



EnerVerse: Envisioning Embodied Future Space for Robotics Manipulation

Siyuan Huang^{2,4*}, Liliang Chen^{1,*†}, Pengfei Zhou^{1,5}, Shengcong Chen¹, Zhengkai Jiang⁶, Yue Hu^{1,7}, Yue Liao³, Peng Gao², Hongsheng Li^{2,3}, Maoqing Yao^{1,‡}, Guanghui Ren^{1,‡}

¹AgiBot, ²Shanghai AI Lab, ³CUHK, ⁴SJTU, ⁵FDU, ⁶HKUST, ⁷HIT,

* Indicates equal contribution, †indicates project leader and ‡indicates corresponding author.

We introduce ENERVERSE, a generative robotics foundation model that constructs and interprets embodied spaces. ENERVERSE employs an autoregressive video diffusion framework to predict future embodied spaces from instructions, enhanced by a sparse context memory for long-term reasoning. To model the 3D robotics world, we propose Free Anchor Views (FAVs), a multi-view video representation offering flexible, task-adaptive perspectives to address challenges like motion ambiguity and environmental constraints. Additionally, we present ENERVERSE-D, a data engine pipeline combining the generative model with 4D Gaussian Splatting, forming a self-reinforcing data loop to reduce the sim-to-real gap. Leveraging these innovations, ENERVERSE translates 4D world representations into physical actions via a policy head (ENERVERSE-A), enabling robots to execute task instructions. ENERVERSE-A achieves state-of-the-art performance in both simulation and real-world settings

Date: January 01, 2025

Website: <https://sites.google.com/view/enerverse>

1 Introduction

Creative AI in vision has achieved significant progress, especially in video generation, where models produce high-quality videos from human instructions Kong et al. (2024); Zheng et al. (2024). This success highlights the model’s spatiotemporal imagination, enabling accurate forecasting of future frames. Similarly, robotic manipulation, a fundamental task in embodied AI, needs accurate prediction of future actions based on language instructions to interact with the physical world. Based on this sharing principle of future space prediction, a natural strategy is to align robotics action prediction with a video generation task to leverage video generation models’ imagination capabilities for policy planning. Motivated by this, recent studies Wen et al. (2024); Rigter et al. (2024); Cheang et al. (2024); Guo et al. (2024) have conducted preliminary explorations by fine-tuning general video generation models on robotic manipulation videos to align feature representations with the robotics domain, and predict physical actions. However, such methods Rigter et al. (2024) often simply adapt general-purpose video generation models to embodied tasks, neglecting the substantial gap between their representation space and the three-dimensional, temporally interconnected robotics environment, thereby hindering accurate action policy prediction.

To bridge the gap, we propose ENERVERSE, a generative robotics foundation model designed to construct and interpret the robotics world. In ENERVERSE, we employ an autoregressive video diffusion framework that iteratively predicts the embodied future space based on a given instruction. Within this generative paradigm, we define a minimal unit of the future space as a ‘chunk’, and the model repeatedly predicts the next chunk to incrementally expand the space. Additionally, to prevent model collapse and enhance the action planning capabilities, we design a sparse context memory mechanism during training. Instead of relying on

consecutive memory, this mechanism preserves essential prior content throughout the generation process in a non-redundant manner, theoretically allowing infinite-length sequence generation. This autoregressive strategy and the sparse memory design enable stable 2D embodied video generation, yet bridging the approach to 3D robotics remains challenging.

To address this, we introduce the novel concept of Free Anchor View (FAV) in ENERVERSE, by simultaneously predicting multi-view robotics videos to represent the 3D robotics world. Unlike hardware-mounted cameras that provide fixed perspectives, FAVs are not physically affixed but freely positioned by the users, offering flexible, task-adaptive multi-view perspectives as shown in Fig. 1. This flexibility is particularly beneficial for robotic tasks, as it mitigates challenges such as: (1) motion modeling ambiguity caused by extrinsic changes in body-mounted cameras during whole-body motions; (2) physical constraints in confined spaces, like collisions in narrow kitchens; and (3) the limited generalization and adaptability of static workspace-mounted cameras. Multi-FAV setups provide richer visual information, constructing implicit 3D spatial representations that significantly improve policy performance.

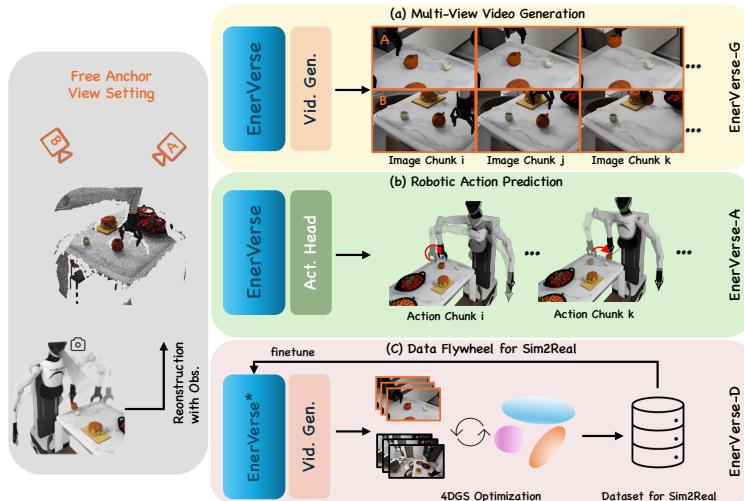


Figure 1 An overview of ENERVERSE. Leveraging camera observations, we first obtain a 3D reconstruction via depth warping and rendering Lassner and Zollhofer (2021), then setup FAVs. ENERVERSE that (a) connects to a video generator head (ENERVERSE-G) to produce multi-view videos, (b) attaches to a robotic action policy head (ENERVERSE-A) for action prediction, and (c) integrates with 4DGS to form a data flywheel (ENERVERSE-D) for Sim2Real adaptation.

However, acquiring precisely calibrated multi-camera observations, along with robotic actions, is costly and labor-intensive, limiting the ENERVERSE’s access to large-scale real-world training data. Simulators can generate abundant synthetic data, but the sim-to-real gap remains a significant challenge. To overcome this, we propose ENERVERSE-D, a data engine that combines a generative model with 4D Gaussian Splatting (4DGS). By leveraging the generative model’s adaptability and 4DGS’s spatial constraints, ENERVERSE-D initiates a self-reinforcing data flywheel, narrowing the sim-to-real gap and enhancing ENERVERSE’s performance.

