LongScape: Advancing Long-Horizon Embodied World Models with Context-Aware MoE

Anonymous Author(s)
Affiliation
Address
email

Abstract

Video-based world models hold significant potential for generating high-quality embodied manipulation data. However, current video generation methods struggle to achieve stable long-horizon generation: classical diffusion-based approaches often suffer from temporal inconsistency and visual drift over multiple rollouts, while autoregressive methods tend to compromise on visual detail. To solve this, we introduce LongScape, a hybrid framework that adaptively combines intra-chunk diffusion denoising with inter-chunk autoregressive causal generation. Our core innovation is an action-guided, variable-length chunking mechanism that partitions video based on the semantic context of robotic actions. This ensures each chunk represents a complete, coherent action, enabling the model to flexibly generate diverse dynamics. We further introduce a Context-aware Mixture-of-Experts (CMoE) framework that adaptively activates specialized experts for each chunk during generation, guaranteeing high visual quality and seamless chunk transitions. Extensive experiments on two embodied benchmarks demonstrate that our method achieves state-of-the-art generation performance. Notably, LongScape can support stable and temporally consistent long-horizon generation over sequences exceeding 1,000 frames. Our code is available at: https://anonymous.4open.science/r/AMSVVD-fdg245.

1 Introduction

Video-based world models have become a prominent research direction in embodied intelligence [1, 2, 3, 4]. By learning environmental dynamics from embodied video data, these models function as powerful synthetic data engines capable of generating large-scale simulated experience to train downstream embodied policy models (such as VLA), thereby alleviating the data scarcity problem in embodied learning [5, 6, 7]. However, current world models are predominantly limited to generating only very short video clips. For more long-horizon embodied tasks in the real world, these models struggle to provide stable generation quality, posing a major challenge to their practical deployment.

Existing video world models broadly fall into diffusion-based, autoregressive, and hybrid categories, each with inherent limitations for long-horizon generation. Diffusion models [8, 2] lack explicit causal structure and often suffer from temporal drift over extended rollouts. Autoregressive models [9, 10] preserve temporal causality but compromise on visual fidelity. Hybrid approaches [11, 12, 13] combine diffusion denoising with autoregressive generation, yet existing methods rely on fixed-length chunks that arbitrarily segment semantically continuous actions or conflate distinct motion patterns (e.g., locomotion and manipulation) into a single chunk. This semantic ambiguity misguides next-chunk prediction and degrades long-range temporal coherence.

To address these challenges, we introduce **LongScape**, an embodied world model designed for long-horizon video generation. Our framework adopts a hybrid approach combining intra-chunk diffusion

denoising with inter-chunk autoregressive causal generation. The key innovation is a context-adaptive, variable-length chunking mechanism guided by robot action priors: frames with gripper state changes or substantial effector motion are partitioned into shorter chunks for fine-grained manipulation, while frames with minimal motion form longer chunks for efficient generation. We pre-process embodied datasets with this action-based rule to create chunks of four distinct granularities, and train multiple Diffusion Transformer (DiT) experts, each specializing in a specific temporal dynamic (e.g., fine manipulation vs. locomotion). Unlike conventional MoE architectures [14, 15] that emphasize load balancing, our experts are explicitly specialized by task context. A context-based dynamic router, conditioned on text instructions and current visual features, selects the appropriate expert at each rollout step, ensuring semantic coherence and causal consistency throughout the generated sequence.

We conduct extensive experiments on the LIBERO [16] and AGIBOT-World [17] benchmarks. Our model achieves state-of-the-art performance in video generation quality compared to existing diffusion, autoregressive, and hybrid baselines, maintaining visual coherence and stability over 50 rollouts and 1000 frames. Our primary contributions are:

- We propose LongScape, a novel embodied world model with adaptive chunk rollout that enables stable long-horizon video generation.
- We introduce an action-prior-guided chunk partitioning scheme that ensures semantic coherence of video chunks for effective causal autoregressive generation.
- Extensive experiments demonstrate state-of-the-art long-horizon generation quality with sustained stability in visual quality and motion correctness.

2 Related Works

2.1 Video Generation Models

Video generation has been approached through diffusion models, autoregressive models, and hybrid frameworks. Diffusion models [18, 19, 20, 21] iteratively denoise latent representations and achieve high visual fidelity but lack explicit temporal causality, leading to drift over long sequences. Autoregressive models [7] formulate generation as next-token prediction, which preserves causality but often yields lower visual quality. Hybrid models [11, 12] combine diffusion denoising within an autoregressive chain to balance both aspects. However, existing hybrid methods employ fixed-length chunks that ignore the varying temporal dynamics of embodied tasks, where manipulation and locomotion demand different temporal granularities.

2.2 Long-Horizon Video Generation

Long-video generation approaches can be grouped into divide-and-conquer strategies (e.g., hierarchical keyframe planning [22, 23, 24]) and temporal autoregressive methods that generate chunk-by-chunk [25, 12, 13]. Both families share a common limitation: fixed temporal granularity. This rigidity prevents adaptation to the heterogeneous motion patterns inherent in embodied tasks. Our LongScape framework introduces adaptive variable-length chunking driven by action priors, applying fine-grained diffusion to short chunks during critical interactions and efficient generation to longer chunks during stable periods.

2.3 Embodied World Models

Embodied world models learn environmental dynamics to support synthetic data generation and policy learning [26]. They broadly follow two paradigms: video generation-based methods that adapt foundation models for text- or action-conditioned synthesis [10, 2, 27, 5, 3, 4], and latent prediction methods that forecast compact state representations for planning [28, 29]. Our work belongs to the text-conditioned video generation paradigm, with a focus on enabling stable long-horizon synthesis for embodied scenarios.

Confidential reviewer copy. This manuscr t s e o n e e f n e g (r 6 r d
sharing, redistribution, or disclosure is strictly prohibited.

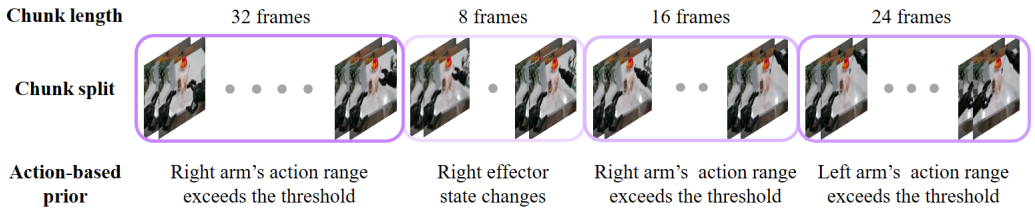


Figure 1: Action-guided chunk partitioning mechanism. This mechanism leverages embodied actions to partition the video, ensuring each chunk contains a distinct semantic unit. It divides the stream into short segments for active manipulation (e.g., significant movement or end-effector state changes) and longer segments for subtle locomotion dynamics.

