# GigaBrain-0: A World Model-Powered Vision-Language-Action Model

**GigaAI**

Project Page: https://gigabrain0.github.io

**GigaBrain Team (alphabetical order):** Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jie Li, Jiagang Zhu, Lv Feng, Peng Li, Qiuping Deng, Runqi Ouyang, Wenkang Qin, Xinze Chen, Xiaofeng Wang, Yang Wang, Yifan Li, Yilong Li, Yiran Ding, Yuan Xu, Yun Ye, Yukun Zhou, Zhehao Dong, Zhenan Wang, Zhichao Liu, Zheng Zhu.
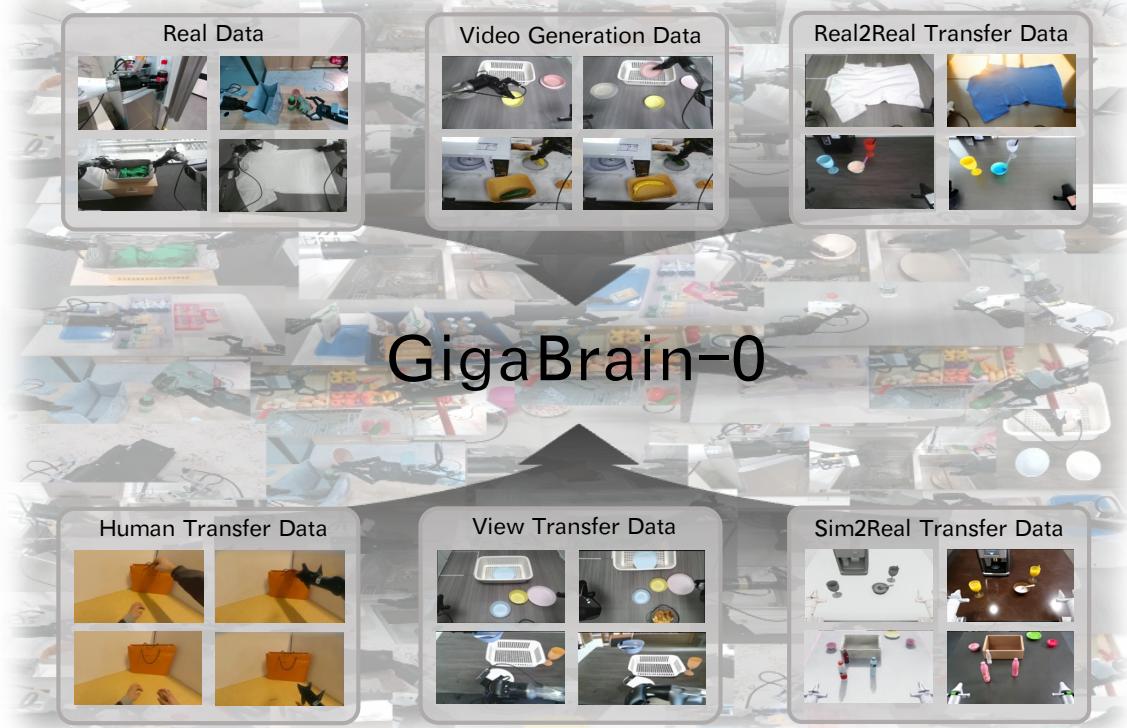
Figure 1: *GigaBrain-0* is a Vision-Language-Action (VLA) model trained on real-world robot data and diverse data generated by world models, including video generation data, Real2Real transfer data, human transfer data, view transfer data, and Sim2Real transfer data, to enhance its generalization in real-world environments.