Building on these designs, ENERVERSE effectively models and interprets the robotic environment in both 3D space and temporal dimensions. With this generative prior, we can directly translate the 4D world (3D spatial with temporal information) representation into physical actions via a policy head, as shown in Fig. 1, allowing the robot to execute task instructions in real-world scenarios. As a result, ENERVERSE-A attains state-of-the-art performance in both simulation and real-world deployments.

The contributions of this work are as follows: (1) We develop an innovative chunk-wise autoregressive diffusion model architecture capable of logically reasoning for embodied future space, benefiting from a sparse contextual memory mechanism. (2) We propose a novel FAV-based embodied future space method associated with policy planning, significantly enhancing spatial understanding. (3) We construct a data flywheel in the robotics domain that integrates 4DGS optimization into multi-view video generation, enabling iterative improvements in data quality.

2 Related work

Video Generation Models. Diffusion-based video generation models have made notable progress, especially in text-to-video (T2V) generation Blattmann et al. (2023); Song et al. (2020). Early T2V approaches Zhang et al. (2023); Chen et al. (2023); Ren et al. (2024); Guo et al. (2023) build on text-to-image (T2I) priors by introducing temporal modules trained on video data. DynamicCrafter Xing et al. (2025) reuses motion priors from T2V diffusion models in an image-to-video (I2V) context. Recent works Kong et al. (2024); Zheng et al. (2024); Bao et al. (2024) explores replacing U-Nets with Diffusion Transformer (DiT) Peebles and Xie (2023a). Other studies Gao et al. (2024) incorporate causal mechanisms to generate longer sequences or extend video-generation models into world modeling by forecasting future states Hu et al. (2023); Bruce et al. (2024); Wang et al. (2023). In this paper, we adopt DynamicCrafter as our base I2V framework due to its open-source availability and widespread use. We also ensure compatibility with modern DiT architectures, although that is not our main focus here.

Video Pretraining for Robotics. GR-2 Cheang et al. (2024) presents a generalizable robot manipulation framework that pretrains on large-scale internet videos, then fine-tunes on both video generation and action prediction for robotic trajectories. LAPA Ye et al. (2024) uses non-robot action videos for representation learning, mapping discrete latent actions (via VQ-VAE) to robotic manipulation tasks through a VLA model. SEER Tian et al. (2024) further explores inverse dynamics pretraining to boost performance. AVID Rigter et al. (2024) employs DynamicCrafter Xing et al. (2025) as its foundation, using an adapter for the robotics domain. VidMan Wen et al. (2024), based on OpenSora Zheng et al. (2024), focuses on environment prediction before action generation but is limited to 2D image space. In contrast, we propose generating long-sequence futures via a novel data generation engine, capturing richer motion information vital for robotics.

4D Generation. Recent progress Chen and Wang (2024) allows reconstruction of dynamic scenes from 2D videos using 3D GS (GS) Kerbl et al. (2023) and Neural Radiance Fields (NeRF) Mildenhall et al. (2021). Prior approaches approximate the spatio-temporal 4D volume with sets of 4D Gaussians Yang et al. (2023), jointly optimizing geometry and motion in canonical space Wu et al. (2024a). More recent advancements Li et al. (2024) employ customized sampling for multi-view video diffusion models, particularly for single dynamic objects. DimensionX Sun et al. (2024) leverages multiple LoRAs Hu et al. (2021) for diverse camera motions, while Cat4D Wu et al. (2024b) uses a single multi-view diffusion model to generate videos for dynamic 3D reconstruction. By contrast, our method produces videos from a Free Anchor View tailored for robotic manipulation tasks. In our offline data flywheel stage, GS complements video-generation models to mitigate the Sim2Real gap.

3 Methods

ENERVERSE comprises several designs, including a chunk-wise autoregressive generation framework and the FAV design for embodied future space generation. We additionally integrate a 4DGS to construct a data flywheel, referred to as ENERVERSE-D, and a policy head to generate physical actions, referred as ENERVERSE-A.

3.1 Next Chunk Diffusion

Chunk-wise Autoregressive Generation. As shown in Fig. 2, the observed latent sequence is represented as $\mathbf{o}_t^{1:K} = [\mathbf{o}_t^1, \dots, \mathbf{o}_t^K] \in \mathbb{R}^{K \times H \times W \times C}$, encoded by a pre-trained Variational Autoencoder (VAE). Here, K denotes the number of observed frames, $H \times W$ represents the spatial resolution, C is the number of channels, and t is the denoising step. Similarly, the latent representation of the rendered image is given by $\mathbf{r}_t^{1:J} \in \mathbb{R}^{J \times H \times W \times C}$. For simplicity, we treat \mathbf{r} as a special case of \mathbf{o} . The predicted latent sequence is denoted as $\mathbf{z}_t^{1:M} = [\mathbf{z}_t^1, \dots, \mathbf{z}_t^M] \in \mathbb{R}^{M \times H \times W \times C}$. The goal is to develop a video diffusion model that generates these predicted latents conditioned on $\mathbf{o}_0^{1:K}$ and a textual prompt \mathbf{c} , following the conditional probability: $p_\theta(\mathbf{z}_t^{1:M} | \mathbf{c}, \mathbf{o}_t^{1:K})$. Here, θ represents the parameters of the denoising network, which is defined as $\epsilon_\theta(\mathbf{z}_t^{1:M}, \mathbf{c}, \mathbf{o}_t^{1:K}, t)$. The network is trained to predict the ground truth noise ϵ from the noisy frame targets