3 Methodology

3.1 Autoregressive video generation with chunk-wise diffusion

Embodied world models broadly fall into two paradigms: action-driven models for online interactive control, and text-driven models for offline data synthesis. LongScape belongs to the latter, targeting stable long-horizon text-conditioned generation as a foundation for embodied data augmentation (validated in Section 4.4).

Inspired by the next-token-prediction autoregressive paradigm of LLMs, we approach video generation by treating a video as a sequence of temporal chunks, with each chunk (single or multiple frames) functioning as a "token". This enables a token-by-token rollout to generate videos, a causal paradigm that inherently preserves causality and helps enforce physical consistency. The full video can be represented as $\mathbf{V} = (\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_N)$, where $\mathbf{S}_t$ is the video chunk with a timestamp t. The objective is to predict the next chunk based on the preceding history:

$$p(\mathbf{V}) = p(\mathbf{S}_1) \prod_{t=1}^{N-1} p(\mathbf{S}_{t+1} | \mathbf{S}_1, \ldots, \mathbf{S}_t), \tag{1}$$

where $\mathbf{S}_t$ is the video chunk at time step t. The central task then becomes modeling the single-step generation process $p(\mathbf{S}_{t+1} | \mathbf{S}_1, \ldots, \mathbf{S}_t)$. Two primary paradigms exist to address this.

The first is to handle visual information similarly to language, using a tokenizer to discretize the image into patches and then extracting discrete representations. These are then arranged in spatiotemporal order to form the chunk's overall token: $\mathbf{X}_t$, where $\mathbf{X}_t = \text{Tokenizer}(\mathbf{S}_t)$. A transformer-based model is then trained on a next-token prediction task to predict the subsequent chunk token $\mathbf{X}_{t+1}$. While this approach is generally efficient, the discretization of continuous visual information can lead to a reduction in generated image quality.

The second paradigm is diffusion-based continuous generation, where the next video chunk, $\mathbf{S}_{t+1}$, is produced through a conditional denoising process guided by the previous chunk, $\mathbf{S}_t$. Operating within a continuous latent space (often from a VAE), the model starts with a noisy latent representation of the target chunk, $\mathbf{z}_{t+1,K}$. This representation is then iteratively refined over K steps. At each denoising step k, a model, typically a U-Net or a Diffusion Transformer (DiT), predicts and removes noise from the current noisy sample, $\mathbf{z}_{t+1,k}$. This operation is conditioned on the preceding chunk $\mathbf{S}_t$, following the process: $\mathbf{z}_{t+1,k-1} = \text{Denoise}(\mathbf{z}_{t+1,k}, \mathbf{S}_t)$. This refinement ultimately yields a clean latent representation, $\mathbf{z}_{t+1,0}$, which is then decoded into the final video chunk, $\mathbf{S}_{t+1}$. This approach preserves finer local visual details, mitigates error accumulation and drift during long-range autoregressive rollout, and benefits from the powerful generative capabilities of modern full-sequence diffusion models. Given these advantages, we adopt a hybrid framework that combines local chunk-based diffusion denoising with global autoregressive generation. In the following section, we elaborate on our design of adaptive chunk partitioning to enable more efficient long-horizon video generation within this framework.

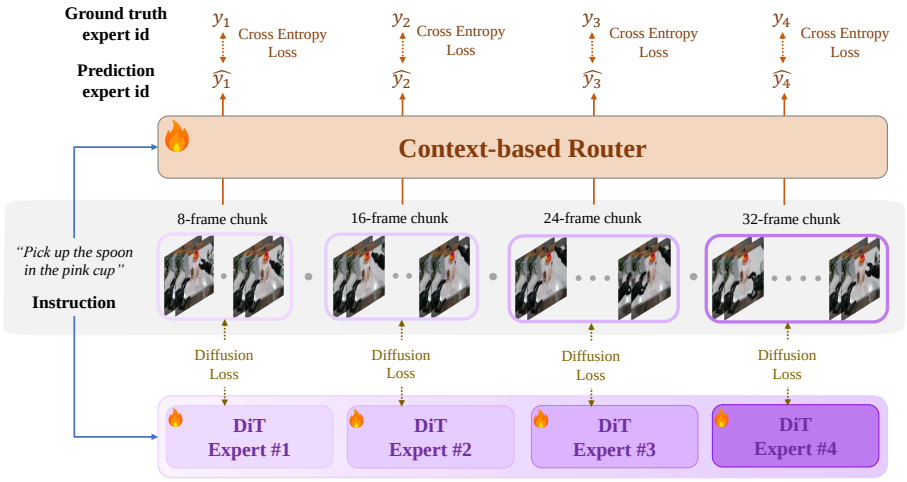


Figure 2: Training paradigm of LongScape. Different types of video chunks are used to train specialized DiT experts, while a context-based dynamic router learns to allocate the next chunk to the appropriate expert based on textual instruction and visual context.

3.2 Action-guided video chunk partition

Our design for the Context-aware Mixture-of-Experts (CMoE) in embodied video generation stems from the observation that different phases of a robotic task involve distinct dynamic patterns. For instance, long-range navigation toward an object requires the model to capture locomotion patterns, while precise manipulation demands an understanding of fine-grained dynamics. Accordingly, we propose to partition video sequences using action-state changes as a structural prior.

We focus on embodied manipulation scenarios, where robotic actions are naturally characterized by the gripper state and end-effector pose, where the gripper state indicates the interaction events, while the end-effector pose captures motion magnitude and direction. Concretely, we posit that robotic action labels provide a reliable signal for video chunk partitioning, based on two key factors: (1) changes in the gripper state reveal whether the robot is interacting with an object, and (2) motion magnitude can be quantified with end-effector pose changes. The video is first divided into non-overlapping base chunks of 8 frames. A final chunk is then formed by grouping 1 to 4 consecutive base chunks, resulting in variable-length chunks of 8, 16, 24, or 32 frames.

