\left(H_i^{(l)}, \mathbf{u}_i\right) = \sum_{k\in\mathcal{A}_i} \alpha_{i,k}^{(l)}\,\mathcal{E}_k^{(l)}\left(H_i^{(l)}\right). \tag{7}$$

The router is implemented as a bias-free linear layer with zero initialization, yielding uniform weights over $\mathcal{A}_i$ at the start of training. We do not introduce an auxiliary load-balancing loss; instead, expert specialization is induced by deterministic modality eligibility and optimized end-to-end with the diffusion objective.

2.4 Worldscape-MoE Training Structure

Grouped Learning. To rapidly adapt the model to new control capabilities while preserving the generation priors inherited from pretraining, we adopt a grouped learning-rate strategy, as illustrated in Fig. 3. Newly introduced or strongly modality-specific modules are optimized with the base learning rate, whereas shared components inherited from the pretrained model, such as the Shared Expert and part of the adapter parameters, are fine-tuned with a smaller learning rate.

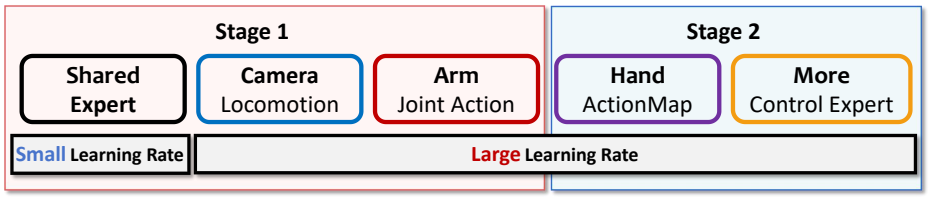


Figure 3: **Worldscape-MoE Tuning.** The shared expert is optimized with a smaller learning rate, while each modality-specific expert uses a larger learning rate for control specialization.

Action Injection. To stably extend the model to multiple control capabilities, we adopt a progressive modality-incremental strategy rather than introducing all experts and control branches at once. The process consists of two stages. In **Stage I**, starting from Wan 2.2 5B [41], we train the model with camera-control and robotic-arm signals to establish reliable generation priors, stable camera-trajectory control, and embodied task capability. In **Stage II**, we progressively introduce new control modalities together with their corresponding experts. Each newly added expert is initialized from the current Shared Expert rather than from random weights. This stage-wise design makes Worldscape-MoE a continually extensible framework that can absorb heterogeneous supervisory signals for world simulation.

3 Experiments

3.1 Experimental Setup

We train our model based on Wan 2.2 5B [41]. All experiments are conducted on a compute cluster equipped with 64 NVIDIA H20 GPUs. To ensure a fair and reproducible comparison, we standardize the evaluation protocol across models whenever their official implementations permit the

corresponding control and resolution settings, with the detailed inference configurations summarized in Appendix Table 5. Below, we provide the detailed evaluation settings for the three supervision modalities.

Locomotion Setting. Following the dataset and metric design of iWorld-Bench [10], we construct a 500-case locomotion evaluation set and adopt its trajectory-following, generation-quality, and consistency metrics.

Manipulation Setting. For robotic-arm manipulation, we follow the WorldArena evaluation protocol [33]. We directly adopt the released WorldArena results and use the corresponding inference configurations for manipulation baselines.

Hand Motion Setting. For hand-motion control, we follow the egocentric hand-motion evaluation setting and randomly select 100 hand-object interaction samples from the EgoDex test set [17]. Following HandWorld [37], we report CLIP-Hand and Hand IoU to measure hand-region semantic consistency and spatial overlap. Following Generated Reality [47], we additionally report L2Err in the 2D image space, PA-MPJPE for Procrustes-aligned 3D joint accuracy, and PA-MPVPE for Procrustes-aligned 3D mesh accuracy. We also report Image Quality to assess overall visual fidelity.

Table 1: **Locomotion Experiments.** Comparison of metric scores for different models on generation quality and trajectory following tasks.

		Generation Quality				Trajectory Following	
Method	Avg.↑	Image Quality ↑	Brightness Consistency ↑	Color Temp. Constraint ↑	Sharpness Retention ↑	Motion Smoothness ↑	Trajectory Accuracy ↑
Worldscape-MoE	**0.7556**	0.6955	0.7758	**0.7639**	0.6745	**0.9941**	0.6300
w/o MoE	0.6869	0.6710	0.6993	0.6613	0.4865	0.9930	0.6100
Matrix-game 3.0	0.6232	0.5633	0.6180	0.6353	0.3852	0.9660	0.5714
HY-World 1.5	0.7322	**0.7128**	0.7027	0.7477	0.5545	0.9908	0.6844
CameraCtrl	0.5521	0.4602	0.4812	0.3076	0.4833	0.9832	0.5970
MotionCtrl	0.5562	0.4583	0.5296	0.2421	0.5182	0.9776	0.6115
CamI2V	0.6137	0.5150	0.5904	0.4513	0.5255	0.9886	0.6115
RealCam-I2V	0.7063	0.6530	0.5712	0.6197	0.6987	0.9901	0.7050
videox-fun-Wan	0.7443	0.6684	0.6856	0.6640	0.6934	0.9899	**0.7645**
AC3D	0.7262	0.4884	**0.7764**	0.7050	**0.7213**	0.9934	0.6729
ASTRA	0.6072	0.5600	0.5916	0.5088	0.5625	0.9826	0.4379

Table 2: **Manipulation Experiments.** EWM Score for dual-arm manipulation across 16 metrics.

Models	EWM Score
Worldscape-MoE	**62.84**
w/o MoE	61.88
CtrlWorld	59.98
Wan 2.6	59.80
CogvideoX	58.79
Veo 3.1	57.77
IRASim	56.14
TesserAct	54.62
Cosmos-Predict 2.5 (action)	54.29
Cosmos-Predict 2.5 (text)	53.06
Vidar	51.92
Wan 2.2	51.71
GigaWorld-0	50.96
RoboMaster	50.35

Table 3: **Hand-motion Experiments.** Comparison of hand-control consistency, spatial alignment, and visual quality.

Method	CLIP-Hand ↑	Hand IoU ↑	L2Err (px) ↓	PA-MPJPE (mm) ↓	PA-MPVPE (mm) ↓	Image Quality ↑
Worldscape-MoE	**0.9475**	**0.7629**	**10.3080**	**5.4189**	**5.1053**	**0.7325**
w/o MoE	0.9276	0.4850	49.0132	7.8311	7.4292	0.7250
HunyuanVideo-1.5	0.9051	0.3624	70.8834	8.8883	8.4693	0.6419
MimicMotion	0.8928	0.5348	36.8214	9.2072	8.7404	0.5324
MagicDance	0.8643	0.3109	72.9801	10.5826	10.1204	0.5739
LOME	0.7382	0.0839	109.5550	10.9962	10.5841	0.5281

3.2 Heterogeneous Control Evaluation

Heterogeneous control performance. Table 1 shows that Worldscape-MoE attains the best overall locomotion score of 0.7556, improving over the strongest baseline VideoX-Fun-Wan by 0.0113 while leading on motion smoothness with 0.9941, indicating that the framework follows camera-level trajectory controls without sacrificing generation quality. On robotic-arm manipulation, Table 2 reports the highest EWM score of 62.84, surpassing the next-best baselines CtrlWorld and Wan 2.6 by $+2.86$ and $+3.04$, respectively, with detailed per-metric breakdowns provided in Appendix Tables 6 and 7. For egocentric hand-motion control, Table 3 further shows that Worldscape-MoE achieves the best performance across all six metrics, obtaining 0.9475 CLIP-Hand, 0.7629 Hand IoU, 10.3080

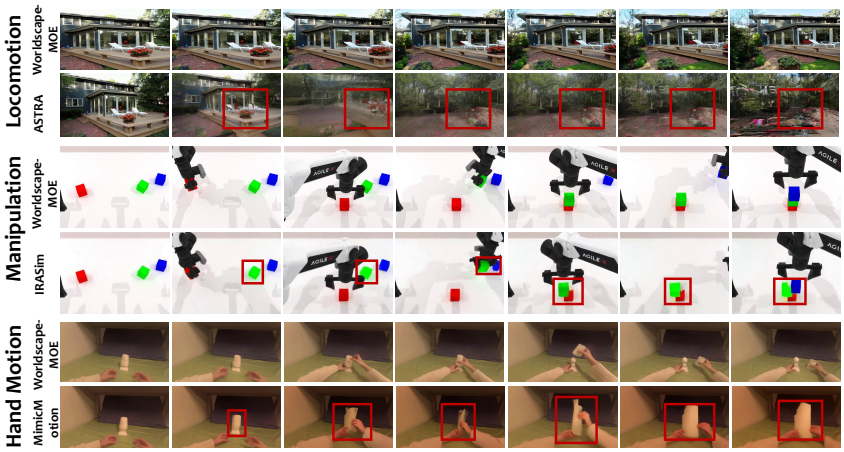


Figure 4: **Qualitative comparison under heterogeneous controls.** Across locomotion, manipulation, and hand-motion settings, Worldscape-MoE produces more controllable and physically coherent world dynamics than representative baselines.

px L2Err, 5.4189 mm PA-MPJPE, 5.1053 mm PA-MPVPE, and 0.7325 Image Quality. Compared with the w/o MoE variant, it improves CLIP-Hand and Hand IoU by 0.0199 and 0.2779, reduces L2Err, PA-MPJPE, and PA-MPVPE by 38.7052 px, 2.4122 mm, and 2.3239 mm, respectively, and improves Image Quality by 0.0075. These results directly demonstrate improvements in visual fidelity, hand-action consistency, and spatial hand alignment, while confirming the effectiveness of the proposed MoE framework.

Across the three heterogeneous control regimes, Worldscape-MoE ranks first on every primary metric, demonstrating a level of cross-modality consistency rarely achieved by prior controllable world models. The simultaneous gains over the strongest single-modality baselines further indicate that joint training on locomotion, manipulation, and hand-motion data leads to mutual reinforcement rather than the cross-modal interference commonly reported in multi-task video generation, which is also reflected in the qualitative comparison in Fig. 4.

3.3 Quantitative Analysis of Expert Routing

To verify whether Worldscape-MoE learns control specialization rather than merely scaling capacity, we record the gate-weighted expert workload at each MoE-FFN layer on the action-only evaluation cases, which reflects how computation is allocated across the shared and modality-specific experts. Concretely, we attach forward hooks to all MoE-FFN routers during inference, accumulate each expert's routing weight over all tokens and denoising steps, and then normalize the resulting layer-wise weighted token counts to draw the all-expert, dedicated-expert, and shared-vs-dedicated loading panels.