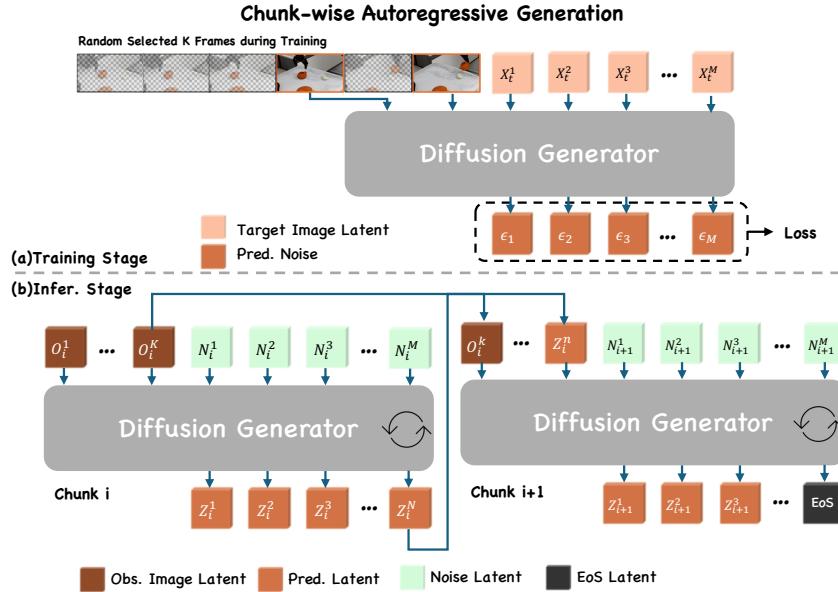


Figure 2 An overview of our chunk-wise autoregressive generation approach. (a) During training, random clean frames selected from consecutive sequences are combined with noisy frames to predict denoised latents. (b) In the inference phase, once new denoised frames are generated, they become the next set of clean frames for subsequent inference steps. This iterative procedure continues until the EoS frame is detected. For clarity, we illustrate only a single view of the autoregressive process, although multi-view generation is fully supported by the model.

by optimizing the loss function:

$$\min_{\theta} \mathbb{E}_{t, \mathbf{z} \sim z_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left\| \epsilon - \epsilon_{\theta}(\mathbf{z}_t^{1:M}, \mathbf{o}_t^{1:K}, t) \right\|_2^2,$$

where ϵ is the sampled ground truth noise, and θ denotes the learnable parameters. Consistent with prior work [Salimans and Ho \(2022\)](#), we predict v in practice. After training, the denoised data \mathbf{z}_0 can be derived from random noise \mathbf{z}_T through iterative denoising.

During inference, the diffusion generator takes both clean and noisy frames as input to produce M denoised frames. The newly generated frames serve as clean inputs for subsequent iterations, and this process repeats until detecting a predefined End-of-Sequence (EOS) frame. As the diffusion generation operates on latent frames, the L1 distance of each frame to the EOS is computed. If this distance falls below a predefined threshold, generation is terminated. In practice, this threshold-based EOS detection is highly effective.

Context Frame Mechanism. Instead of the conventional approach of using consecutive frames as the clean frame context for chunk prediction during training, we propose using sparsely sampled frames as the clean frame context. This approach leverages the redundancy often present in video data, allowing approximately 80% of frames to be discarded without compromising training effectiveness. Additionally, the high frame-dropping ratio enhances the model's robustness, particularly in handling out-of-distribution (OOD) scenarios such as covariant shift problems commonly encountered in the robot learning domain. From a representation learning perspective, this randomized sampling strategy promotes a deeper understanding of chunk prediction, potentially outperforming methods that rely on continuous frames.

During inference, clean frames are derived from observed or rendered frames and denoised using a sliding window approach. This technique ensures a smooth transition between observed and generated frames while improving efficiency and reducing GPU memory consumption.

3.2 4D Embodied Space Generation

Single-view or fixed multi-anchor view [Goyal et al. \(2024\)](#) approaches face challenges in handling occlusion and physical constraints in complex 3D environments, such as kitchens, where fixed cameras may be obstructed or infeasible. Similarly, wrist-mounted cameras complicate policy learning by coupling environmental dynamics

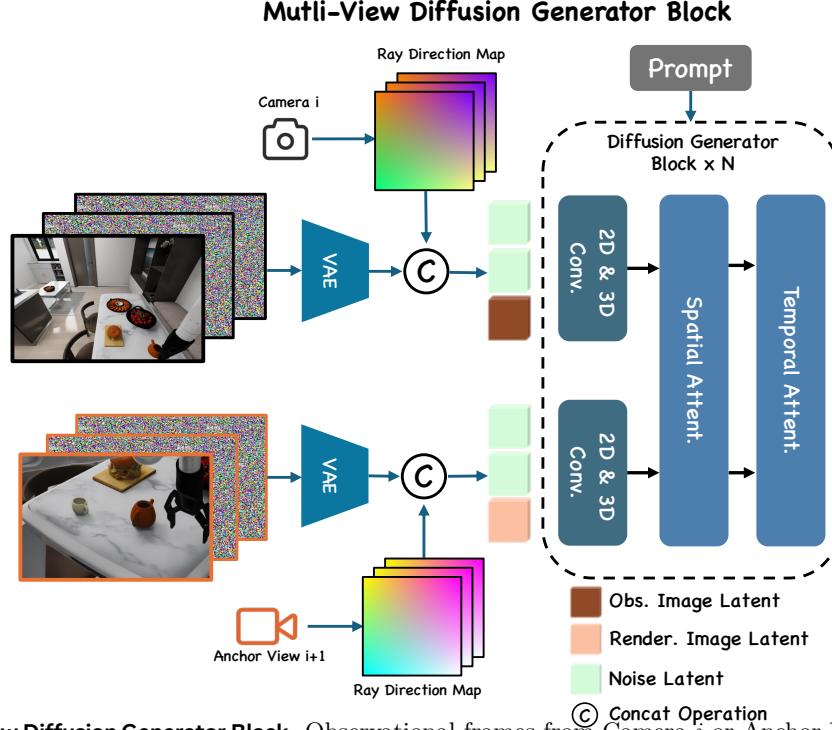


Figure 3 Multi-View Diffusion Generator Block. Observational frames from Camera i or Anchor View $i + 1$ is encoded using a VAE and the corresponding ray direction maps are then concatenated with the video latents. Subsequently, a combination of convolutional layers and attention mechanisms is applied. Notably, the observational frames from the camera are optional.

with the robot’s motion. To address these limitations, we extend the diffusion generator described in Sec. 3.1 to a Free Anchor View video generation pipeline.