The partitioning algorithm processes the video sequentially from the initial frame. Starting with the first unassigned base chunk, it iteratively evaluates a candidate chunk consisting of the next n consecutive base chunks ($1 \leq n \leq 4$). The partitioning steps proceed as follows:

1. **Substantial Motion Detection:** If the range of motion in any dimension within the candidate chunk exceeds θ_d (θ_d is empirically set to half of the global motion amplitude), all n base chunks are grouped into a single chunk.
2. **Maximum Length Check:** If $n = 4$, the candidate chunk is immediately assigned as a chunk to avoid exceeding the maximum allowed length.
3. **Gripper State Change Detection:** If the gripper state changes in the last base chunk of the candidate chunk, the first $n - 1$ base chunks form one chunk, and the last base chunk is treated as a separate chunk.
4. **Continue Expansion:** If none of the above conditions are met, n is incremented by 1, and the process repeats—unless n already equals 4.

This approach ensures that video segments with significant motion or critical gripper state changes are assigned shorter chunks, while segments with less dynamic activity are assigned longer chunks. The overall process is formalized in Algorithm 1.

Table 1: Quantitative comparison of different video world models in terms of video visual quality on LIBERO and AGIBOT-World datasets.

Method	LIBERO						AGIBOT-World					
	PSNR↑	LPIPS↓	SSIM↑	FVD↓	Subject Consistency↑	Background Consistency↑	PSNR↑	LPIPS↓	SSIM↑	FVD↓	Subject Consistency↑	Background Consistency↑
CogVideoX	19.315	0.1402	0.7729	184.69	0.8974	0.9582	15.800	0.4028	0.6733	267.39	0.8224	0.8874
Genie	19.300	0.2192	0.7267	383.82	0.9098	0.9592	16.166	0.3620	0.6058	480.45	0.8743	0.9180
Cosmos-Predict 2.5	10.047	0.5005	0.1587	413.53	0.7068	0.8832	12.523	0.4212	0.5331	268.71	0.8597	0.9093
NOVA	13.007	0.4488	0.4262	346.91	0.9114	0.9580	16.043	0.4201	0.6998	273.60	0.8148	0.9018
MAGI-1	15.330	0.3520	0.4721	541.39	0.8847	0.9353	14.448	0.4849	0.5961	490.53	0.9092	0.9265
CausVid	19.915	0.1383	0.7860	276.25	0.9160	0.9566	15.873	0.4934	0.6920	358.33	0.6965	0.8378
Genie-Envisioner	18.204	0.2832	0.6848	417.55	0.9175	0.9569	14.468	0.5193	0.6323	495.93	0.9054	0.9192
Ours	**19.977**	**0.1231**	**0.7883**	**153.72**	**0.9297**	**0.9622**	**16.493**	**0.3613**	**0.7015**	**256.16**	**0.9207**	**0.9341**

3.3 Context-aware MoE for adaptive rollout

Following the partitioning of video into chunks of varying lengths, we introduce a **Context-aware Mixture-of-Experts (CMoE)** framework to reconcile specialized DiT networks for denoising each chunk type. This approach mitigates conflicts and forgetting issues that arise from training on diverse motion patterns, e.g., large-scale locomotion versus fine-grained object manipulation. Our model employs a set of K DiT experts, denoted as $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_K\}$. Each expert is initialized from the parameters of a pre-trained CogVideoX [19]. In our implementation, we set $K = 4$, corresponding to four distinct chunk types. Each expert network is composed of a stack of transformer blocks, featuring 3D full attention layers and feed-forward networks connected by Adaptive Layer Normalization (AdaLN) layers. We train each expert $\mathcal{E}_i$ to perform a denoising task on its corresponding chunk type, optimizing the objective function:

$$\mathcal{L}_i = \mathbb{E}_{\mathbf{z}_0, \epsilon \sim \mathcal{N}(0,\mathbf{I}), \tau \sim [1,T]} \left[\left\| \epsilon - \mathcal{E}_i(\mathbf{z}_\tau, \tau, \mathbf{c}, \mathbf{S}_t) \right\|^2 \right], \tag{2}$$

where $\mathbf{z}_0$ is the clean latent representation of the target chunk $\mathbf{S}_{t+1}$, ϵ is the added noise, $\mathbf{z}_\tau$ is the noisy latent at diffusion time step τ, T is the total number of diffusion steps, $\mathbf{c}$ is the global text condition, and $\mathbf{S}_t$ is the preceding video chunk. The expert model $\mathcal{E}_i$ predicts the noise ϵ to be removed, conditioned on the text prompt and the preceding visual context.

To enable adaptive inference-time expert selection, we introduce a **dynamic contextual router** $\mathcal{R}$, which is implemented by a single cross-attention transformer network, taking the global text instruction $\mathbf{c}$ and visual features from the current video chunk $\mathbf{S}_t$ as input, predicting which expert to activate for the next chunk $\mathbf{S}_{t+1}$. The router is trained using a cross-entropy loss:

$$\mathcal{L}_{\text{router}} = \mathbb{E}_{(\mathbf{c}, \mathbf{S}_t, i)} \left[-\log p(i|\mathbf{c}, \mathbf{S}_t) \right], \tag{3}$$

where i is the ground-truth expert index. The training process of experts and the router is presented in Figure 2 and this process is highly efficient, requiring only a small subset of partitioned video chunks and their corresponding chunk length labels.

The inference pipeline is illustrated in Figure 3. At each rollout step, the global text prompt and visual features from the last frame of the current chunk are fed into the router to predict the appropriate expert:

$$i^* = \underset{i \in \{1,\ldots,K\}}{\arg\max} \left(\mathcal{R}(\mathbf{c}, \mathbf{S}_t)_i \right). \tag{4}$$

The activated expert $\mathcal{E}_{i^*}$ then generates the next chunk of the corresponding length $\mathbf{S}_{t+1}$:

$$\mathbf{S}_{t+1} = \mathcal{E}_{i^*}(\mathbf{S}_t, \mathbf{c}). \tag{5}$$

Although the current design focuses on robotic manipulation scenarios, the proposed CMoE framework is generally applicable to long-horizon video generation with structured dynamics, and can be extended to other action-annotated domains such as mobile navigation and human motion. We leave large-scale validation on general long-form video benchmarks to future work.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on LIBERO [16] and AGIBOT-World [17], both focusing on long-horizon composite embodied manipulation tasks (e.g., folding clothes). Detailed descriptions are provided in Appendix A.2.