## Abstract

Training Vision-Language-Action (VLA) models for generalist robots typically requires large-scale real-world robot data, which is expensive and time-consuming to collect. The inefficiency of data collection severely limits the scalability, and generalization capacity of current VLA systems. Therefore, we introduce *GigaBrain-0*, a novel VLA foundation model empowered by world model-generated data. By leveraging world models to generate diverse data at scale, *GigaBrain-0* significantly reduces reliance on real robot data while improving cross-task generalization. Our approach further improves policy robustness through RGBD input modeling and embodied Chain-of-Thought (CoT) supervision, enabling the model to reason about spatial geometry, object states, and long-horizon dependencies during task execution. This leads to substantial gains in real-world performance on dexterous, long-horizon, and mobile manipulation tasks. Extensive experiments demonstrate that *GigaBrain-0* achieves superior generalization across variations in appearances (e.g., textures, colors), object placements, and camera viewpoints. Additionally, we present *GigaBrain-0-Small*, an optimized lightweight variant designed to run efficiently on devices such as the NVIDIA Jetson AGX Orin.

# 1. Introduction

Recent advances in vision-language-action (VLA) models (Bjorck et al., 2025; Black et al., 2024; Bu et al., 2025; Cheang et al., 2025; Intelligence et al., 2025; Jiang et al., 2025; Zhai et al., 2025) have shown promising results in enabling generalist robots to understand high-level instructions, perceive their environments, and execute complex manipulation tasks. These models aim to bridge the gap between symbolic reasoning and embodied action by integrating visual inputs, natural language commands, and motor control into a unified framework. However, training such models typically relies on large-scale datasets of real-world robot interactions, and such data collection is not only expensive and time-consuming but also limited in diversity and scalability. The inefficiency of physical data collection poses a fundamental bottleneck to the development of robust, general-purpose robotic systems capable of operating across a wide range of environments, objects, and task configurations.

To overcome these limitations, we introduce *GigaBrain-0*, a novel VLA foundation model that leverages world model-generated data to reduce reliance on costly real-world robot data while improving generalization and data efficiency. By training on synthetic yet realistic trajectories generated by world models, as shown in Fig. 1, *GigaBrain-0* accesses a vast and diverse set of experiences, including variations in object materials, colors, lighting, and viewpoints, far beyond what is feasible to collect physically. This scalable data generation pipeline enables the model to learn robust representations that transfer effectively to real-world environments.

Our approach further enhances policy robustness through two framework innovations: RGBD input modeling and embodied Chain-of-Thought (CoT) supervision. By incorporating depth information, the model gains a richer understanding of 3D geometry and spatial layout, which is crucial for precise manipulation. Meanwhile, the embodied CoT framework encourages the model to generate intermediate reasoning steps, such as manipulation trajectories and subgoal planning, mimicking the cognitive processes underlying human problem-solving. This structured reasoning enables effective handling of long-horizon tasks and fine-grained actions that require sustained attention and sequential decision-making.

We evaluate *GigaBrain-0* through extensive real-world robotic deployments, including dexterous manipulation tasks (e.g., laundry folding, paper towel preparation), long-horizon tasks (e.g., table bussing, juice preparation), mobile manipulation tasks (e.g., boxes moving, laundry baskets moving). Results show that *GigaBrain-0* delivers consistently strong performance across this broad range of tasks and exhibits exceptional generalization under diverse conditions, including changes in appearances (e.g., texture, color), object placements, and camera viewpoints. Furthermore, we introduce *GigaBrain-0-Small*, an efficient variant optimized for deployment on hardware like the NVIDIA Jetson AGX Orin. Our work highlights the potential of world model-generated data as a scalable and effective alternative to traditional data collection paradigms, marking a significant step toward versatile, general-purpose robotic systems.

# 2. Related Works

## 2.1. Vision-Language-Action Models

Developing general-purpose robotic manipulation policies that can interpret high-level instructions and operate effectively in diverse physical environments remains a fundamental challenge in robotics. Prior approaches (Karamcheti et al., 2023; Nair et al., 2022; Radford et al., 2021; Xiao et al., 2022) have focused on leveraging heterogeneous datasets to learn rich, transferable representations that enhance policy robustness in complex and unseen scenarios. Inspired by advances in large language models, VLA frameworks (Bjorck et al., 2025; Black et al., 2024; Cheang et al., 2025; Doshi et al., 2024; Intelligence et al., 2025; Kim et al., 2024; Li et al., 2024; Liu et al., 2024; O'Neill et al., 2024; Pertsch et al., 2025; Qu et al., 2025; Team et al., 2024; Wang et al., 2024) have emerged as a promising paradigm, scaling model capacity and data volume to improve generalization across tasks and embodiments. These models typically build upon pretrained vision-language models (Alayrac et al., 2022; Bai et al., 2025; Beyer et al., 2024; Liu et al., 2023; Peng et al., 2023), integrating