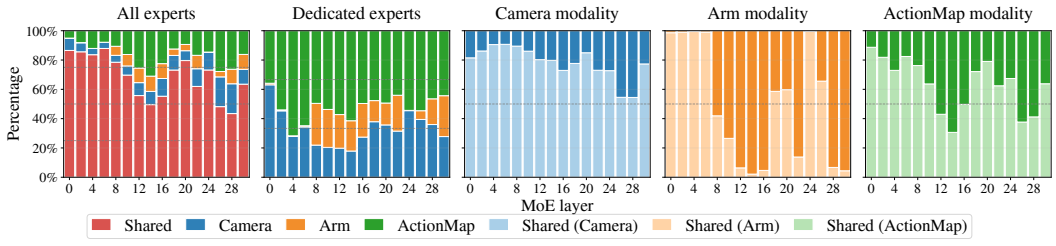


Figure 5: **Layer-wise MoE expert loading analysis.** Gate-weighted expert workload across MoE layers. The panels show all-expert loading, dedicated-expert loading, and shared-vs-dedicated routing for locomotion, manipulation, and hand-motion control.

As shown in Fig. 5, the shared expert remains the dominant path in the overall workload, accounting for 69.48% of gate-weighted computation across the sampled action-only cases. The dedicated experts, in contrast, exhibit clear modality-dependent usage: they account for 20.91% of the workload for locomotion, 47.95% for manipulation, and 35.97% for hand-motion. This pattern suggests that camera-like locomotion can still be largely handled by the shared spatiotemporal prior with moderate expert-specific correction, while manipulation and action-map hand control rely more strongly on their corresponding modality experts. These updated routing statistics confirm that the more fully trained Worldscape-MoE jointly maintains a unified backbone and learns specialized experts for heterogeneous action controls.

3.4 Ablation on MoE Effectiveness and Scalability

MoE Effectiveness. To isolate the contribution of the MoE design, we compare Worldscape-MoE with a direct mixed-training baseline without MoE that jointly optimizes all control modalities using the same dense backbone. As shown in Fig. 6 and Fig. 30, the dense baseline suffers from cross-modal control interference, resulting in weaker control fidelity and less coherent world dynamics. Although this Without MoE baseline remains competitive overall, Worldscape-MoE achieves a higher EWM Score and stronger interaction/control-related metrics, indicating that expert routing helps reduce cross-modal interference. In contrast, Worldscape-MoE better preserves modality-specific control behavior while maintaining stronger visual and physical consistency. Worldscape-MoE and the w/o MoE baseline are trained for the same number epochs, yet Worldscape-MoE consistently outperforms the dense baseline across all metrics. This suggests that the acquisition of world knowledge is shared across control modalities: the model can learn scene generation, physical causality, and dynamic consistency not only from the target modality but also from other forms of control supervision.

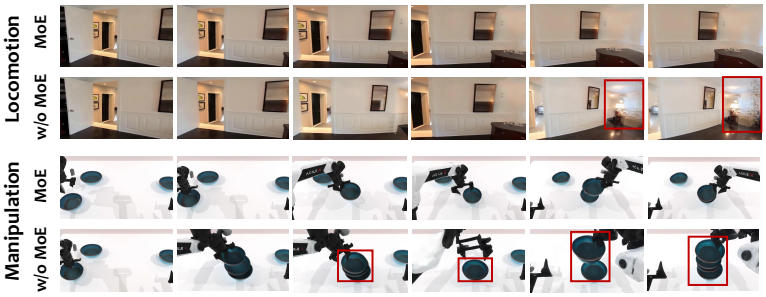


Figure 6: **Qualitative evidence for the effectiveness of MoE.** Worldscape-MoE produces more stable results than direct mixed training without MoE under heterogeneous controls.

MoE Scalability. We further evaluate whether the proposed framework can be extended to additional control modalities without compromising previously acquired capabilities. After introducing a new modality, Worldscape-MoE successfully learns the new control signal while largely preserving generation quality and control fidelity on existing modalities. We observe a brief degradation in locomotion performance at the early stage of extension, consistent with a temporary adaptation phase as shared and modality-specific parameters are re-balanced to absorb the new control knowledge. This transient effect is gradually mitigated as training proceeds, and the original locomotion capability recovers. We also conduct an additional extension experiment on single-arm LIBERO manipulation, with the setting and qualitative results provided in Appendix D.4. Additional evidence from parameter-update dynamics is provided in Appendix D.3. Overall, these results suggest that the proposed MoE framework can continuously absorb new control types without substantially degrading existing capabilities.

3.5 Out of Distribution Detection

As shown in Fig. 7, Worldscape-MoE demonstrates strong out-of-distribution generalization across locomotion, robotic manipulation, and hand-motion domains. It generates coherent locomotion rollouts under large visual shifts, generalizes simulated Robotwin arms to realistic furniture-service scenarios, and transfers hand-pose control trained from a single dataset to diverse unseen scenes. These results suggest that data and physical knowledge are shared across modalities, enabling cross-domain scene transfer and richer embodied dynamics modeling.

Beyond single-modality control, Worldscape-MoE benefits from generating from broad and diverse world knowledge rather than being constrained to one specific control form. This leads to stronger modeling of physical regularities, as illustrated in Fig. 31, and further enables coupled reasoning across different actions. In particular, we observe that the model can synthesize loco-manipulation behaviors in diverse environments, as shown in Fig. 32. More qualitative out-of-distribution cases are provided in Appendix D.5.

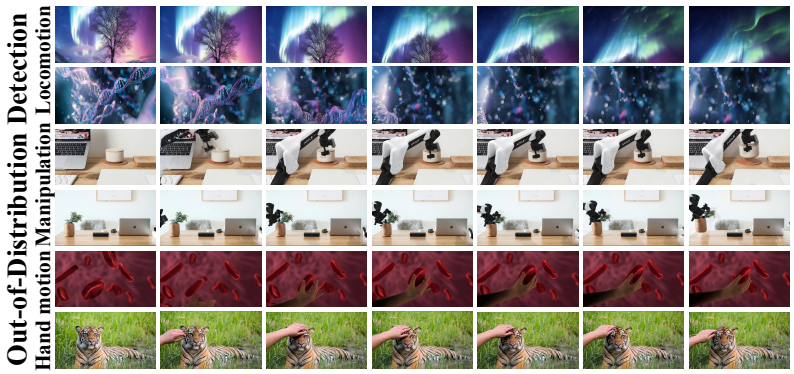


Figure 7: **Out-of-distribution detection across three motion domains.** Representative OOD across locomotion, robotic manipulation, and hand-motion tasks.

4 Related Work

Video Generation World Models. Recent progress in video generation world models has been enabled by high-quality open-source video foundation models, such as the Wan series [41], Cosmos [2], CogVideoX [49], and HunyuanVideo [45]. Existing world models typically generate future videos with DiT-based diffusion conditioned on current observations, text, and control signals, and can be broadly divided into **Locomotion**-based methods, such as the Genie series [4], the Matrix-Game series [16, 50, 44], and the HY-World series [19], and **Manipulation**-based methods, such as Hand2World [43], GigaWorld-0 [39], and Cosmos-Predict 2.5 [3]. These approaches differ in how control signals are injected, but all focus on a single control paradigm. In contrast, our framework unifies multiple mainstream control types within a single world model.

Mixture-of-Experts (MoE). Mixture-of-Experts (MoE) architectures have shown strong performance in Transformer-based models [34, 21, 12, 7] and have also been successfully applied to vision-language models for handling heterogeneous modalities, as exemplified by LLaVA [24] and DeepSeek [46]. MoE designs have further demonstrated promise in image diffusion generation, e.g., Nucleus-Image [1]. More recently, MoE concepts have started to appear in generative world modeling: CogVideoX [49] enhances text-video fusion with expert transformers and adaptive LayerNorm, Astra [54] introduces mixture-of-action experts for action encoding, and Wan2.2 14B [6] employs noise-aware experts across denoising stages. However, existing works still focus on isolated forms of expert specialization, and the use of MoE to unify heterogeneous control modalities within a single world model remains largely unexplored. Our work addresses this gap by introducing an MoE-based framework for unified heterogeneous-control world modeling.

5 Conclusion and Future Work

In this work, we presented Worldscape-MoE, the first world model framework that unifies heterogeneous action controls within a single model. By jointly learning from heterogeneous control data, Worldscape-MoE improves physical consistency, temporal coherence, and overall world-simulation quality over single-control training, offering a promising direction for alleviating the data bottleneck of current world models. Through the combination of a shared expert for cross-control world knowledge and modality-specific experts for control specialization, our framework achieves strong generalization and competitive performance across diverse benchmarks and downstream tasks. In future work, we plan to further improve its efficiency and scalability by compressing the DiT architecture and extending the framework to broader control modalities.

Acknowledgments

This paper was supported by the Natural Science Foundation of China under Grant 62371269, Shenzhen Science and Technology Program ZDCY202517012, Meituan Academy of Robotics Shenzhen.

References

[1] Chandan Akiti, Ajay Modukuri, Murali Nandan Nagarapu, Gunavardhan Akiti, and Haozhe Liu. Nucleus-image: Sparse moe for image generation. *arXiv preprint arXiv:2604.12163*, 2026.

[2] Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025.

[3] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

[4] 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.

[5] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

[6] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Feng Wang, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

[7] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

[8] 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.

[9] Marc O Ernst and Heinrich H Bülthoff. Merging the senses into a robust percept. *Trends in cognitive sciences*, 8(4):162–169, 2004.

[10] Jianjie Fang, Yingshan Lei, Qin Wan, Ziyou Wang, Yuchao Huang, Yongyan Xu, Baining Zhao, Weichen Zhang, Chen Gao, Xinlei Chen, and Yong Li. iworld-bench: A benchmark for interactive world models with a unified action generation framework. *arXiv preprint arXiv:2605.03941*, 2026.

[11] Jianjie Fang, Xvyuan Liu, Ziyou Wang, Rongze Tang, Zhaolu Wang, Zhuohang Li, Xin Zhang, Haisheng Su, Chen Gao, Wei Wu, Xinlei Chen, and Yong Li. CAER: Causal action effect reweighting for world model training. *arXiv preprint arXiv:2608.30897*, 2026. URL `https://arxiv.org/abs/2608.30897`.

[12] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2022.

[13] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, 44(5): 701–739, 2025.

[14] Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, Xinlei Chen, and Yong Li. EmbodiedCity: A benchmark platform for embodied agent in real-world city environment. *arXiv preprint arXiv:2410.09604*, 2024. URL https://arxiv.org/abs/2410.09604.

[15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[16] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

[17] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[18] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

[19] Team HY-World, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, et al. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv preprint arXiv:2604.14268*, 2026.

[20] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[21] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

[22] Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675*, 2025.