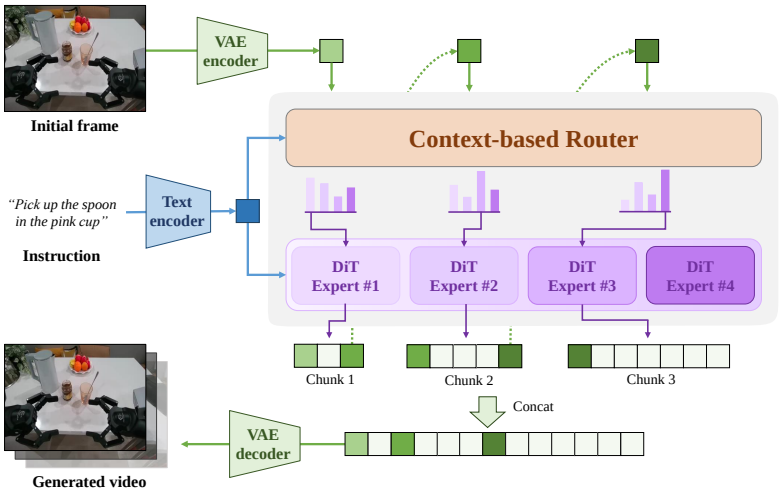


Figure 3: Inference pipeline of LongScape. At each autoregressive step, the dynamic router leverages the global text instruction and the visual features of the current chunk to select the appropriate DiT expert. The selected expert then generates the subsequent video chunk via diffusion denoising, with this iterative process continuing to produce a long video sequence.

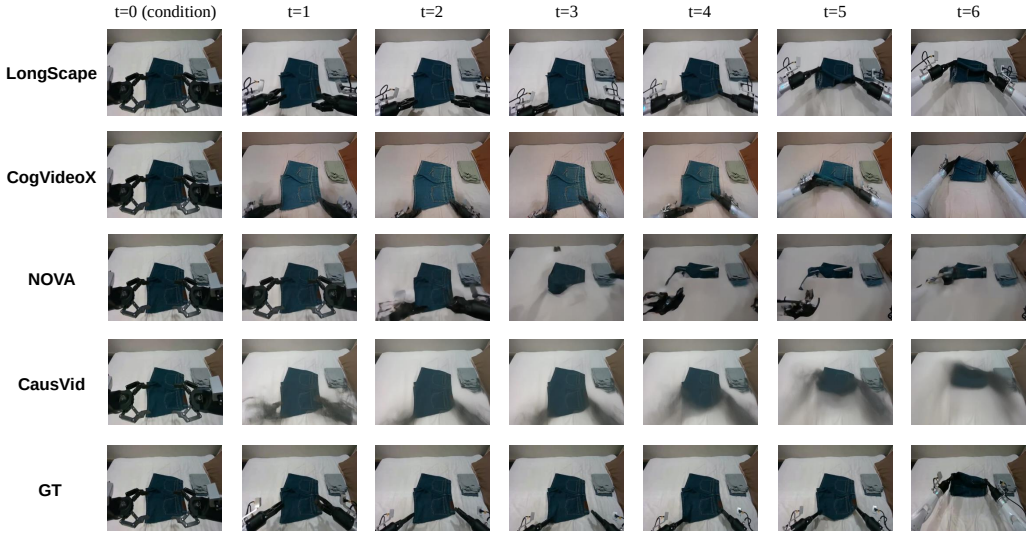


Figure 4: Qualitative comparison on AGIBOT-World dataset with text instruction "grasp the lower pant leg and waistband of denim shorts and fold them over the upper pant leg and waistband". Here we select several models with good results and sample six frames from the generated long video sequences for better presentation.

Baselines. We compare LongScape against three categories of state-of-the-art models:

- **Foundational Video Generation Models:** CogVideoX [19], Genie [9], Cosmos-Predict 2.5 [7], and NOVA [11].
- **Long-Horizon Video Generation Models:** MAGI-1 [12] and CausVid [13].
- **Embodied World Models:** Genie-Envisioner [2].

This selection covers both general video generation architectures and specialized long-horizon methods. Further model details are in Appendix A.3.

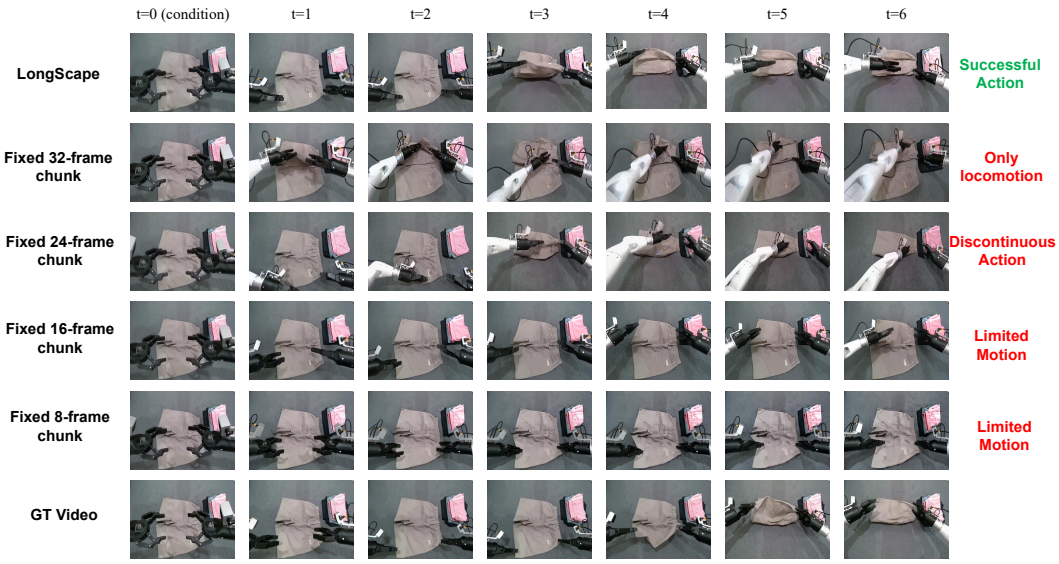


Figure 5: Visualization of the ablation study on AGIBOT-World dataset. We compare LongScape with variants using only fixed-length chunks. These fixed-length approaches often exhibit discontinuous or physically implausible actions. In contrast, LongScape, by incorporating an adaptive expert mechanism, ensures action coherence and stable generation.

Table 2: Ablation study of using various fixed-length chunk generation and the MoE designs.

Method	LIBERO						AGIBOT-World					
	PSNR↑	LPIPS↓	SSIM↑	FVD↓	Subject Consistency↑	Background Consistency↑	PSNR↑	LPIPS↓	SSIM↑	FVD↓	Subject Consistency↑	Background Consistency↑
LongScape	**19.977**	**0.1231**	**0.7883**	**153.72**	**0.9297**	**0.9622**	**16.493**	**0.3613**	**0.7015**	**256.16**	**0.9207**	**0.9341**
32-frame chunk	19.755	0.1285	0.7845	180.31	0.9091	0.9610	15.641	0.4286	0.6702	282.38	0.8535	0.9111
24-frame chunk	19.620	0.1314	0.7798	191.02	0.9104	0.9620	15.903	0.4034	0.6814	303.75	0.8952	0.9328
16-frame chunk	19.640	0.1388	0.7783	176.55	0.9129	0.9614	15.894	0.3852	0.6917	373.50	0.9090	0.9328
8-frame chunk	18.893	0.1629	0.7518	229.71	0.8999	0.9599	15.578	0.4052	0.6669	387.49	0.8789	0.9178
Unified Expert	12.895	0.4277	0.4387	365.35	0.9028	0.9583	–	–	–	–	–	–

Metrics. We evaluate using PSNR, SSIM, LPIPS, FVD, subject consistency, and background consistency (see Appendix A.4).