As illustrated in Fig. 3, our method directly generates multi-view latents, represented as $\mathbf{z}_t^{1:M} \in \mathbb{R}^{M \times V \times H \times W \times C}$, where V denotes the number of views. To ensure consistency across different views, a ray direction map, encoding intrinsic and extrinsic camera parameters, is concatenated with the corresponding image latents along the channel dimension, enabling view-aware generation through ray casting ???. Additionally, spatial attention is applied along the V dimension to model cross-view relationships, ensuring coherent and consistent outputs while preserving the geometric relationships among objects in the scene.

Real-World Data Flywheels. Collecting calibrated multi-camera observations in real-world settings is expensive and labor-intensive. Consequently, we rely on data from simulators, which often exhibit significant domain gaps when applied to real-world scenarios. These gaps, including discrepancies in visual appearance and metric accuracy, hinder their direct applicability. To address these issues, we propose a multi-stage data generation pipeline that utilizes sparse observations to generate multi-view videos of a scene as shown in Fig. 4. First, a base model ENERVERSE is fine-tuned to be capable of receiving a **complete offline** observation sequence. When inferring, observations captured from multiple mounted cameras, covering robotic arm motion and scene dynamics to ensure cross-view consistency, are used to construct a 4D Gaussian representation via GS. Once the 3D scene reconstruction is complete, the anchor views are rendered to obtain higher-precision, geometrically consistent observations. These rendered observations are iteratively refined with ENERVERSE-D to generate pseudo-ground truth. After collecting sufficient real-world multi-view video data, we further fine-tune the multi-view video generator. This iterative process reduces noise, improves reconstruction accuracy, and facilitates Sim2Real domain adaptation, ultimately producing large-scale, high-quality datasets critical for training 4D generation models.

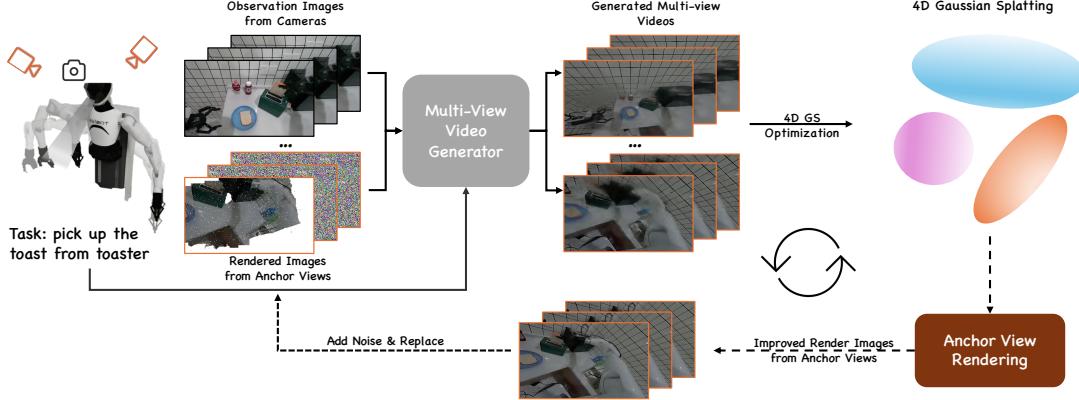


Figure 4 The pipeline for ENERVERSE as a data engine. Observation images from multiple cameras and rendered anchor view images are processed by the multi-view video generator to produce denoised multi-view videos. These videos, along with their camera poses, are used in 4DGS for 4D scene reconstruction. The reconstructed 3D content is rendered into anchor views to generate high-precision images. These high-quality rendered images are iteratively refined and fed back into the pipeline, improving motion consistency and reconstruction accuracy. This iterative process ensures geometric consistency and produces high-fidelity outputs suitable for applications.

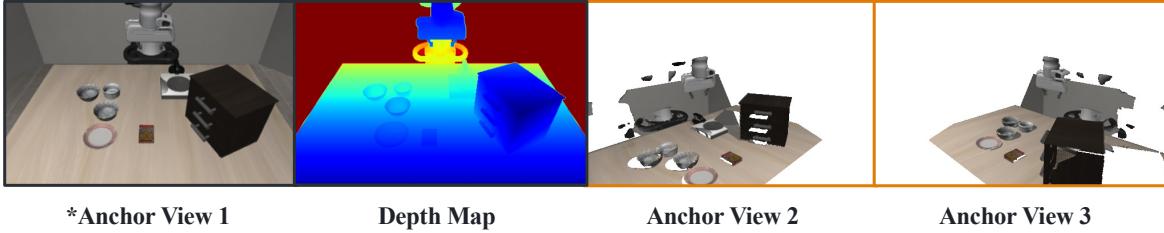


Figure 5 Visualization of FAVs on the LIBERO. Anchor View 1 represents the observation image captured by a mounted camera. Anchor View 2 and Anchor View 3 are generated by rendering from a point cloud reconstructed from Anchor View 1 using depth wrapping.

3.3 From 4D Embodied Space to Physical Action

In addition to video generation, we integrate a policy head into the diffusion generator networks, enabling the generation of videos and corresponding actions after the extensive future space generation pretraining. The input latent vector of the policy head is extracted from the middle layer of the video diffusion generator, as shown in Fig. 1. To improve efficiency, this latent vector is taken from the noisiest step of the diffusion process, i.e., the feature vector after the first denoising step, reducing computational overhead. During policy prediction, the sparse memory stores visual images that are either observed or reconstructed FAVs under a multi-view setup as shown in Fig. 5. Actions are predicted in chunks, making the approach well-suited for time-sensitive robotic control tasks.

4 Experiments

To demonstrate the effectiveness of proposed method, we evaluate ENERVERSE in two different domains, e.g., video generation quality and robotic policy performance.