[23] Zhuohang Li, Yiding Ma, Yu Shang, Lei Jin, Shengyuan Wang, Jianjie Fang, Zhaolu Wang, Xvyuan Liu, Yinzhou Tang, Xin Jin, Ziyou Wang, Weikang Su, Xin Zhang, Haisheng Su, Weizhen He, Wei Wu, Haoyi Duan, Gordon Wetzstein, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Xin Wang, Wenwu Zhu, Chen Gao, and Yong Li. WorldArena 2.0: Extending embodied world model benchmarking on modality, functionality and platform. *arXiv preprint arXiv:2605.17912*, 2026. URL https://arxiv.org/abs/2605.17912.

[24] Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

[25] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024.

[26] 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.

[27] Xvyuan Liu, Jianjie Fang, Wei Wu, Chen Gao, and Yong Li. Aim short to reach far: Your frozen world model can plan better than you think. *arXiv preprint arXiv:2609.30036*, 2026. URL https://arxiv.org/abs/2609.30036.

[28] Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, et al. Robotwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the computer vision and pattern recognition conference*, pages 27649–27660, 2025.

[29] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024.

[30] Qwen Team. Qwen3.6-Plus: Towards real world agents, April 2026. URL `https://qwen.ai/blog?id=qwen3.6`.

[31] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017.

[32] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

[33] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

[34] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[35] Haisheng Su, Zongdai Liu, Xin Jin, Haoxuan Dou, Chengming Hu, Baorun Li, Zhanwang Liu, Ruiyan Xu, Jianjie Fang, Xin Zhang, Zhenjie Yang, Xue Yang, Chen Gao, Junchi Yan, Yong Li, and Wei Wu. WorldScape Policy 2.0: Empowering steerable world action modeling with reasoning-augmented memory and in-context learning. *arXiv preprint arXiv:2607.18840*, 2026. URL `https://arxiv.org/abs/2607.18840`.

[36] Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

[37] Zhihao Sun, Zhiying Du, Xitong Yang, and Zuxuan Wu. Handworld: Hand-centric unified video action generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15976–15985, June 2026.

[38] Rongze Tang, Jianjie Fang, Zhaolu Wang, Ziyou Wang, Xvyuan Liu, Haisheng Su, Xin Zhang, Wei Wu, Chen Gao, Yong Li, and Zhibo Chen. IMPACT: Attention is the interaction map for scalable interaction-aware world model training. *arXiv preprint arXiv:2609.00161*, 2026. URL `https://arxiv.org/abs/2609.00161`.

[39] GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, et al. Gigaworld-0: World models as data engine to empower embodied ai. *arXiv preprint arXiv:2511.19861*, 2025.

[40] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

[41] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[42] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025.

[43] Yuxi Wang, Wenqi Ouyang, Tianyi Wei, Yi Dong, Zhiqi Shen, and Xingang Pan. Hand2world: Autoregressive egocentric interaction generation via free-space hand gestures. *arXiv preprint arXiv:2602.09600*, 2026.

[44] Zile Wang, Zexiang Liu, Jaixing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

[45] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

[46] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[47] Linxi Xie, Lisong C Sun, Ashley Neall, Tong Wu, Shengqu Cai, and Gordon Wetzstein. Generated reality: Human-centric world simulation using interactive video generation with hand and camera control. *arXiv preprint arXiv:2602.18422*, 2026.

[48] Jingbo Xu, Jianjie Fang, Zhuohang Li, Zihao Song, Zhaolu Wang, Xvyuan Liu, Xin Zhang, Haisheng Su, Yinzhou Tang, Chen Gao, Wei Wu, and Yong Li. UniW(A)M: A unified mixture-of-experts framework for world and world-action models. *arXiv preprint arXiv:2610.18697*, 2026. URL `https://arxiv.org/abs/2610.18697`.

[49] 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.

[50] Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

[51] Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and Yong Li. UrbanVideo-Bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32400–32423, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1558. URL `https://aclanthology.org/2025.acl-long.1558/`.

[52] Baining Zhao, Jiacheng Xu, Weicheng Feng, Xin Zhang, Zhaolu Wang, Haoyang Wang, Shilong Ji, Ziyou Wang, Jianjie Fang, Zhiheng Zheng, Weichen Zhang, Yu Shang, Wei Wu, Chen Gao, Xinlei Chen, and Yong Li. WorldVLN: Autoregressive world action model for aerial vision-language navigation. *arXiv preprint arXiv:2605.15964*, 2026. URL `https://arxiv.org/abs/2605.15964`.

[53] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[54] Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General interactive world model with autoregressive denoising. *arXiv preprint arXiv:2512.08931*, 2026. URL `https://arxiv.org/abs/2512.08931`.

A Broader Impacts

This work may have several positive broader impacts. By unifying heterogeneous control modalities within a single world model, Worldscape-MoE provides a scalable framework for learning from broader and more diverse data sources, which may facilitate the development of more general and data-efficient intelligent agents. The resulting improvements in simulation fidelity and controllability could benefit robotics, embodied AI, and interactive decision-making [52, 14], while reducing reliance on costly real-world data collection [51]. In addition, the shared-expert design may promote transferable physical knowledge across tasks and control settings, supporting more robust downstream learning.

At the same time, the high fidelity of generated videos also requires careful ethical consideration. In particular, such models could be misused to generate deceptive synthetic videos, including deepfakes, or to misrepresent safe procedures and robot behaviors. To mitigate these risks, future work should explore safeguards such as synthetic-content detection, provenance or watermarking mechanisms, and anomaly detection methods for identifying unsafe or non-standard behaviors. We also plan to incorporate stronger safety-oriented generation constraints and detection mechanisms into future public releases of our models to support more responsible deployment in safety-critical applications.

B Limitation

Despite the strong performance of our framework, Worldscape-MoE remains computationally expensive to train and requires substantial GPU resources. In addition, the training speed decreases as more control modalities are incorporated, which limits the practical scalability of the current framework. Improving training efficiency is therefore an important direction for future work. One possible solution is to pre-encode VAE features in advance to reduce the computational overhead during training. Another limitation is that the current framework is not yet optimized for real-time deployment. In particular, how to distill Worldscape-MoE into a faster model and adapt it to existing acceleration or distillation pipelines remains an open problem. Addressing these challenges is essential for building a real-time, high-performance world model that supports heterogeneous control at scale.

C Data Process Details

C.1 Open Sources Datasets Details

RealEstate10K. RealEstate10K [53] is a large-scale dataset providing camera trajectories for roughly 10 million frames, collected from about 80,000 clips sourced from 10,000 YouTube videos. For each clip, camera poses are organized as continuous trajectories describing spatial position and orientation over time. These poses are estimated via large-scale reconstruction pipelines based on SLAM and bundle adjustment.

DL3DV-10K. DL3DV-10K [25] is a real-world dataset containing over 10,000 high-quality videos with manual annotations of key scene structures and complexity. In addition to camera poses, it includes rich geometric information such as NeRF-based depth, point clouds, and 3D meshes. The dataset supports tasks including 3D reconstruction, scene consistency modeling, and vision-language learning.

Sekai. Sekai [22] is a large-scale, diverse video dataset that focuses on rich scene dynamics and varied visual content, making it suitable for tasks involving motion understanding and generation. The dataset contains a wide range of environments and object interactions, providing complex camera movements and temporal variations. Each video clip is accompanied by structured annotations, including captions and motion-related information, which facilitate downstream tasks such as trajectory modeling and video synthesis.

SpatialVID. SpatialVID [42] is a large-scale video dataset built for multi-modal generation tasks such as text-to-video and image-to-video. It spans about 10 TB and is split into 545 subsets, each containing video data along with corresponding annotations. Each clip is paired with structured metadata that can be easily processed, and the annotations include captions, dynamic masks, camera parameters, poses, and motion instructions, supporting detailed spatial-temporal modeling.

EgoDex. EgoDex [17] is an egocentric video dataset designed for learning dexterous human manipulation from first-person observations. It contains 829 hours of 30 Hz, 1080p videos collected with Apple Vision Pro and ARKit/visionOS, covering 194 active tabletop manipulation tasks such as grasping, placing, folding, and arranging objects. Each video is paired with frame-level 3D pose annotations for the head, upper body, and both hands, along with camera intrinsics, camera transforms, joint SE(3) transforms, confidence values, and natural-language annotations; from EgoDex, we obtain 150K hand-joint control training clips.

Ego4D. Ego4D [15] is a large-scale open-domain egocentric video dataset capturing unscripted daily-life activities across diverse environments. It contains 3,670 hours of first-person videos collected by 931 camera wearers from 74 locations in 9 countries, spanning household activities, workplaces, outdoor scenes, leisure activities, social interactions, and commuting. Although Ego4D provides broad scene diversity and supports multiple first-person perception benchmarks, it does not include native frame-aligned 3D hand-joint poses for all videos, so we generate hand-joint control supervision using the same automatic hand reconstruction pipeline and obtain 100K training clips.

LIBERO. LIBERO [26] is a language-conditioned benchmark and dataset for lifelong single-arm robot manipulation. It provides an extendible procedural generation pipeline together with high-quality human-teleoperated demonstrations for 130 tasks organized into four task suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-100. These suites are designed to evaluate knowledge transfer under controlled changes in spatial relations, object categories, task goals, and their entangled combinations. For our single-arm setting, we process the released LIBERO demonstrations into robot-action-conditioned video clips with aligned task language and action trajectories, and use the corresponding train/validation metadata to study transfer to held-out manipulation cases.

C.2 Synthetic Data

In this study, we adopt the 81 camera configurations and data collection strategies from iWorld-Bench [10], and perform large-scale stylized data acquisition in Unreal Engine. Specifically, we leverage more than 50 diverse scenes spanning both indoor and outdoor environments, covering a wide range of visual styles such as urban, natural, and fantasy settings. Across these scenes, we sample approximately 1,000 camera viewpoints in total, and further generate around 80K training clips through systematic trajectory execution and multi-view capture. As summarized in Table 4, we present a representative subset of the selected scenes, along with their scene categories (indoor/outdoor) and the corresponding number of sampling points, illustrating the diversity and coverage of our data collection process.