Implementation. Each DiT expert has 5.57B parameters; the router has 108.57M (AGIBOT-World) and 71.35M (LIBERO) parameters. All models are trained on 30,000 clips sampled at 10 Hz. Full training configurations and resource requirements are detailed in Appendix A.5.

4.2 Main results

Quantitative comparison. Quantitative evaluations in Table 1 demonstrate that our model achieves superior performance across all four metrics on both the LIBERO and AGIBOT-World datasets. These benchmarks focus on **long-horizon composite embodied manipulation tasks** (each task lasting approximately 20 seconds), which generally require multiple rollout steps to generate the full video sequence. Compared with the most competitive baseline, LongScape achieves an average 8.6% improvement on LIBERO dataset and 5.8% improvement on AGIBOT-World dataset. For diffusion-based models such as CogVideoX, we generate the entire long frame sequence in a single forward pass. For autoregressive and hybrid models, long sequences were produced through iterative rollouts. However, these methods have significant drawbacks. Diffusion models often generate physically implausible content due to their parallel frame generation mechanism, while the autoregressive models suffer from error accumulation, where early inaccuracies propagate and amplify over time. These limitations hinder both baseline types from producing coherent long-range videos. Our model's architecture avoids these issues by flexibly generating chunks with full semantic integrity. This approach ensures high-quality generation within each chunk and maintains strong causal relationships between them, making our model exceptionally well-suited for long-horizon generation tasks.

Table 3: Per-frame generation time (in seconds) on LIBERO dataset.

Method	32-frame chunk	24-frame chunk	16-frame chunk	8-frame chunk	LongScape
Time/Frame (s)	0.5598	0.6189	0.7323	1.0912	0.7974

Table 4: Quantitative comparison of different models on out-of-domain tasks on AGIBOT-World dataset.

Method	PSNR↑	LPIPS↓	SSIM↑	FVD↓	Subject Consistency↑	Background Consistency↑
CogVideoX	15.528	0.3877	0.6465	301.80	0.8649	0.8916
Genie	11.863	0.6234	0.2716	385.78	0.8186	0.9028
NOVA	13.167	0.5406	0.4601	435.50	0.7816	0.8895
CausVid	9.833	0.6754	0.2376	470.38	0.7585	0.8449
LongScape	**15.936**	**0.3717**	**0.6585**	**297.16**	**0.9172**	**0.9167**

Qualitative results. To provide a more intuitive comparison of our method's effectiveness against the baselines, we offer a set of visualizations of the generated results. Figure 4 presents a challenging cloth-folding task from the AGIBOT-World dataset. This task requires a high level of spatiotemporal coherence and a strong understanding of physical laws. Here, CogVideoX produces blurry artifacts. NOVA suffers from objects disappearing from the scene. By comparsion, our model consistently maintains a plausible deformation of the cloth and generates a continuous, coherent motion sequence even with multiple rollouts. Figure 6 shows the generation of a complex manipulation task from the LIBERO dataset. This task involves multiple steps, such as picking and placing objects and closing a drawer.

As shown in the visualization, CogVideoX suffers from ghosting artifacts in later frames. CausVid again blurs subject. NOVA misinterprets the instruction, picking up the wrong object, while Genie fails in the very first pickup action. In contrast, our model maintains coherent actions across multiple rollouts, producing a logical and successfully executed task sequence.

4.3 Ablation Study

Effectiveness of the context-based MoE. Effectiveness of the context-based MoE. To validate our MoE design, we compare LongScape against four variants each using a single expert trained on fixed-length chunks (8, 16, 24, or 32 frames). As shown in Table 2, all fixed-length variants underperform our CMoE on both datasets. Longer chunks (e.g., 32 frames) perform slightly better than shorter ones (e.g., 8 frames), which we attribute to very short chunks failing to capture semantically complete actions.

To disentangle variable-length chunking from additional model capacity, we further compare against a **Unified Expert** baseline: a single DiT trained on mixed chunks of all four lengths. The Unified Expert achieves substantially worse performance (e.g., 12.895 vs. 19.977 PSNR on LIBERO) and even underperforms the best fixed-length expert, confirming that a single model cannot effectively learn all temporal granularities simultaneously. Multi-expert specialization is therefore necessary to realize the benefits of variable-length chunking.

Qualitative results (Figure 5 and Figure 7) further illustrate that fixed-length designs suffer from mode collapse: long-chunk models produce only coarse locomotion, while short-chunk models exhibit poor motion dynamics. Our CMoE flexibly handles both patterns, producing consistently coherent generation.

Effectiveness of the dynamic router. We evaluate the router as a four-class classification task on 12,700 LIBERO test samples, achieving 91.4% accuracy. We also examine the effect of routing errors during autoregressive rollout: when an incorrect expert is selected, the transition remains visually plausible due to high overlap between adjacent contexts, and such errors do not accumulate—the visual features of the newly generated frame naturally steer the router back within a few steps, preserving long-range stability.

Table 5: Downstream policy success rates (%) on LIBERO_10 when training $\pi 0.5$ with varying amounts of LongScape-generated data. Short video augmentation (CogVideoX, 5 s) is shown for comparison.

Real Data	Augmented Data	Success Rate (%)
500	0 (baseline)	58.4
500	500 long (LongScape)	62.8
500	1000 long (LongScape)	66.0
500	1500 long (LongScape)	68.2
500	1500 short (CogVideoX)	44.6

We further verify the robustness of our action-guided chunking by conducting a sensitivity analysis on the motion threshold θ_d. Across $\theta_d \in [0.3, 0.7]$, LongScape consistently outperforms the best fixed-length expert (details in Appendix A.6). We use $\theta_d = 0.5$ throughout.

Throughput analysis of the MoE architecture. Table 3 reports per-frame generation time on LIBERO. Fixed-chunk models become slower as chunk length decreases; LongScape achieves 0.7974s per frame, an intermediate value balancing fine-grained detail and parallel efficiency. The multi-expert design incurs negligible overhead: training uses data-parallel fine-tuning (total compute equals one full fine-tune), inference activates only one expert at a time, and storing four experts requires approximately 42 GB in half precision—feasible on a single NVIDIA RTX 4090D GPU.