4.1 Experiment Settings

Training Data. We selected several public datasets characterized by well-defined task logic, including RT-1 [Brohan et al. \(2022\)](#), Taco-Play [Rosete-Beas et al. \(2022\)](#), ManiSkill [Gu et al. \(2023\)](#), BridgeV1 [Walke et al. \(2023\)](#), LanguageTable [Lynch et al. \(2023\)](#) and RoboTurk [Mandlekar et al. \(2019\)](#), for pretraining. During

Method	Atomic Task						Long Task
	PSNR↑	FVD↓	Quality↑	Seman.↑	Consist.↑	Continuity↑	
DC-FN	25.42	445.94	54	97	92	80	×
EnerVerse	26.1	404.65	59	97	89	90	✓

Table 1 Performance comparison between DynamiCrafter (FN) and our proposed approach across Atomic Task metrics (Quantitative Comparison and User Study) and Long Task ability. The proposed method outperforms DynamiCrafter (FN) in most metrics, demonstrating its effectiveness in video generation and task performance.

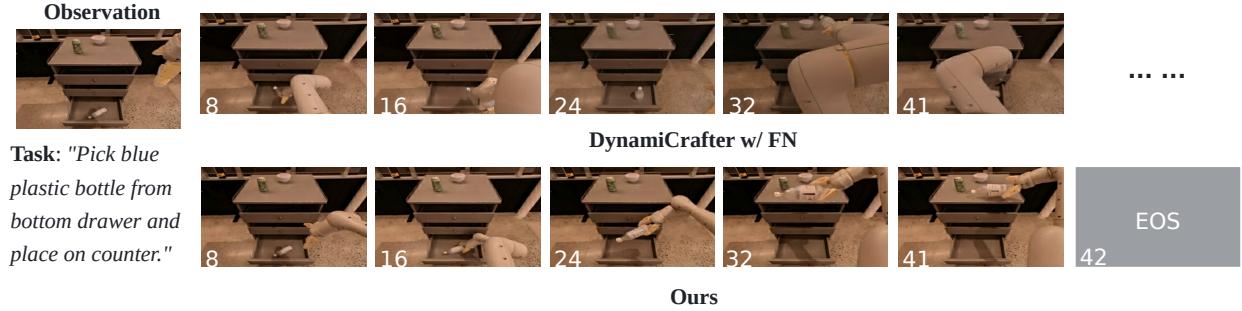


Figure 6 Qualitative comparison for single view video generation between ENERVERSE and DynamiCrafter(FN) on RT-1 dataset. Since ENERVERSE predict EOS frame at 42th frame for this task, we visualize up-to 42th frame sampled from both generated sequence. The sequences generated by DynamiCrafter(FN) did not maintain the logic and produce many hallucinations as the sequence grew. In contrast, the sequence generated by ENERVERSE was logically coherent, continuously and completely generating the future space of the entire task, and accurately predicting the EOS frame.

pretraining , only video frames were utilized for video generation training. Furthermore, we constructed a dataset containing multi-anchor view video ground truths using the Isaac Sim simulator Mittal et al. (2023). The FAV generation model was trained by leveraging the weights derived from the single-view video generation model. For the policy planning task, fine-tuning with a limited quantity of demonstration data from specific scenarios proved sufficient to attain state-of-the-art performance. To mitigate domain gaps encountered when training with heterogeneous data, we employed domain embeddings inspired by Wang et al. (2024). Specifically, distinct domain embeddings were allocated to each sub-dataset. In subsequent space generation and policy planning, these embeddings were integrated with the diffusion timestep embeddings prior to input into the diffusion model. This methodology effectively alleviated conflicts arising from discrepancies in entities, task types, and visual styles.

Training Details. Our model is conducted based on UNet-based Video Diffusion Models (VDM) Xing et al. (2025), and can be easily adapted to DiT Peebles and Xie (2023b) architectures. In our experiments on generating embodied future spaces, we identified that chunk size significantly influences model performance. Comparative analyses utilizing chunk sizes of 1, 4, 8, and 16 revealed that the model exhibited optimal robustness when employing a chunk size of 8 (further details regarding these experiments can be found in the supplementary material). Following the methodology outlined in Bruce et al. (2024), we introduced corruptive noise to the frames within the memory context. To alleviate degradation in autoregressive generation, the intensity of this noise was modulated in a cosine-related manner relative to the distance from the current moment. In the policy prediction experiment, the action head adopts the Diffusion Policy (DP) architecture Chi et al. (2023), with a total of 190M parameters. For the condition of the DP head, we utilize the feature before middle block of the UNet in the first denoise step, and calculate the mean value over spatial dimension, resulting in a final shape of $T \times C$, where T is the length of video and C is the number of channels before middle block. The rendered FAV images are with 512×320 image sizes and the action header predicts the delta pose.

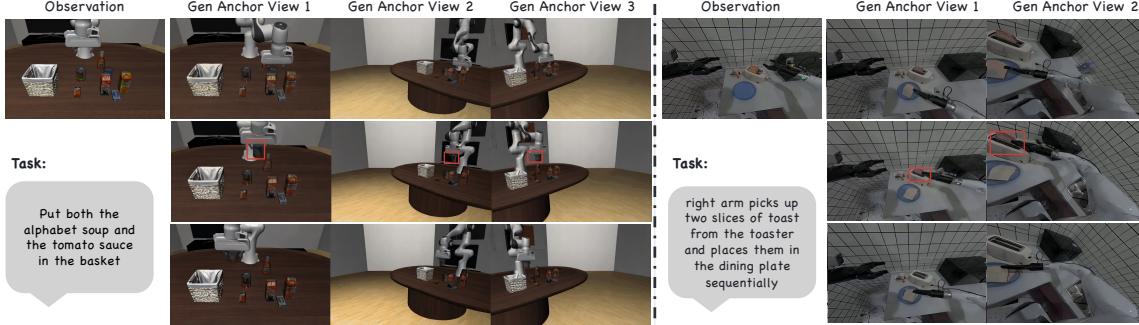


Figure 7 Qualitative results for multi-anchor view generation on LIBERO (left) and real-world manipulation data (right), from AgiBot World [AgiBot \(2024\)](#). One view is overlapped with a fixed RGB sensor and other views are manually set. Visualized Frames are uniform. The consistency of objects across views by a red rectangle.