C.3 Embodied Data Collection on the RoboTwin Platform

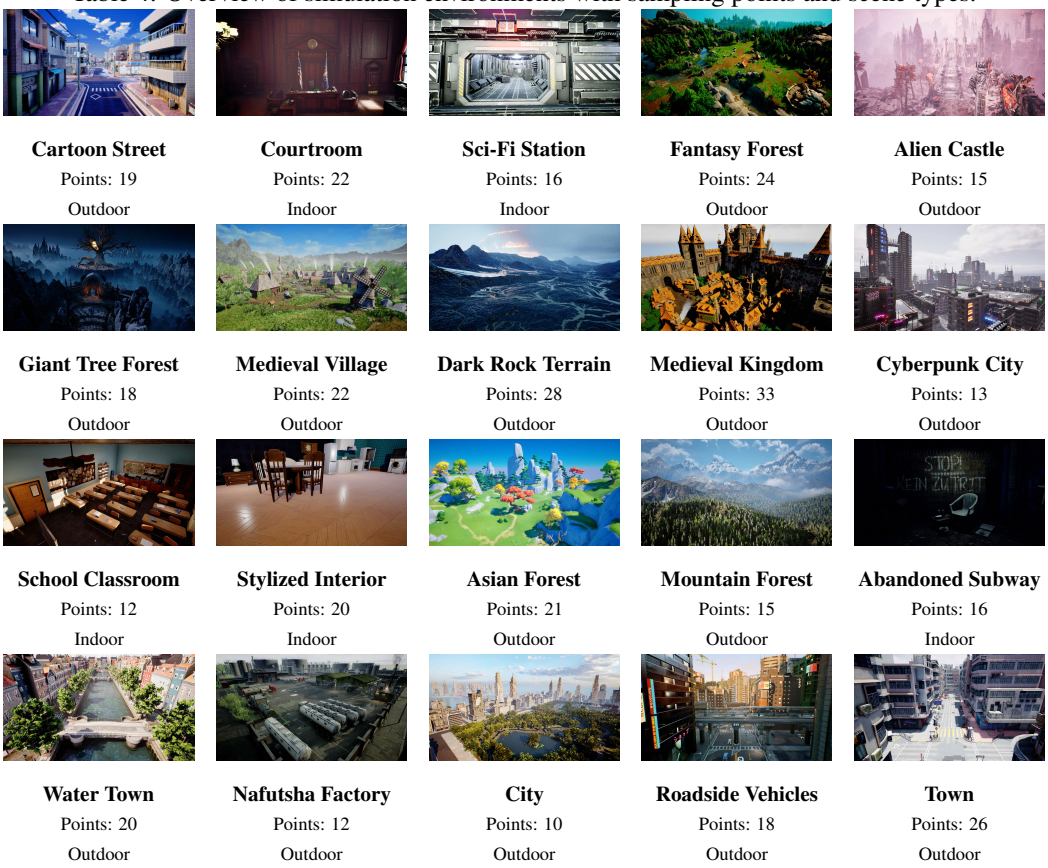
RoboTwin 2.0.[5] RoboTwin 2.0 is an embodied platform for bimanual manipulation in simulation, integrating unified interfaces, robot embodiments, and success-checking logic. By instantiating tasks from predefined configurations, the simulator records synchronized observations, actions, and metadata for every successful episode. This design ensures scalable data collection while preserving the spatial-temporal and semantic information essential for downstream policy learning and reasoning.

Training data is collected on machines equipped with NVIDIA GeForce RTX 4090 GPUs using a distributed pipeline that replays successful seeds. For each episode, the system archives multi-modal data, including HDF5 files, rendered videos, and action trajectories. All successfully parsed videos are rendered at 30 FPS with a resolution of 1280x704, with video durations ranging from approximately 2.50s to 51.63s.

C.4 Seedance 2.0 Embodied Data Augmentation Pipeline

Collecting real-world embodied interaction videos is costly and often suffers from limited visual diversity. Models trained solely on raw data are therefore prone to overfitting to specific object colors, appearances, materials, or scene textures, rather than learning the underlying relationship between control signals and physical interaction dynamics. To improve visual generalization while preserving the validity of action supervision, we build an embodied data augmentation pipeline based on Seedance 2.0 [32], as illustrated in Fig. 8.

Table 4: Overview of simulation environments with sampling points and scene types.



Cartoon Street	**Courtroom**	**Sci-Fi Station**	**Fantasy Forest**	**Alien Castle**
Points: 19	Points: 22	Points: 16	Points: 24	Points: 15
Outdoor	Indoor	Indoor	Outdoor	Outdoor
Giant Tree Forest	**Medieval Village**	**Dark Rock Terrain**	**Medieval Kingdom**	**Cyberpunk City**
Points: 18	Points: 22	Points: 28	Points: 33	Points: 13
Outdoor	Outdoor	Outdoor	Outdoor	Outdoor
School Classroom	**Stylized Interior**	**Asian Forest**	**Mountain Forest**	**Abandoned Subway**
Points: 12	Points: 20	Points: 21	Points: 15	Points: 16
Indoor	Indoor	Outdoor	Outdoor	Indoor
Water Town	**Nafutsha Factory**	**City**	**Roadside Vehicles**	**Town**
Points: 20	Points: 12	Points: 10	Points: 18	Points: 26
Outdoor	Outdoor	Outdoor	Outdoor	Outdoor

Given an original embodied interaction video together with its control annotations, we use Seedance 2.0 as an instruction-guided video editing model to generate visually augmented versions of the same interaction process. We use Qwen3.6-Plus [30] to batch-generate such instructions. The editing process is explicitly restricted to local appearance variations. For example, the color of the manipulated object can be changed, such as replacing the color of a shoe, while keeping surrounding objects, the background, object geometry, camera trajectory, hand or robot motion, contact relations, and the final task state unchanged. Similarly, for grasping or lifting sequences, the appearance of the manipulated object can be replaced or modified while preserving the original grasp trajectory, the object motion pattern, and the final lifted pose. In this way, the augmented samples introduce new visual variations without altering the underlying embodied interaction behavior.

Since the augmentation process is designed to preserve the physical interaction process, the generated videos inherit the control annotations of the original samples, including camera trajectories, action maps, and low-dimensional robot action vectors. When the modified visual attributes are explicitly mentioned in the text description, we update the caption accordingly; otherwise, we retain the original high-level task description. In addition, we filter out edited samples that change the task semantics, alter object motion, break contact relations, modify the final task outcome, or introduce obvious temporal inconsistency artifacts, thereby ensuring that the augmented data remain aligned with the original demonstrations in both temporal consistency and physical semantics.

This pipeline provides a practical way to scale the training data for embodied world models without recollecting robot or human demonstrations. By decoupling visual appearance variation from action supervision, the model can observe the same control signal under different visual realizations, which encourages it to learn action-conditioned physical dynamics rather than memorizing specific object textures, colors, or scene appearances. As a result, this data augmentation strategy improves robustness to object appearance changes, scene style variations, and out-of-distribution visual conditions.



Figure 8: **Seedance 2.0 Data Augmentation.** We use Seedance 2.0 to diversify the embodied data in terms of object color, tabletop material, and manipulated objects. The results demonstrate consistent generation quality and interaction dynamics while keeping the robotic arm motion unchanged.

C.5 Hand-Joint Data Construction

Unified Hand Annotation with HaMeR. To remove annotation-format discrepancies between EgoDex and Ego4D, we process all selected video frames with a unified HaMeR annotation pipeline. HaMeR [29] is a fully Transformer-based method for hand mesh recovery. It uses a large-scale Vision Transformer backbone and a Transformer decoder to regress hand parameters from a monocular RGB image. The method adopts the MANO parametric hand model [31] to represent hand geometry. Given a hand crop I_t, HaMeR predicts the MANO pose parameters, shape parameters, and camera parameters:

$$\Theta_t = f_{\text{HaMeR}}(I_t) = \{\theta_t, \beta_t, \pi_t\}.$$

Here, θ_t denotes the articulated hand pose, β_t denotes the hand shape, and π_t denotes the camera-related parameters. The MANO layer then produces a hand mesh and 21 3D joints:

$$(M_t, X_t) = \text{MANO}(\theta_t, \beta_t), \quad X_t \in \mathbb{R}^{21\times 3}.$$

We process the left and right hands separately. For each frame t and hand $h \in \{L, R\}$, the annotation consists of 3D joint coordinates $X_{t,h} \in \mathbb{R}^{21\times 3}$, joint pose matrices $R_{t,h} \in SO(3)^{21}$, and 2D pixel coordinates $U_{t,h} \in \mathbb{R}^{21\times 2}$ obtained by projecting the 3D joints into the video plane. Both hands preserve the standard 21-joint hand topology, including the wrist and finger joints. Frames with failed hand detection, severe occlusion, implausible reprojection, or unstable temporal behavior are filtered using detection confidence, joint reprojection validity, and temporal consistency checks.

Camera-Aware 3D-to-2D Projection. To align the hand-control signal with the RGB video pixels, we transform the reconstructed 3D joints into the camera coordinate system of each video frame and project them to the image plane with the corresponding camera intrinsics. Let $X^w_{t,h,j} \in \mathbb{R}^3$ be the 3D coordinate of the j-th joint of hand h at frame t in the world or reconstruction coordinate system, let $T_t^{c\leftarrow w} = [R_t^{c\leftarrow w} \mid \mathbf{t}_t^{c\leftarrow w}]$ be the camera extrinsic matrix, and let K_t be the camera intrinsic matrix. The camera-space joint is computed as

$$\tilde{X}^c_{t,h,j} = R_t^{c\leftarrow w} X^w_{t,h,j} + \mathbf{t}_t^{c\leftarrow w}.$$

The homogeneous image projection is

$$\tilde{u}_{t,h,j} \sim K_t \tilde{X}^c_{t,h,j},$$

and the pixel coordinate is obtained by perspective division:

$$u_{t,h,j} = \left(\frac{\tilde{u}^{(x)}_{t,h,j}}{\tilde{u}^{(z)}_{t,h,j}}, \frac{\tilde{u}^{(y)}_{t,h,j}}{\tilde{u}^{(z)}_{t,h,j}} \right).$$

For EgoDex, the original release provides camera intrinsics and per-frame camera/joint SE(3) transforms, which allows direct reprojection into the video frame. For Ego4D, we apply the same projection procedure using the available video metadata or estimated camera parameters. After this step, each hand joint is represented by a 2D coordinate in the same pixel coordinate system as the corresponding RGB frame.

Action-Map Representation. Given the projected left- and right-hand joints

$$U_t = \{U_{t,L}, U_{t,R}\}, \quad U_{t,h} \in \mathbb{R}^{21\times 2},$$

we construct a hand-joint action map A_t for each video frame. The action map is a pixel-aligned control condition that binds the 3D hand motion signal to the 2D video space. We rasterize the 42 hand joints from both hands in the original image coordinate system while preserving hand identity and joint topology. The action map can be represented as multi-channel joint heatmaps, sparse coordinate maps, or skeleton-rendered maps. During Worldscape-MoE training, A_t shares the same temporal index and spatial coordinate frame as the RGB frame I_t, and can therefore be directly used as the control input for the diffusion world model. For a video clip $\{I_t\}_{t=1}^T$, we obtain a synchronized control sequence $\{A_t\}_{t=1}^T$, enabling the model to learn the dynamic correspondence among hand motion, object response, and scene evolution.

D Additional Experimental Details

D.1 Model Inference Configurations

Table 5 summarizes the detailed inference configurations of the evaluated world-generation models, including model version, capability type, generation resolution, video length, and frame rate. These settings are provided to ensure a fair comparison across models with different official releases and inference protocols.