Out-of-Domain Generalization on Unseen Tasks. We evaluate generalization by training exclusively on "cloth folding" and testing on "pick-and-place" in an unseen supermarket environment. As shown in Table 4, LongScape outperforms all baselines across all metrics, confirming its ability to learn task-agnostic dynamics.

4.4 Generated Data for Downstream Policy Learning

To verify that LongScape's long-horizon generation translates to practical embodied learning benefits, we evaluate its synthetic data in downstream policy training. Using $\pi 0.5$ on LIBERO_10, we augment a baseline set of 500 real demonstrations with LongScape-generated trajectories of varying sizes. We also compare with an equal volume of short clips (approx. 5 s) from CogVideoX, representing typical short-horizon generation.

As shown in Table 5, LongScape-generated long videos consistently improve policy success rate, reaching a 9.8% absolute gain over the real-data-only baseline. Conversely, short video augmentation degrades performance to 44.6%, likely because truncated clips fail to capture the full temporal dynamics of real manipulation tasks. These results confirm that LongScape's stable long-horizon generation directly benefits embodied policy learning, rather than serving merely as a video quality benchmark.

5 Conclusion and Future Works

In this work, we introduce LongScape, a novel framework for long-horizon embodied world modeling that adaptively integrates diffusion denoising with autoregressive generation. Our method leverages action priors to guide video chunk partitioning and employs a dynamic router to activate specialized generation experts. This approach enables the production of video tailored to the semantic context of each chunk, significantly enhancing visual quality and coherence. Experimental results demonstrate that LongScape outperforms existing approaches by maintaining stable and coherent video generation over extended horizons.

Despite these advancements, several directions merit further exploration. First, while the current action-guided chunking mechanism relies on heuristic rules, future work could investigate learning-based chunking strategies to achieve more flexible and task-aware video partitioning. Second, extending LongScape to support multi-view camera inputs would further enhance its applicability in complex real-world scenarios.

References

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

[2] Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

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

[4] Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. Roboscape: Physics-informed embodied world model. *arXiv preprint arXiv:2506.23135*, 2025.

[5] Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, et al. Enerverse-ac: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*, 2025.

[6] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through neural trajectories. *arXiv e-prints*, pages arXiv–2505, 2025.

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

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

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

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

[11] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024.

[12] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

[13] Shaobin Zhuang, Zhipeng Huang, Ying Zhang, Fangyikang Wang, Canmiao Fu, Binxin Yang, Chong Sun, Chen Li, and Yali Wang. Video-gpt via next clip diffusion. *arXiv preprint arXiv:2505.12489*, 2025.

[14] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

[15] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[16] 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.

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

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

[19] 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.

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

[21] 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.

[22] Yanhui Wang, Jianmin Bao, Wenming Weng, Ruoyu Feng, Dacheng Yin, Tao Yang, Jingxu Zhang, Qi Dai, Zhiyuan Zhao, Chunyu Wang, Kai Qiu, Yuhui Yuan, Xiaoyan Sun, Chong Luo, and Baining Guo. Microcinema: A divide-and-conquer approach for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8414–8424, June 2024.

[23] Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*, 2025.

[24] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Ming Gong, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. NUWA-XL: Diffusion over diffusion for eXtremely long video generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1309–1320, Toronto, Canada, July 2023. Association for Computational Linguistics.

[25] Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, and Jun Xiao. Vid-gpt: Introducing gpt-style autoregressive generation in video diffusion models. *arXiv preprint arXiv:2406.10981*, 2024.

[26] 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*, 2024.

[27] Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Zhengkai Jiang, Yue Hu, Yue Liao, Peng Gao, Hongsheng Li, Maoqing Yao, et al. Enerverse: Envisioning embodied future space for robotics manipulation. *arXiv preprint arXiv:2501.01895*, 2025.

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

[29] Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: Dino as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025.

[30] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

A Appendix

A.1 Reproducibility statement

To ensure reproducibility, we have made our code and some visualization cases available in the following anonymous repository for review: `https://anonymous.4open.science/r/AMSVVD-fdg245`..

A.2 Details of Datasets

We conduct experiments on LIBERO and AGIBOT-World datasets, both focusing on long-horizon composite embodied manipulation tasks.

- **LIBERO** [16]: A benchmark for lifelong robot learning, including 130 diverse manipulation tasks organized into specialized suites (LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-100) to study knowledge transfer across variations in object spatial relationships, object types, and task goals. In our experiments, we performed training and evaluation on each of the 130 tasks individually.
- **AGIBOT-World** [17]: A large-scale real-world dataset collected with a wheeled dual-arm robot. It covers over 100 real-world scenarios across five core environments (household, catering, industrial, commercial, office) and contains a high proportion of long-horizon tasks. Our experiments utilize a subset focusing on two representative long-horizon task categories: folding clothes and pick-and-place operations.

A.3 Details of Baselines

We compare our approach with seven advanced video generation world models, covering purely diffusion-based, autoregressive, and combined diffusion-autoregressive architectures, which are detailed as follows.

- **CogVideoX** [19]: An advanced open-source diffusion-based text-to-video model. Its core innovation is a DiT-based expert transformer that employs 3D full attention to effectively model complex spatiotemporal relationships, achieving high visual fidelity. We use the 5B parameter version for a fair comparison.
- **Genie** [9]: A large-scale autoregressive foundation world model. It operates by encoding video frames into discrete tokens and employs a spatial-temporal transformer for next-token prediction. As a seminal work, it demonstrates the potential of scaling autoregressive paradigms for diverse video generation tasks.
- **Cosmos-Predict 2.5** [7]: An autoregressive world model for physical scene reasoning. Built upon a Llama 3-style architecture, it is specifically designed to predict the future states of dynamic scenes, making it a strong baseline for tasks requiring physical consistency and long-horizon forecasting.
- **NOVA** [11]: A hybrid diffusion-autoregressive model designed for streaming video generation. It integrates per-frame diffusion denoising within an autoregressive framework, ensuring both high intra-frame detail and robust inter-frame temporal coherence across long sequences.
- **MAGI-1** [12]: A state-of-the-art hybrid model for long video generation. It performs fixed-length diffusion denoising within an autoregressive chain, effectively balancing the visual quality advantages of diffusion models with the long-range consistency of autoregressive approaches.
- **CausVid** [13]: A video generation model that emphasizes causal consistency. It employs a teacher-student distillation framework, which transfers the knowledge from a pre-trained diffusion model to a fast autoregressive generator for efficient streaming generation.
- **Genie-Envisioner** [2]: A unified world foundation platform specifically designed for embodied scenarios. It facilitates realistic video prediction for robotic manipulation by supporting multi-view camera inputs and action conditioning, serving as a strong baseline in the domain of embodied AI.