Model	Visual Input	Spatial	Object	Goal	Long	Avg.
Diffusion Policy	One Third View	78.3	92.5	68.3	50.5	72.4
Octo	One Third View	78.9	85.7	84.6	51.1	75.1
OpenVLA	One Third View	84.7	88.4	79.2	53.7	76.5
MDT	One Third & One Wrist View	78.5	87.5	73.5	64.8	76.1
MAIL	One Third & One Wrist View	74.3	90.1	81.8	78.6	83.5
EnerVerse	One FAV	92.1	93.2	78.1	73.0	84.1
EnerVerse	Three FAVs	91.2	97.7	85.0	80.0	88.5

Table 2 Evaluation results on the LIBERO benchmark across four task suites.

4.2 Comparison Results

Embodied Future Space Generation. Following AVID [Rigter et al. \(2024\)](#), we assess video generation quality utilizing the RT-1 [Brohan et al. \(2022\)](#) dataset. To create a comparable baseline, we fine-tune DynamicCrafter on the RT-1 dataset and run inference iteratively with FreeNoise [Qiu et al. \(2023\)](#) to enable long video generation(DC-FN). For evaluation, we generate 200 synthetic videos with varied lengths by conditioning the models on the initial frame and task instructions, subsequently comparing the generated videos against the ground truth using standard metrics such as PSNR and FVD. However, while these metrics primarily evaluate visual quality, embodied tasks necessitate additional considerations, including semantic alignment with instructions, workspace consistency across frames, and motion continuity. To address these higher-order aspects, we execute a user study involving robotics experts, assessing the generated videos based on semantic accuracy, frame consistency, and motion continuity.

Tab. 1 illustrates that our method substantially outperforms DynamicCrafter (FN) in both quantitative and qualitative evaluations. In terms of quantitative metrics, our approach achieves a higher PSNR and a lower FVD. These findings indicate that our method produces videos of superior visual quality and enhanced temporal dynamics. In the user study, our method secures a higher quality score and exceeds DynamicCrafter in motion continuity, which is essential for robotic manipulation tasks. Although both methods attain equivalent semantic accuracy, this suggests that our approach effectively preserves instruction alignment while delivering superior overall performance. Moreover, our method uniquely accommodates long tasks, as evidenced by its successful execution of long-range manipulation scenarios, whereas DynamicCrafter falters in this domain. We also provide a qualitative comparison in Fig. 6.

Multi-View Consistency. In this section, we qualitatively demonstrate the capability of ENERVERSE to generate multi-view videos of the same scene while ensuring consistency across anchor views. Furthermore, each view attains high-quality image generation, thereby highlighting the robustness of our approach. As shown in Fig. 7, ENERVERSE could generate high-quality multi-anchor view videos in both simulator and real-world settings.

Robotic Policy Evaluation Following the evaluation protocol in OpenVLA [Kim et al. \(2024\)](#), we evaluate robotic policies using the LIBERO [Liu et al. \(2024\)](#) benchmark, which consists of four distinct task suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long. Each suite contains 10 tasks, each with 50 human demonstrations. For each task suite, a separate policy model is fine-tuned. We compare our method against five baselines: *Diffusion Policy* [Chi et al. \(2023\)](#), a direct action learning policy trained from scratch; *Octo Team* [Team et al. \(2024\)](#), a transformer-based policy model fine-tuned on the target dataset; *OpenVLA*, a 7B vision-language-action (VLA) model fine-tuned on the target dataset; *MDT* [Reuss et al. \(2024\)](#), a diffusion transformer-based policy with an auxiliary MAE loss; *MAIL* [Jia et al. \(2024\)](#), a policy model with Mamba [Gu and Dao \(2023\)](#) in an encoder-decoder structure. For evaluation, all models are tested across tasks using 50 rollouts per task, with results averaged over three random seeds. Experiments with ENERVERSE-A are conducted under two setups: a single static-camera, consistent with OpenVLA, and a three-static-camera configuration as shown in Fig. 5.

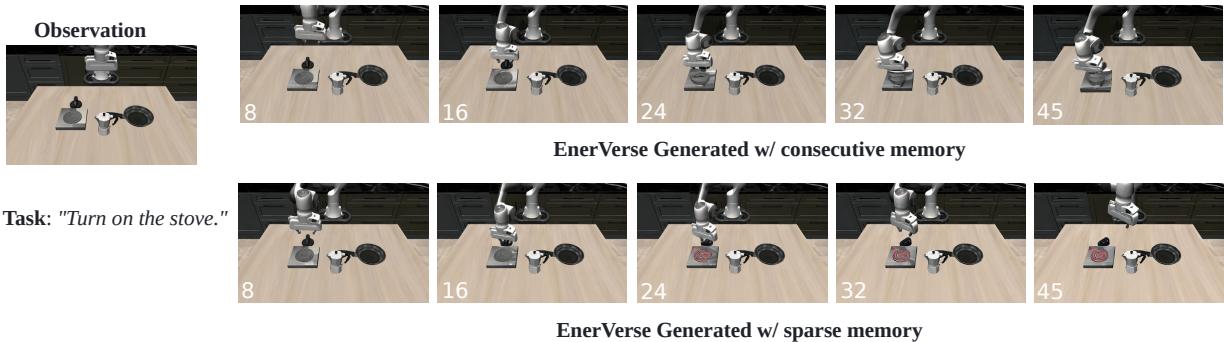


Figure 8 Ablation results for context memory mechanism in video generation. Providing history information to the generation model with consecutive context (first line) often leads to unexpected model collapse while the model with sparse memory (second line) shows robust performance and save much computing resources.

As shown in Tab. 2, ENERVERSE achieves state-of-the-art performance across the LIBERO benchmark, significantly surpassing all baselines. With a One Third View input, it achieves an average score of 84.0, outperforming strong baselines like MAIL (83.5) and OpenVLA (76.5). The Three Third View configuration further enhances performance, achieving the highest average score of 88.5, demonstrating the value of richer visual input. The model’s balanced performance across all tasks, particularly excelling in Object and Goal tasks, underscores its robustness and adaptability.