Table 5: Detailed inference configurations of world generation models evaluated in our benchmark. Version denotes the official release date of each model.

Method	Version	Ability	Resolution	Length (s)	FPS
Matrix-Game 3.0	25.08.12	I2V	640×352	3.4	24
HY-World 1.5	25.12.17	I2V	832×480	3.2	24
CameraCtrl	24.04.02	I2V	256×256	7.6	10
MotionCtrl	23.12.06	I2V	256×256	7.6	10
CamI2V	25.07.12	I2V	512×320	7.6	10
RealCam-I2V	25.07.12	I2V	896×512	5	16
VideoX-Fun-Wan	25.10.16	I2V	832×480	5	16
AC3D	25.04.01	I2V	720×480	6	8
ASTRA	25.12.09	I2V	832×480	4	20
HunyuanVideo-1.5	25.11.20	I2V	848×480	5	24
Cosmos-Predict 2.5	25.10.06	I2V	1280×704	5	16
MimicMotion	24.07.08	I2V	1024×576	4.8	15
MagicDance	24.04.03	I2V	512×512	4	15
LOME	26.04.05	I2V	832×480	5	15

D.2 Details of EWM Score

Embodied World Models (EWMs) are generative models designed to predict future environment states conditioned on current observations and external instructions or robot actions [48]. Unlike general-purpose video generation models, EWMs must capture physically grounded and action-consistent dynamics, effectively functioning as mental simulators for action planning [35, 27] or environment proxies for scalable robotic training.

To systematically evaluate the capabilities of our proposed framework and existing baselines, we adopt the EWMScore, a holistic metric originating from the WorldArena [33] benchmark. WorldArena provides a unified evaluation framework specifically designed for embodied world models, assessing both their perceptual fidelity and functional utility in downstream decision-making tasks. The EWMScore integrates multi-dimensional performance into a single interpretable index by computing the arithmetic mean of 16 normalized video metrics.

Table 6: Video quality evaluation results across visual quality, motion quality and content consistency dimensions.

Models	Visual Quality			Motion Quality			Content Consistency		
	Image Quality	Aesthetic Quality	JEPA Similarity	Dynamic Degree	Flow Score	Motion Smoothness	Subject Consist.	Background Consist.	Photometric Consist.
Worldscape-MoE	0.4566	0.3795	0.8920	0.4373	0.2632	0.7717	0.8333	0.9043	0.1439
w/o MoE	0.5220	0.4053	0.8779	0.4432	0.2457	0.7776	0.8282	0.8990	0.1126
GigaWorld-0	0.5041	0.3991	0.4413	0.6709	0.3118	0.7811	0.7303	0.8563	0.1756
TesserAct	0.3322	0.4590	0.4579	0.5150	0.2447	0.7579	0.8250	**0.9238**	0.2491
RoboMaster	0.3487	0.3842	0.2966	0.6124	0.1484	0.6940	0.8295	0.9123	0.3356
Vidar	0.4145	0.4068	0.5608	0.2767	0.1426	0.7973	0.7629	0.8300	0.2350
Cosmos-Predict 2.5 (text)	0.6668	0.4501	0.3126	0.5911	0.4302	0.7882	0.7488	0.8511	0.1383
Cosmos-Predict 2.5 (action)	0.4489	0.3576	0.9296	0.3994	0.0573	0.7100	0.8197	0.8894	0.3528
CtrlWorld	0.3522	0.3893	0.9185	0.4257	0.3449	0.7377	**0.8411**	0.9057	0.1729
Wan 2.2	0.3884	0.3963	0.7575	0.4349	0.1269	0.7019	0.8388	0.9042	**0.4776**
CogvideoX	0.3582	0.3777	**0.9384**	0.3166	0.2189	0.7391	0.8083	0.8773	0.3580
IRASim	0.3489	0.3623	0.9330	0.4139	0.2083	0.7052	0.8312	0.9068	0.3522
Veo 3.1	0.6605	**0.4632**	0.5694	0.5450	0.1396	0.6989	0.7878	0.8710	0.3247
Wan 2.6	**0.6824**	0.4433	0.7229	**0.7421**	**0.4532**	**0.8539**	0.7517	0.8687	0.1904

Table 7: Video quality evaluation results across physics adherence, 3D accuracy and controllability dimensions.

Models	Physics Adherence		3D Accuracy		Controllability		
	Interaction Quality	Trajectory Acc.	Depth Acc.	Perspectivity	Instruction Following	Semantic Alignment	Action Following
Worldscape-MoE	**0.8008**	0.4610	0.9030	0.9686	**0.9348**	**0.9039**	0.0955
w/o MoE	0.7622	0.3540	0.9038	**0.9744**	0.8703	0.8914	0.0324
GigaWorld-0	0.5368	0.1552	0.6316	0.7596	0.6156	0.8591	0.1134
TesserAct	0.5800	0.1396	0.7159	0.7920	0.6152	0.8783	0.0311
RoboMaster	0.5364	0.1158	0.8335	0.7588	0.5772	0.8761	0.0352
Vidar	0.5348	0.1928	0.7872	0.7592	0.5912	0.8826	0.0819
Cosmos-Predict 2.5 (text)	0.3872	0.0816	0.7051	0.7964	0.2664	0.7733	**0.1418**
Cosmos-Predict 2.5(action)	0.5500	0.2945	0.8862	0.7644	0.5840	0.8879	0.0133
CtrlWorld	0.6212	**0.4766**	0.9300	0.7960	0.7272	0.8912	0.0210
Wan 2.2	0.5184	0.1627	0.7768	0.7660	0.5376	0.8877	0.0512
CogvideoX	0.5940	0.3526	0.9097	0.7828	0.7268	0.8977	0.0076
IRASim	0.5656	0.3639	**0.9312**	0.7788	0.6604	0.8849	0.0526
Veo 3.1	0.7872	0.1231	0.7421	0.8276	0.9328	0.8607	0.0852
Wan 2.6	0.7280	0.1182	0.7144	0.8032	0.8536	0.8728	0.0992

These metrics are systematically categorized into six major evaluation dimensions:

- **Visual Quality**: Assesses whether the generated videos are perceptually reliable for embodied scenarios, evaluating low-level fidelity, perceptual appeal, and similarity to real data. The sub-metrics include Image Quality, Aesthetic Quality, and JEPA Similarity.
- **Motion Quality**: Reflects whether the model captures physically meaningful and temporally coherent dynamics. The sub-metrics consist of Dynamic Degree, Flow Score, and Motion Smoothness.

- **Content Consistency**: Measures the stability of objects and scenes throughout the video at both semantic and appearance levels. The sub-metrics are Subject Consistency, Background Consistency, and Photometric Consistency.
- **Physics Adherence**: Evaluates whether generated behaviors conform to real-world physical constraints rather than merely appearing visually plausible. The sub-metrics are Interaction Quality and Trajectory Accuracy.
- **3D Accuracy**: Assesses the preservation of real-world spatial structures and geometric consistency beyond 2D image appearance. The sub-metrics include Depth Accuracy and Perspectivity.
- **Controllability**: Measures the model's ability to accurately respond to and execute external action instructions. The sub-metrics consist of Instruction Following, Semantic Alignment, and Action Following.

The detailed quantitative evaluation results of our model and all other baselines across these dimensions are provided in the appendix tables. Specifically, Table 6 reports the video quality evaluation results across the visual quality, motion quality, and content consistency dimensions. Meanwhile, Table 7 details the evaluation results across the physics adherence, 3D accuracy, and controllability dimensions.

Analysis of Worldscape-MoE on EWMScore. As shown in Tables 6 and 7, Worldscape-MoE delivers consistently leading performance on the dimensions that most directly characterize an *embodied* world model. In terms of **Physics Adherence**, it attains the best Interaction Quality (**0.8008**) among all baselines, surpassing the strongest closed-source models Veo 3.1 (0.7872) and Wan 2.6 (0.7280) and exceeding action-conditioned EWMs such as Cosmos-Predict 2.5 (action) (0.5500) and IRASim (0.5656) by a substantial margin, while its Trajectory Accuracy (0.4610) ranks second only to CtrlWorld (0.4766). On **3D Accuracy**, the Without MoE baseline obtains the highest Perspectivity (0.9744), while Worldscape-MoE remains highly competitive on Perspectivity (0.9686) and Depth Accuracy (0.9030), demonstrating that the generated futures faithfully preserve real-world spatial structure rather than relying on 2D appearance shortcuts. On **Controllability**, Worldscape-MoE further establishes new state-of-the-art results on both Instruction Following (**0.9348**) and Semantic Alignment (**0.9039**), outperforming all baselines, which indicates that Worldscape-MoE grounds external textual instructions far more reliably than appearance-driven generators. On the appearance-oriented dimensions, Worldscape-MoE remains competitive, with a strong JEPA Similarity (0.8920) and high Subject Consistency (0.8333) and Background Consistency (0.9043). Taken together, these results show that Worldscape-MoE advances the frontier of EWMs along the physically grounded, geometry-consistent, and instruction-controllable axes that the EWMScore is specifically designed to measure.

D.3 MoE Scalability Analysis



Figure 9: **Transient Degradation and Recovery During Modality Extension.** Qualitative results from 500 to 6500 training steps show the transient degradation of locomotion capability after introducing a new control modality, followed by gradual recovery as training proceeds. This effect is particularly pronounced in scenarios involving substantial scene changes.

We analyzed the model convergence status and feature evolution under multimodal control by tracking the weight updates of each expert module across different training checkpoints.

Weight update analysis. We track the cumulative L2 variation of each expert module across consecutive 500-step intervals in Table 8. The Shared expert remains the most stable throughout training, while Locomotion, Manipulation, and ActionMap exhibit larger but comparable update

Table 8: Expert weight L2 norm variation across different training step segments.

Expert	250-750	750-1.25k	1.25-1.75k	1.75-2.25k	2.25-2.75k	2.75-3.25k	3.25-3.75k	3.75-4.25k
Shared	67.78	70.25	71.13	65.81	64.35	64.50	67.92	67.99
Locomotion	115.60	122.62	125.02	116.48	116.34	119.70	138.31	130.17
Manipulation	134.89	137.50	139.52	126.24	123.10	122.23	126.33	125.71
ActionMap	155.30	144.04	135.25	134.97	134.43	132.53	144.95	138.01

magnitudes, with ActionMap showing the strongest early adaptation and all three modality-specific experts gradually stabilizing after the first few segments. This pattern indicates that joint mixed training enables balanced co-adaptation across experts, rather than suppressing the locomotion pathway when new control modalities are introduced.