A.4 Details of Metrics

We evaluate the generated videos using the following metrics:

- **PSNR**: It evaluates the pixel-level similarity between generated and ground-truth frames.
- **LPIPS**: It measures the image feature similarity of generated and ground truth frames.
- **SSIM**: It assesses the brightness, contrast, and structural consistency between generated and ground truth frames.
- **FVD**: It measures the distance between real and generated video feature distributions.
- **Subject Consistency**: It measures the perceptual stability of the main subject across the video frames, following VBench [30] protocol..
- **Background Consistency**: It assesses the temporal coherence and invariance of the background scene throughout the video, following VBench [30] protocol.

A.5 Details of Implementation

For both datasets, each DiT expert contains 5.57B parameters, initialized from the CogVideoX-5B checkpoint [19]. The dynamic router has 108.57M parameters for AGIBOT-World and 71.35M parameters for LIBERO. We sample videos at 10 Hz from LIBERO and AGIBOT-World, yielding approximately 30,000 training clips from each source. Each expert is fine-tuned independently on its designated chunk type for 2 epochs using 4 NVIDIA H20 GPUs, requiring around 24 hours per expert. The router is trained for a similar duration (24 hours) on 4 NVIDIA H20 GPUs. Inference throughput measurements are conducted on a single NVIDIA GeForce RTX 4090D GPU.

Algorithm 1 Algorithm of chunk partition

Input: Video frames F, action sequence A for each frame
Output: Partitioned chunks set $S = \{S_1, S_2, \ldots, S_k\}$

Divide video into base chunks C of 8 frames each
Calculate action thresholds θ_d for each dimension $d \in \{x, y, z, \text{pitch}, \text{yaw}, \text{roll}\}$
while not all base chunks processed **do**:
    Check candidate chunk $C_{1:n}$ of length n (1-4 base chunks):
    if $\Delta A_{d,1:n} > \theta_d$ **then**: add $C_{1:n}$ to S
    else if n=4 **then**: add $C_{1:n}$ to S
    else if gripper state changes in C_n **then**: add $C_{1:n-1}$ and C_n to S
    else: increase n and continue checking

A.6 Sensitivity Analysis of Motion Threshold θ_d

We evaluated $\theta_d \in \{0.3, 0.5, 0.7\}$ on the complete LongScape pipeline. All three settings outperform the best single fixed-length expert, with $\theta_d = 0.5$ yielding the optimal trade-off.

Table 6: Performance across different θ_d values on LIBERO.

θ_d	PSNR↑	LPIPS↓	SSIM↑	FVD↓
0.3	19.865	0.1246	0.7883	158.93
0.5	**19.977**	**0.1231**	**0.7883**	**153.72**
0.7	19.923	0.1243	0.7856	172.74
Best single DiT	19.755	0.1285	0.7845	180.31

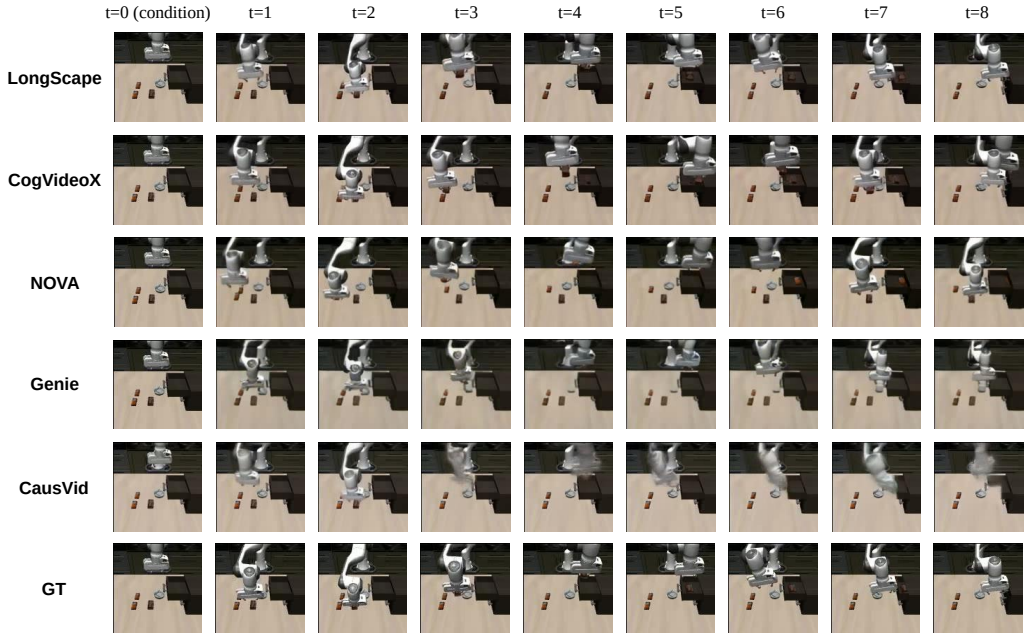


Figure 6: Qualitative comparison on LIBERO dataset with text instruction "put the chocolate pudding in the top drawer of the cabinet and close it". Here we select several models with good results and sample eight frames from the generated long video sequences for better presentation.