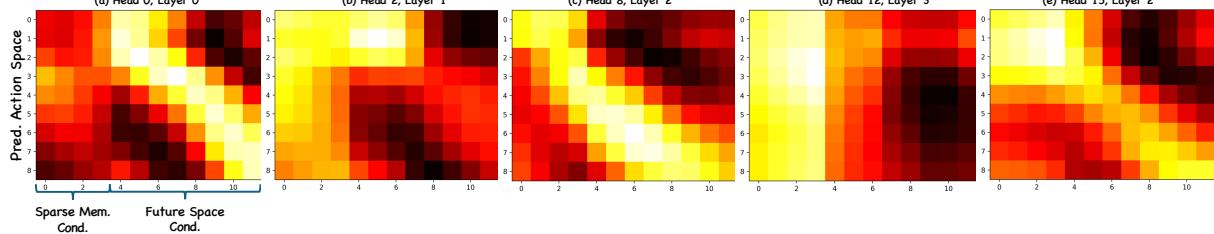
4.3 Further Studies

In this section, we explore several key design choices for ENERVERSE. First, we examine the significance of the proposed sparse memory mechanism, which plays a critical role in both policy learning and video generation. Second, we discuss the training strategy utilized in ENERVERSE. Third, we analyze the alignment between the predicted action spaces and visual spaces through attention map analysis. Finally, we introduce the real-world experiment setup.

Sparse Memory Mechanism. We evaluate the effectiveness of our sparse memory mechanism in both policy learning and video generation. The evaluation is conducted on the LIBERO-Long task suite, as this suite involves significantly longer task execution steps, requiring the policy to exhibit strong long-range memory and task reasoning capabilities. The evaluation is performed with a single visual input. As shown in Tab. 3, the absence of the *sparse memory* results in significant performance degradation, with the policy achieving only 30.8 compared to 73 when the sparse memory mechanism is applied. Similarly, Fig. 8 demonstrates that when the video generator operates without sparse memory, the model experiences unexpected collapse and fails to recover in out-of-distribution (OOD) scenarios. In contrast, the sparse memory mechanism ensures robust performance while also saving computational resources.

Training Strategy Analysis. To analyze the impact of different training strategies on robotic policy learning, we trained four robotic policies on the LIBERO-Spatial task suite using the following approaches: (1) training the entire ENERVERSE from scratch using only policy loss optimization; (2) training the entire ENERVERSE as in (1) but initialized with pretrained weights from a general video generator, e.g. DynamiCrafter(DC) [Xing](#)

Setup	w/o Sparse Memory	w Sparse Memory
LIBERO-Long-SV	30.8	73

Table 3 Sparse Memory analysis on LIBERO-Long.**Figure 9** Attention maps from different heads and layers of the model. The y-axis (Query) represents the predicted action space (8 steps), while the x-axis (Key-Value) spans Sparse Memory (first 4 columns) and predicted future space (last 8 columns). Bright yellow indicates high attention, showing how the model focuses on memory (left) and future predictions (right) when generating actions.

et al. (2025), which is trained with the general natural videos; (3) co-training ENERVERSE by optimizing both the robotic policy action loss and the video generation loss simultaneously; and (4) the default two-stage training strategy, where the video generator is pretrained first, followed by fine-tuning ENERVERSE using only robotic policy loss optimization.

Strategy	All-Scratch	With DC Pretrain.	One-Stage Co-Train	Two-Stage Finetune
LIBERO-Spatial	Failed	79	86.3	92.1

Table 4 Performance comparison of different training strategies on the LIBERO-Spatial task suite.

As shown in Tab. 4, training ENERVERSE from scratch without loading pretrained weights failed to converge, underscoring the importance of robust initialization. Another possible reason for this failure could be the relatively limited training data compared to the number of network parameters. Initializing with pretrained weights improved performance (79), while jointly optimizing the policy loss and video generation loss in a one-stage co-training setup further increased performance to 86.3. This demonstrates that the video generation task enhances policy learning. Our default Two-Stage Fine-tuning strategy, which involves pretraining the video generator followed by fine-tuning ENERVERSE with policy loss optimization, achieved the best performance.

Attention Map Analysis. To further analyze the alignment between the predicted action space and the visual space, including the visual observations cached by our Sparse Memory Mechanism and the generated future space, we visualized the attention maps from the first several layers of the Cross-Attention Block in our policy head.

Fig. 9 illustrates attention maps from different heads and layers, showcasing the model’s hierarchical focus and the impact of our proposed embodied future space generation in facilitating robust action prediction. In Fig. 9(a), attention is distributed almost entirely across the future space, reflecting the model’s ability to leverage sparse memory conditions and generated predictions from the outset. In contrast, Fig. 9(d) shows the attention sharply focused on the sparse memory space, with minimal reliance on the generated future space, indicating that the model has transitioned to memory-based reasoning. Interestingly, Figures 9(c,e) demonstrate that the model effectively integrates information from both the sparse memory space and the predicted future space. Moreover, these attention maps reveal that earlier decision steps tend to prioritize sparse memory, while later action steps shift focus to the generated future space. These results validate that our generative pretraining effectively enhances the model’s ability to integrate temporal information, align predicted actions with future visual contexts, and make robust decisions.

Real-World Experiments. To evaluate the manipulation capabilities of ENERVERSE, we conducted real-world

experiments using commercial robotics in two challenging industrial scenarios. Unlike the evaluation tasks described in Sec. 4.2, these scenarios required precise manipulation and robust decision-making. In the first task, the robot placed blocks into designated compartments of a foam worktable, demanding accuracy due to the tight fit and visual similarity between the foam and table. In the second task, the robot sorted several transparent plastic objects, including a measuring cup and plate, where the transparency added complexity to object recognition and manipulation. For details, please refer to Appendix A.

5 Conclusion

In conclusion, ENERVERSE is a generative robotics foundation model that tackles multi-view video generation and long-range policy execution by modeling embodied future spaces. With sparse contextual memory and Free Anchor Views (FAVs), ENERVERSE enhances spatial reasoning and task adaptability. The ENERVERSE-D pipeline, combining generative modeling with 4DGS, bridges the sim-to-real gap, reducing reliance on real-world data. Integrated with a policy head, ENERVERSE-A achieves state-of-the-art performance in manipulation tasks.