Joint Training Strategy. Experimental results show that the expert initialization strategy significantly impacts the multimodal decoupling effect.If parameters are cloned from a pre-trained Locomotion expert to initiate mixed training for other experts, it leads to a significantly smaller magnitude of weight updates for the Locomotion expert. This approach distorts the original feature space, resulting in a partial collapse of locomotion control performance. In contrast, uniformly cloning and initializing all experts from a Dense Model backbone for multi-expert joint training facilitates a more stable and balanced decoupling effect across all experts.

Transient Degradation During Modality Extension. When a new control modality is introduced, we observe a short period of locomotion degradation in the early stage of training, as shown in Fig. 9. We attribute this behavior to a temporary adaptation phase, during which the shared and modality-specific parameters are jointly re-adjusted to absorb the newly introduced control knowledge. This phenomenon is particularly pronounced in locomotion cases that require large scene changes. Importantly, the degradation is not persistent: as training continues, the model gradually restores its original locomotion capability while retaining the newly learned modality. In the following, we provide qualitative cases from 500 to 6500 training steps to visualize this recovery process.

D.4 Generalization to Single-Arm LIBERO Manipulation

To further examine whether the MoE-trained shared expert learns transferable embodied dynamics, we conduct an additional extension experiment on the LIBERO single-arm manipulation benchmark. In this setting, each sample consists of an initial observation, a task language instruction, and the corresponding low-dimensional single-arm robot action trajectory from the LIBERO demonstrations. We process these demonstrations into action-conditioned video clips and use the official train/validation metadata for fine-tuning and held-out evaluation. Specifically, we initialize a single-FFN model from the MoE shared expert backbone and fine-tune it on the LIBERO single-arm split. As shown in Fig. 10, after only 5K fine-tuning steps, the shared-initialized model already produces action-conditioned rollouts that closely follow the ground-truth demonstrations in object motion, end-effector movement, and scene consistency. Importantly, this adaptation does not compromise the model's generative capability on other control modalities, which remains effective after fine-tuning, suggesting that the shared expert preserves broadly reusable world dynamics rather than overfitting to the single-arm manipulation domain.

This result provides qualitative evidence that the shared expert does not merely serve the original heterogeneous-control training tasks. Instead, it captures manipulation-relevant priors that can be transferred to a new single-arm setting with limited additional training. We emphasize this as a qualitative generalization study rather than a large-scale quantitative comparison: the goal is to show that the MoE-shared backbone can quickly adapt to unseen LIBERO manipulation cases and generate physically plausible futures under robot-action conditioning.

D.5 More Qualitative Results of Worldscape-MoE

In this section, we present more qualitative results of Worldscape-MoE.

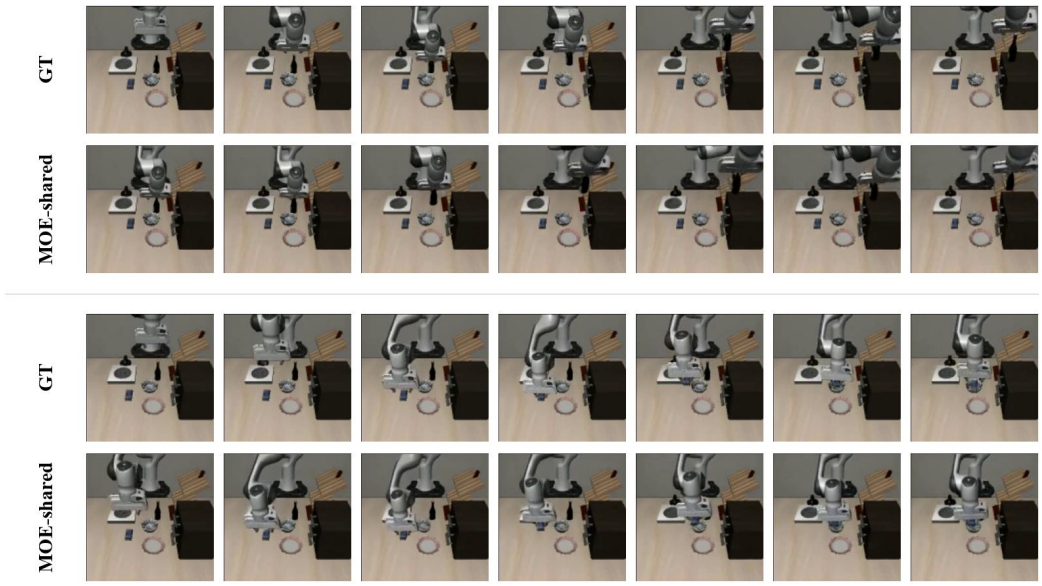


Figure 10: **Single-arm LIBERO generalization.** Each panel compares the ground-truth demonstration with the model initialized from the MoE shared expert after 5K fine-tuning steps. The figure is pre-rendered from normalized comparison videos by uniformly sampling seven frames from each rollout.

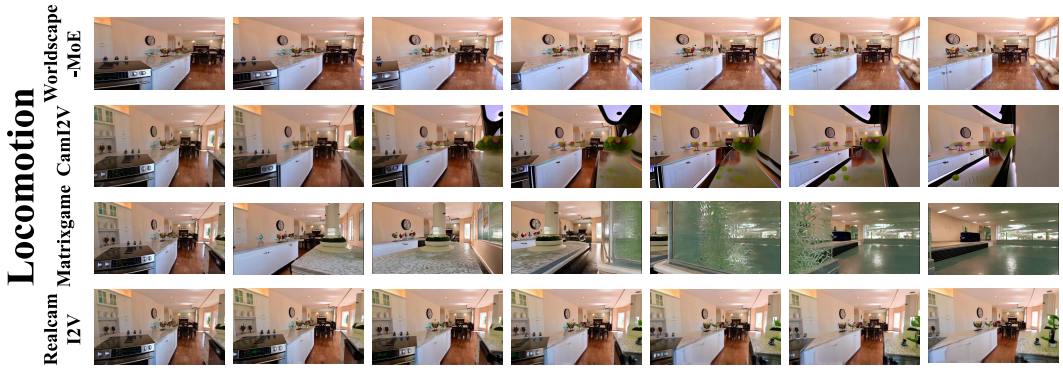


Figure 11: **Locomotion comparison.** Each row compares video frames generated by Worldscape-MoE, CamI2V, Matrixgame, and RealCam-I2V under the same locomotion trajectory, with columns uniformly sampled from each rollout.



Figure 12: **Locomotion comparison.** Each row compares video frames generated by Worldscape-MoE, Astra, Matrixgame, and RealCam-I2V under the same locomotion trajectory, with columns uniformly sampled from each rollout.



Figure 13: **Locomotion comparison.** Each row compares video frames generated by Worldscape-MoE, Astra, Matrixgame, and MotionCtrl under the same locomotion trajectory, with columns uniformly sampled from each rollout.

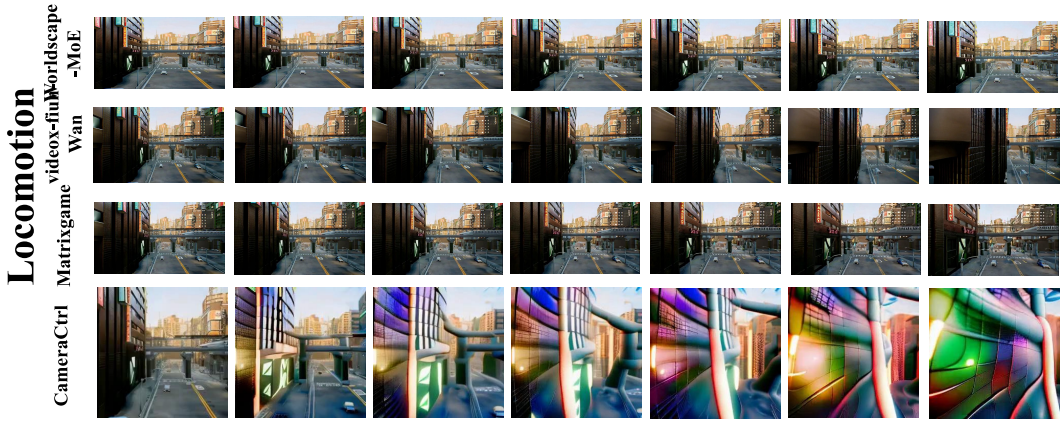


Figure 14: **Locomotion comparison.** Each row compares video frames generated by Worldscape-MoE, VideoX-fun-Wan, Matrixgame, and CamTrackCtrl under the same locomotion trajectory, with columns uniformly sampled from each rollout.



Figure 15: **Locomotion comparison.** Each row compares video frames generated by Worldscape-MoE, CamI2V, Matrixgame, and VideoX-fun-Wan under the same locomotion trajectory, with columns uniformly sampled from each rollout.

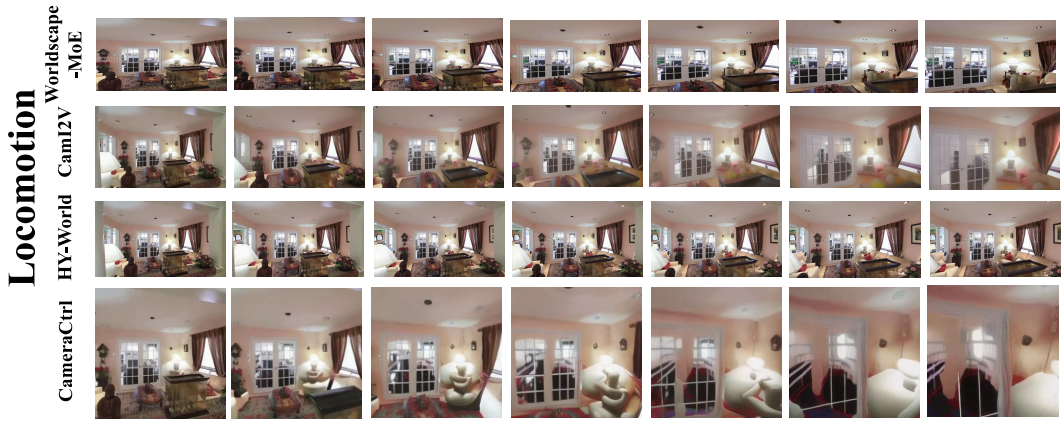


Figure 16: **Locomotion comparison.** Each row compares video frames generated by Worldscape-MoE, CamI2V, HY-World, and CameraCtrl under the same locomotion trajectory, with columns uniformly sampled from each rollout.