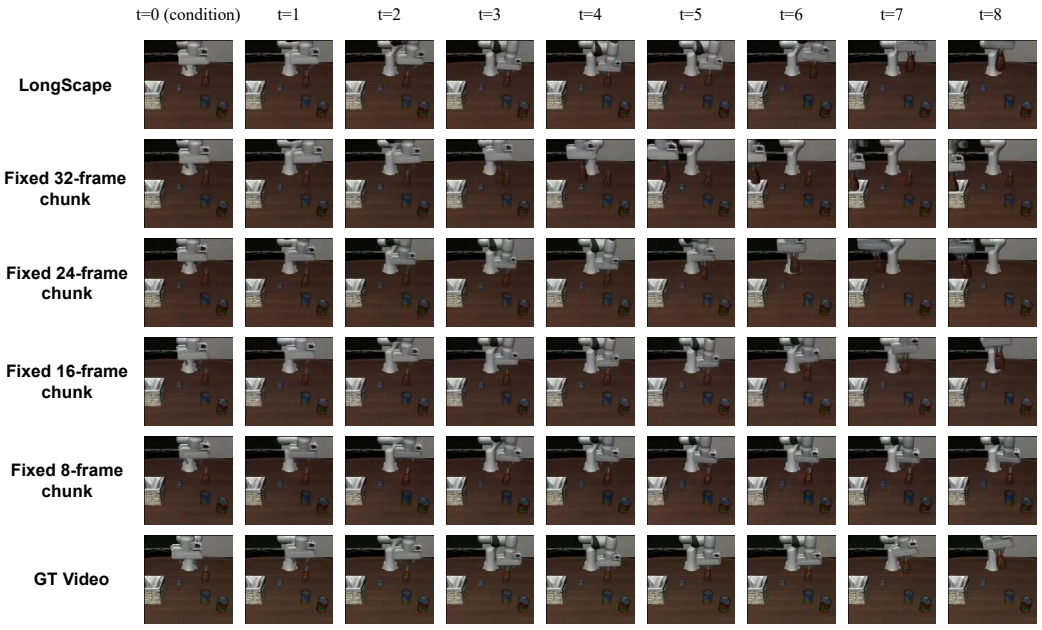


Figure 7: Visualization of the ablation study on LIBERO dataset. We compare LongScape with variants using only fixed-length chunks. These fixed-length approaches often exhibit discontinuous or physically implausible actions. In contrast, LongScape, by incorporating an adaptive expert mechanism, ensures action coherence and stable generation.

B Limitations

Despite the strong performance of LongScape, several limitations merit discussion.

Heuristic chunking. Our action-guided chunk partitioning relies on manually defined rules (gripper state changes and motion thresholds). While we have verified its robustness across a wide range of θ_d values (see Appendix A.6), the heuristic nature may limit applicability to domains where such structured action labels are unavailable, such as in-the-wild human demonstration videos. Extending LongScape with learned or vision-based chunking strategies is a promising direction for future work.

Text-driven paradigm. LongScape is designed as a text-conditioned generative world model for offline data synthesis, rather than an action-conditioned interactive world model. It does not natively support real-time closed-loop control or online correction during rollout. Extending the architecture to incorporate explicit action conditioning is feasible but has not been validated in this work.

Domain scope. Our experiments focus on robotic manipulation tasks (LIBERO and AGIBOT-World). While we demonstrate zero-shot transfer between manipulation task categories, performance on other embodied domains (e.g., navigation, quadruped locomotion) remains unexplored.

Computational storage. Although inference activates only one expert at a time, storing all four DiT experts requires approximately 42 GB in half precision. This is manageable for server-grade hardware but may be burdensome for resource-constrained deployments.

We believe these limitations represent meaningful directions for future investigation rather than fundamental flaws, and we openly discuss them to guide subsequent research.

C Broader Impacts

Positive societal impacts. LongScape advances the capability of embodied world models to generate long, coherent video trajectories from text instructions. This can lower the cost and improve the safety of training robotic agents by providing high-quality synthetic data for simulation-based learning, potentially reducing the need for expensive and potentially hazardous real-world data collection. Applications in assistive robotics, manufacturing automation, and accessible robot learning tools could benefit from this line of research.

Negative societal impacts and risks. As with any technology that generates realistic visual sequences, there are potential risks. First, LongScape-generated videos, while visually coherent, may contain physically implausible or unsafe actions that are not immediately apparent. If used directly for policy training without sufficient real-world validation, this could lead to policy failures when deployed on physical robots. Second, the ability to synthesize plausible manipulation sequences could be misused to create misleading demonstrations of robotic capabilities, e.g., generating videos that falsely suggest a robot can perform a dangerous or unethical task. Third, over-reliance on synthetic data could reduce incentives for collecting real-world safety-critical data.

Mitigation strategies. We encourage the following safeguards: (i) all policies trained on LongScape-generated data should undergo rigorous real-world validation before deployment; (ii) generated videos should be clearly labeled as synthetic when shared; (iii) evaluation protocols for physical plausibility and safety should complement standard visual quality metrics. We will release our model checkpoint with a responsible use guideline that outlines these considerations.

Intended use. LongScape is a research prototype intended for academic investigation of world modeling and simulation-based learning. It is not designed or validated for direct real-world deployment without comprehensive safety testing.

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 abstract and introduction clearly state our contribution of LongScape, an action-guided variable-length chunking framework with context-aware MoE for long-horizon embodied video generation, and all claims are supported by the experimental results.

   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 dedicated "Limitations" section iscusses the heuristic chunking rules, text-driven paradigm, domain scope, and storage cost. We also outline directions for future mitigation. See 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: The paper does not include 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: We provide anonymized code (link in abstract and Appendix A.1), detailed model architectures, training configurations, and hyperparameters in Section 4.1 and Appendix A.5. The datasets used are publicly available.

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: Justification: An anonymous code repository is provided, containing scripts for data preprocessing, training, and evaluation. The datasets (LIBERO, AGIBOT-World) are publicly available.

   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: Training details (sampling rate, number of clips, expert/routing training epochs, GPU type, batch size, learning rate schedules) are described in Section 4.1 and Appendix A.5.

   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: [No]

   Justification: Error bars are not reported. Due to the high computational cost of training multiple DiT experts and performing long-horizon rollouts, each experiment was run once with a fixed random seed. We do not claim statistical significance based on multiple runs, and our main conclusions are drawn from consistent improvements across two datasets and multiple metrics.

   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: Section 4.1 and Appendix A.5 specify the GPU models (NVIDIA H20, RTX 4090D), training time (approx. 24h per expert), and inference throughput (Table 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: Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research adheres to it. No human subjects were involved, and we discuss potential societal impacts in the Impact Statement.

   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: The "Broader Impacts" section discusses positive impacts (e.g., safer robotic simulation, data augmentation) and negative risks (e.g., physical implausibility, misleading demos, reduced real-world data collection), along with mitigation strategies and intended use. See C.

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: We do not release a new pre-trained model or scraped dataset with high misuse risk. Our model is a fine-tuned version of an existing open-source video model for research purposes only.

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: All datasets (LIBERO, AGIBOT-World) and baseline models (CogVideoX, etc.) are properly cited. We have respected their licenses for academic research.

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: We provide an anonymized code repository with documentation on how to run the experiments. The code includes the data preprocessing, training, and evaluation pipelines.

    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: The paper does not involve crowdsourcing or research with human subjects.

    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: The paper does not involve research with human subjects.

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: LLMs were used solely for writing assistance and do not form any part of the core method or scientific contribution.

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.

Confidential reviewer copy. This manuscr t s e o n e e f n e g (r 6 r d
sharing, redistribution, or disclosure is strictly prohibited.