References

- AgiBot. Agibot world. <https://agibot-world.com>, 2024.
- Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.
- 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.
- Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.
- Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024.
- Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023.
- Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- 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.
- Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. Rvt-2: Learning precise manipulation from few demonstrations. *arXiv preprint arXiv:2406.08545*, 2024.
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
- Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023.
- Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. *arXiv preprint arXiv:2411.18179*, 2024.
- Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.
- Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Xiaogang Jia, Qian Wang, Atalay Donat, Bowen Xing, Ge Li, Hongyi Zhou, Onur Celik, Denis Blessing, Rudolf Lioutikov, and Gerhard Neumann. Mail: Improving imitation learning with mamba. *arXiv preprint arXiv:2406.08234*, 2024.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

- Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanyvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- Christoph Lassner and Michael Zollhofer. Pulsar: Efficient sphere-based neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1440–1449, 2021.
- Bing Li, Cheng Zheng, Wenxuan Zhu, Jinjie Mai, Biao Zhang, Peter Wonka, and Bernard Ghanem. Vivid-zoo: Multi-view video generation with diffusion model. *arXiv preprint arXiv:2406.08659*, 2024.
- 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, 2024.
- Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.
- Ajay Mandlekar, Jonathan Booher, Max Spero, Albert Tung, Anchit Gupta, Yuke Zhu, Animesh Garg, Silvio Savarese, and Li Fei-Fei. Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1048–1055. IEEE, 2019.
- Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8(6):3740–3747, 2023. doi: 10.1109/LRA.2023.3270034.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023a.
- William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023b. <https://arxiv.org/abs/2212.09748>.
- Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling, 2023.
- Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhui Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.
- Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.
- Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. Avid: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822*, 2024.
- Erick Rosete-Beas, Oier Mees, Gabriel Kalweit, Joschka Boedecker, and Wolfram Burgard. Latent plans for task agnostic offline reinforcement learning. 2022.
- Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024.
- Octo Model Team, Dibya Ghosh, Homer Walk, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

- Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. <https://arxiv.org/abs/2412.15109>, 2024.
- Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.
- Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *arXiv preprint arXiv:2409.20537*, 2024.
- Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.
- Youpeng Wen, Junfan Lin, Yi Zhu, Jianhua Han, Hang Xu, Shen Zhao, and Xiaodan Liang. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. *arXiv preprint arXiv:2411.09153*, 2024.
- Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024a.
- Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2024b.
- Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025.
- Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023.
- Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.
- Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qing, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models, 2023.
- Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024. <https://github.com/hpcalitech/Open-Sora>.

Appendix

A Real-World Robotic Experiments

To evaluate the manipulation capabilities of ENERVERSE-A, we conducted real-world experiments. The robot is instructed to place blocks into designated compartments of a foam worktable, requiring accuracy due to the tight fit and visual similarity between the foam and table, as shown in Figure 10.

Compared with the general "Pick and Place" task, this task has additional challenges:

- The robot must follow natural language instructions, such as "Row One, Column Two," to identify the required compartment.
- The compartments are only slightly larger than the magnet blocks, transforming the pick-and-place task into a highly precise "insertion" operation.
- The magnet blocks are relatively heavy, requiring the robot gripper to grasp near the center of the block to ensure stability during manipulation.

Correspondingly, we define four evaluation metrics:

- **Grasp:** Indicates whether the robotic gripper holds the suitable part of the block and transfers it stably during manipulation. It has binary values: 0 for failure, 1 for success.
- **Place:** Determines whether the robot places the block into a possible compartment. A score of 0 indicates failure, 1 indicates a perfect placement, and 0.5 indicates that the block has some collisions with the foam during manipulation.
- **Instruction Following:** Evaluates whether the robot places the block into the desired compartment as instructed. It has binary values: 0 for failure, 1 for success.

The overall **Success** is calculated as the product of the individual factors. The policy was executed five times for each compartment, and the average scores are presented in Table 5. ENERVERSE-A demonstrates strong performance in most target positions. However, it fails to handle positions (3, 2) and (3, 3). We hypothesize that this limitation arises because these positions are located near the boundary of the robot's action space, making them challenging to reach. Demonstration videos are provided in the supplementary materials.

In addition to the block placement task, we conducted experiments on sorting transparent plastic objects, such as measuring cups and plates. Demonstration videos for these experiments are also included in the supplementary materials. For additional videos showcasing multi-view video generation and policy rollouts, please refer to the supplementary materials.

Target Position	Grasp	Place	Ins. Following	Success
(1,1)	1	1	1	1
(1,2)	1	1	1	1
(1,3)	1	0.8	1	0.8
(2,1)	1	0.7	1	0.7
(2,2)	1	1	1	1
(2,3)	1	0.8	1	0.8
(3,1)	1	0.7	1	0.7
(3,2)	1	1	0	0
(3,3)	1	1	0	0

Table 5 Performance of the robotic system in placing blocks into designated compartments. The task demands high precision due to the tight fit and visual similarity between the foam and table.

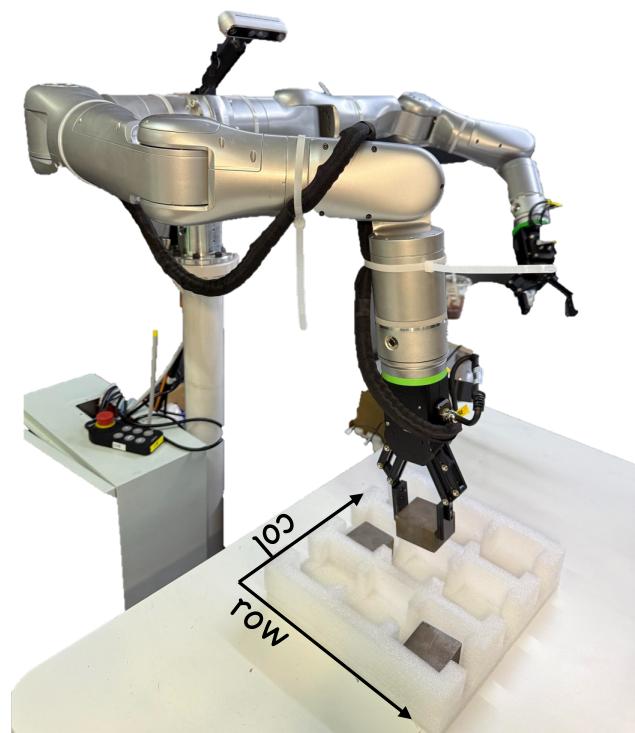


Figure 10 Real-world experimental setup. The overhead camera is the sole visual input used for the robot's operation.