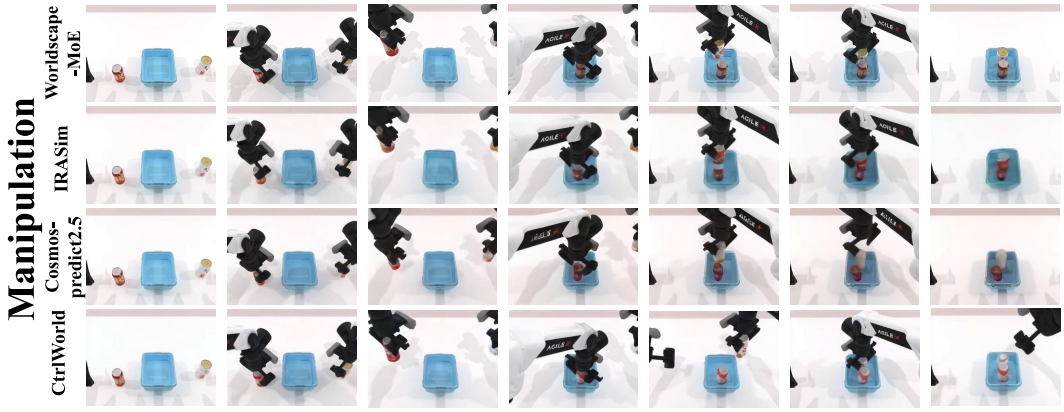


Figure 17: **Manipulation comparison.** Each row compares video frames generated by Worldscape-MoE, IRASim, Cosmos-predict2.5, and CtrlWorld under the same manipulation trajectory, with columns uniformly sampled from each rollout.

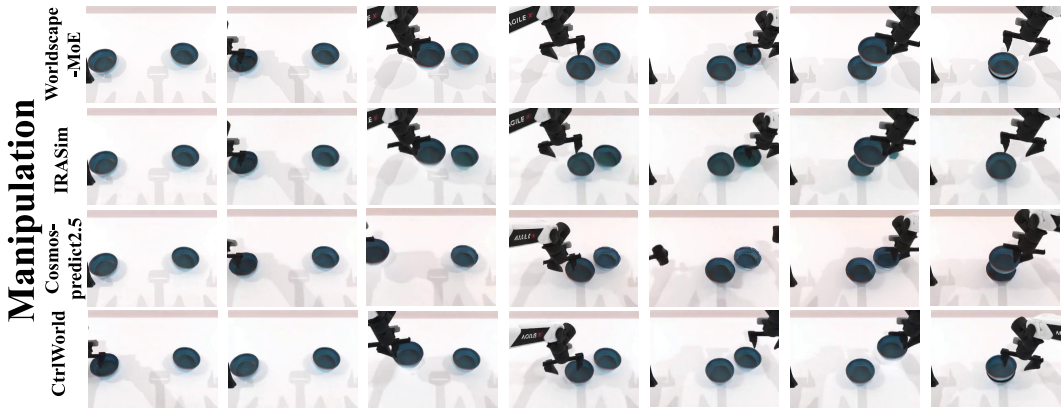


Figure 18: **Manipulation comparison.** Each row compares video frames generated by Worldscape-MoE, IRASim, Cosmos-predict2.5, and CtrlWorld under the same manipulation trajectory, with columns uniformly sampled from each rollout.

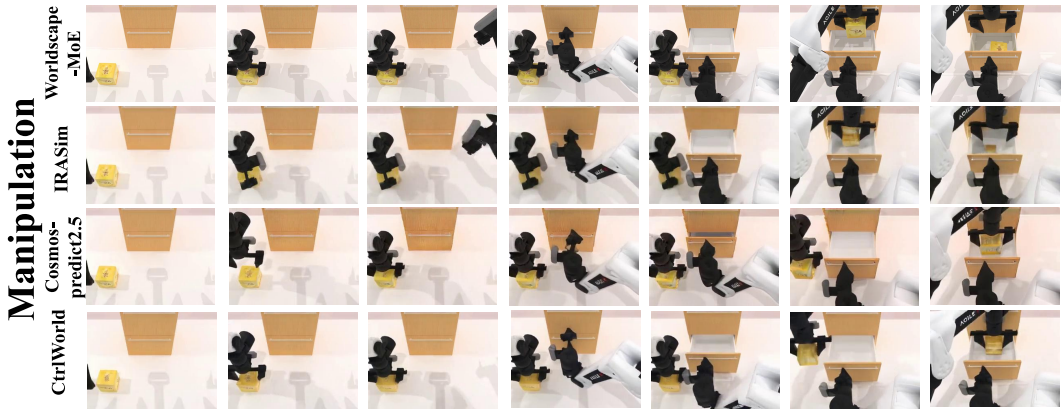


Figure 19: **Manipulation comparison.** Each row compares video frames generated by Worldscape-MoE, IRASim, Cosmos-predict2.5, and CtrlWorld under the same manipulation trajectory, with columns uniformly sampled from each rollout.

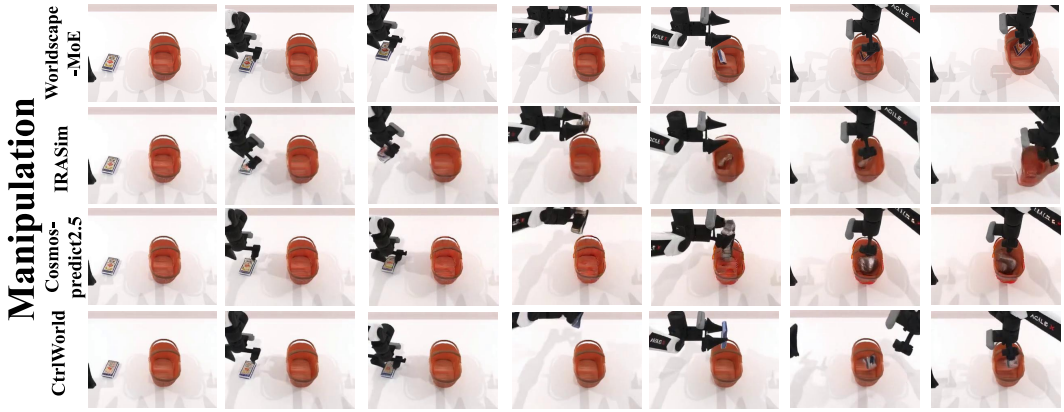


Figure 20: **Manipulation comparison.** Each row compares video frames generated by Worldscape-MoE, IRASim, Cosmos-predict2.5, and CtrlWorld under the same manipulation trajectory, with columns uniformly sampled from each rollout.

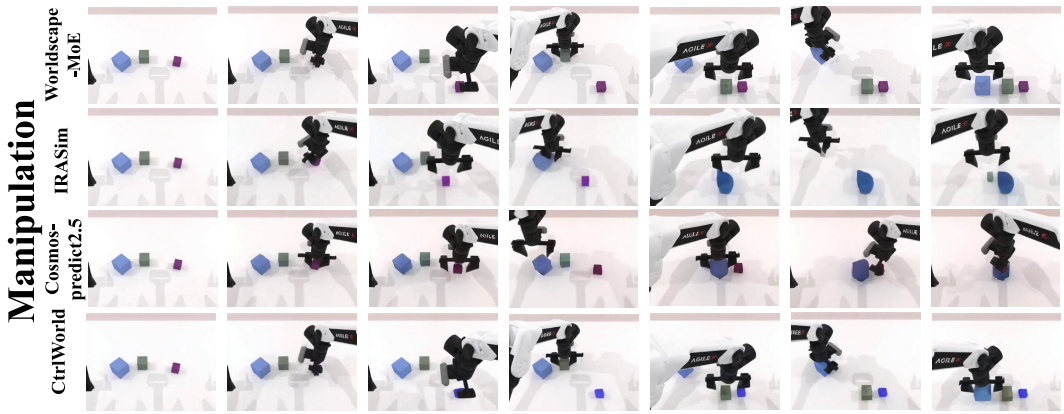


Figure 21: **Manipulation comparison.** Each row compares video frames generated by Worldscape-MoE, IRASim, Cosmos-predict2.5, and CtrlWorld under the same manipulation trajectory, with columns uniformly sampled from each rollout.

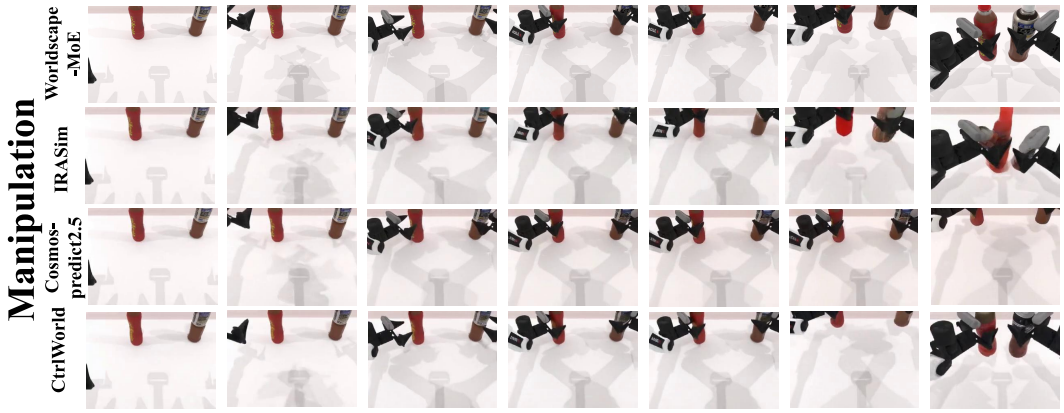


Figure 22: **Manipulation comparison.** Each row compares video frames generated by Worldscape-MoE, IRASim, Cosmos-predict2.5, and CtrlWorld under the same manipulation trajectory, with columns uniformly sampled from each rollout.

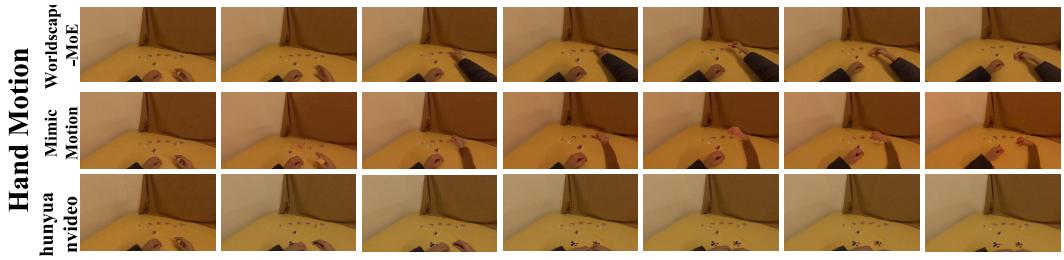


Figure 23: **Hand motion comparison.** Each row compares video frames generated by Worldscape-MoE, Motion, and nvideo under the same hand motion trajectory, with columns uniformly sampled from each rollout.



Figure 24: **Hand motion comparison.** Each row compares video frames generated by Worldscape-MoE, MimicMotion, and HunyuanVideo under the same hand motion trajectory, with columns uniformly sampled from each rollout.

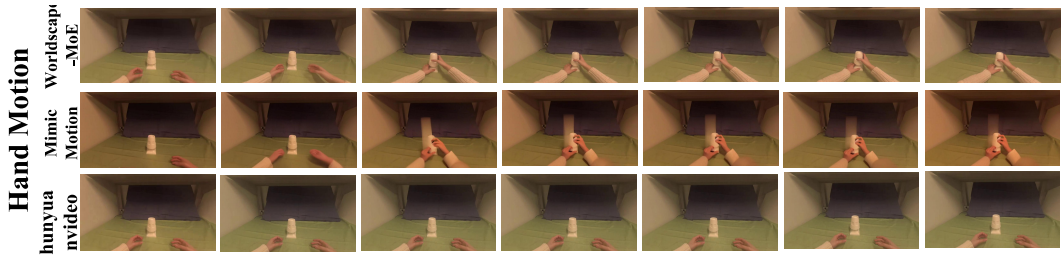


Figure 25: **Hand motion comparison.** Each row compares video frames generated by Worldscape-MoE, MimicMotion, and HunyuanVideo under the same hand motion trajectory, with columns uniformly sampled from each rollout.

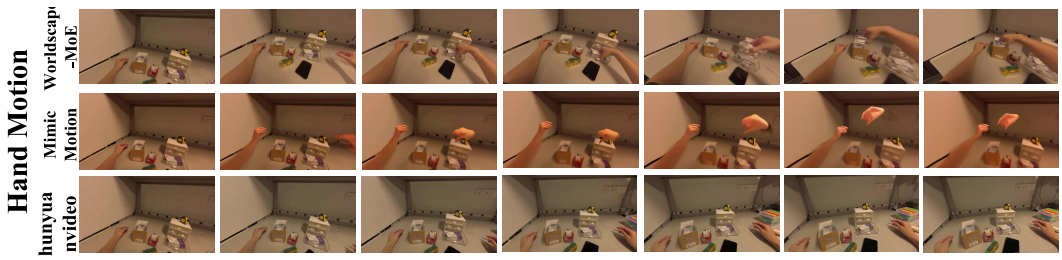


Figure 26: **Hand motion comparison.** Each row compares video frames generated by Worldscape-MoE, MimicMotion, and HunyuanVideo under the same hand motion trajectory, with columns uniformly sampled from each rollout.

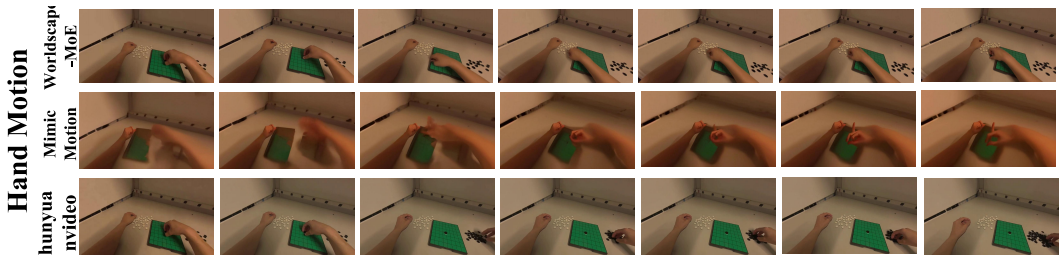


Figure 27: **Hand motion comparison.** Each row compares video frames generated by Worldscape-MoE, MimicMotion, and HunyuanVideo under the same hand motion trajectory, with columns uniformly sampled from each rollout.

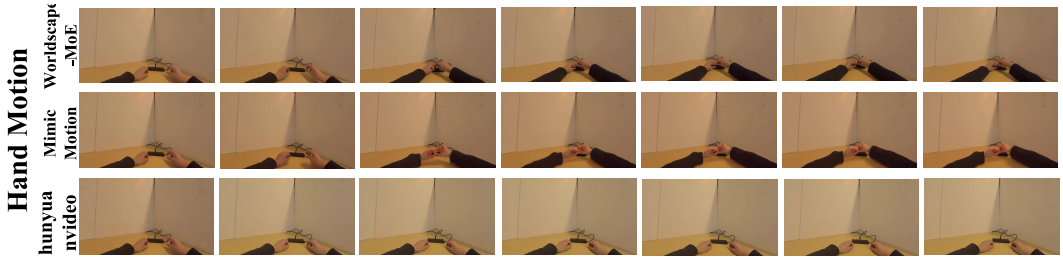


Figure 28: **Hand motion comparison.** Each row compares video frames generated by Worldscape-MoE, MimicMotion, and HunyuanVideo under the same hand motion trajectory, with columns uniformly sampled from each rollout.

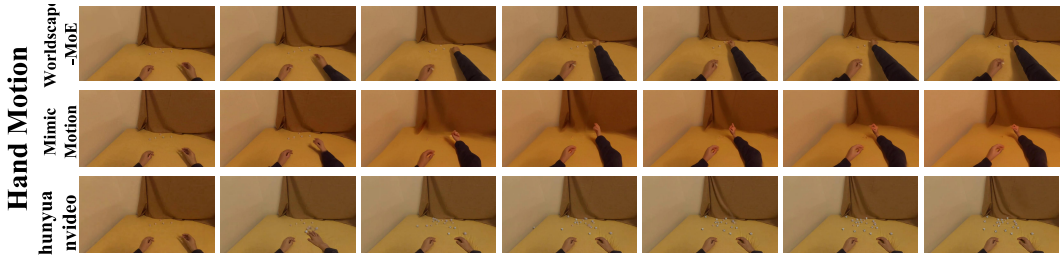


Figure 29: **Hand motion comparison.** Each row compares video frames generated by Worldscape-MoE, MimicMotion, and HunyuanVideo under the same hand motion trajectory, with columns uniformly sampled from each rollout.

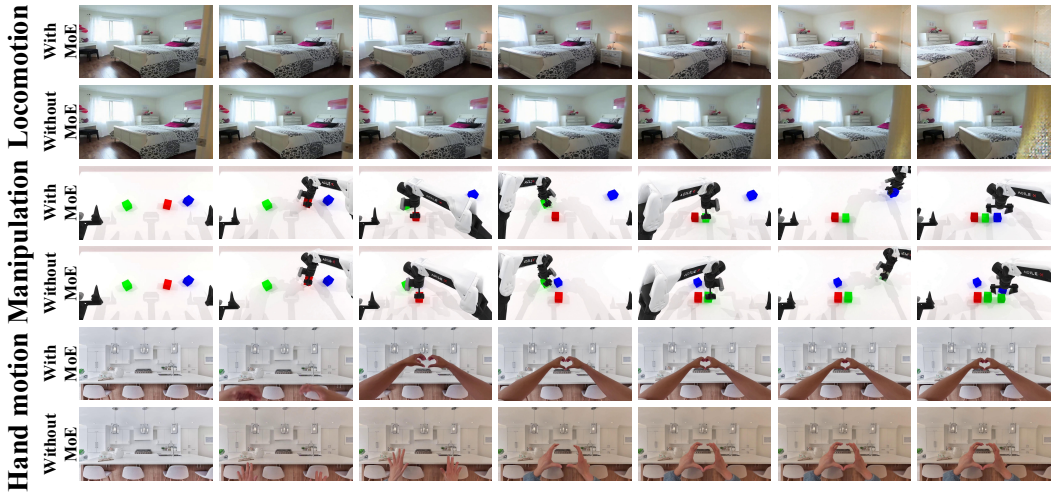


Figure 30: **Effect of MoE across diverse motion scenarios.** We compare generated rollouts with and without MoE under three representative settings: locomotion, manipulation, and hand motion. Each row shows uniformly sampled frames from one rollout, illustrating the effect of MoE on temporal consistency and motion-conditioned generation quality.

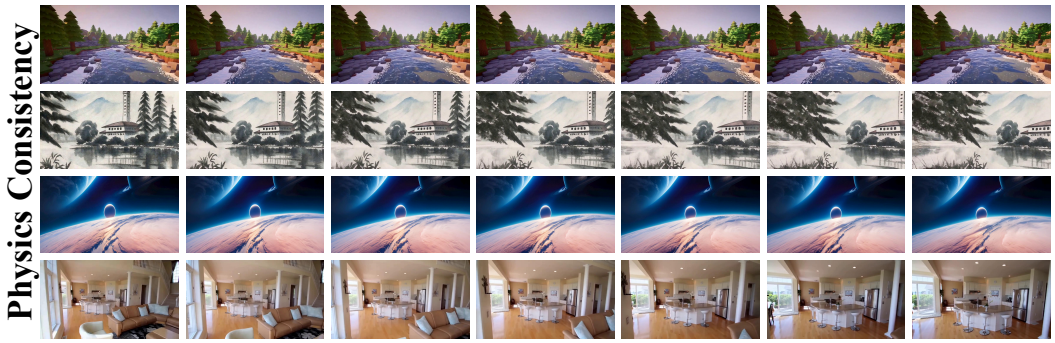


Figure 31: **Physics consistency across diverse visual domains.** The figure shows representative generated rollouts under different scene styles, including natural landscapes, ink-wash scenes, space environments, and indoor rooms. Each row contains uniformly sampled frames from one rollout, illustrating that the generated trajectories preserve temporal and physical consistency across visually distinct domains.



Figure 32: **Loco-manipulation generation across diverse scenes.** The figure presents representative rollouts involving coordinated locomotion and object interaction. Each row contains uniformly sampled frames from one rollout, showing how the agent approaches and manipulates targets under different visual environments.

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 novel training framework for world models, as evidenced by both the methodology and the experimental validation.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: A detailed discussion of the limitations is provided in Appendix B.

   Guidelines:

   - The answer [N/A] 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: [N/A]

Justification: This paper does not involve theoretical results.

Guidelines:

- The answer [N/A] 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: The information provided in the paper is sufficient to reproduce the main experimental results underlying the paper's primary claims and conclusions.

   Guidelines:

   - The answer [N/A] 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: Please refer to the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides detailed experimental settings and implementation details in Section 3 and Appendix D.

   Guidelines:

   - The answer [N/A] 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: Please refer to the implementation details in Section 3.

   Guidelines:

   - The answer [N/A] 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: Please refer to the implementation compute resources details in Section 3.

   Guidelines:

   - The answer [N/A] 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 work fully complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] 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: Please refer to Section A for a discussion of broader impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not involve high-risk data. All training data are derived from open-source datasets. Under this category, the trained model does not present identifiable risks.

    Guidelines:

    - The answer [N/A] 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 provide a detailed description of all data sources and codebases in the appendix, all of which are derived from the open-source community.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: Please refer to Appendix C.2 and C.4 for a detailed description.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This work does not involve any human participants.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This work does not involve any human participants.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

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

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.