multimodal inputs to generate action sequences, either through autoregressive token prediction or continuous action modeling via flow matching (Lipman et al., 2022; Liu, 2022). This shift from purely vision-to-action architectures to semantically grounded, language-conditioned policies has significantly expanded the scope of achievable behaviors. To support such data-hungry models, researchers have increasingly relied on cross-embodiment datasets (Dasari et al., 2019; Ebert et al., 2021; Khazatsky et al., 2024; O'Neill et al., 2024; Walke et al., 2023), combining publicly available logs from heterogeneous robotic platforms as well as large proprietary collections, such as those used in (Bjorck et al., 2025; Black et al., 2024; Cheang et al., 2025; Intelligence et al., 2025). Despite the performance gains, this reliance on extensive real-world interaction data raises practical concerns regarding scalability and cost.

## 2.2. World Models as Data Engines

The recent advancement in world model development (Agarwal et al., 2025; Alhaija et al., 2025; Assran et al., 2025; Jang et al., 2025; Jiang et al., 2025; Kong et al., 2024; Liao et al., 2025; Wang et al., 2025) has catalyzed a shift toward using synthetic data to narrow the sim-to-real divide in embodied intelligence (Zhu et al., 2024). In fields like autonomous driving, generative models are increasingly employed to simulate complex traffic scenarios, as evidenced by approaches such as (Gao et al., 2023, 2024; Hu et al., 2023; Ni et al., 2024, 2025; Ren et al., 2025; Russell et al., 2025; Wang et al., 2024; Zhao et al., 2024, 2025,). In robotics, where data acquisition is often limited by hardware and labor intensity, generative world models offer a promising alternative. Works including (Du et al., 2023; Feng et al., 2025; Jang et al., 2025; Yang et al., 2023; Zhou et al., 2024) leverage natural language prompts to generate plausible future trajectories, which are then converted into low-level control signals via inverse dynamics estimation. To improve the structural fidelity of synthesized outputs, TesserAct (Zhen et al., 2025) and Robot4DGen (Liu et al., 2025) introduce a unified multimodal generation framework that produces synchronized streams of RGB frames, depth maps, surface normals, or 3D point clouds. This enables the construction of temporally coherent 4D reconstructions, significantly advancing policy learning compared to models trained on RGB video alone. Further enhancements in scene diversity are achieved through techniques like background inpainting (Fang et al., 2025; Yuan et al., 2025), which alters environmental textures, and video-to-video translation methods (Alhaija et al., 2025; Dong et al., 2025; Li et al., 2025; Liu et al., 2025; Wang et al., 2025; Xu et al., 2025) that stylize or adapt visual dynamics. Notably, *GigaBrain-0* exploits the generative capacity of world models to produce highly varied data across texture, material, illumination, object placement, and camera viewpoints, providing rich, generalizable data source for training VLA models.

## 3. GigaBrain-0 Model

*GigaBrain-0* is an end-to-end VLA model $g_\theta$ that, given visual observations and high-level language instructions, reasons over embodied scenarios to generate compliant action sequences for controlling a wheeled bi-manual robot (e.g., Agilex, G1). To enhance prompt-following fidelity and enable smoother action generation, as illustrated in Fig. 2, *GigaBrain-0* adopts a mixture-of-transformers (Liang et al., 2024) architecture, it leverages a pretrained Vision-Language Model (VLM), PaliGemma2 (Steiner et al., 2024), to encode multimodal inputs, and employs an action Diffusion Transformer (DiT) (Peebles and Xie, 2023) with flow matching (Lipman et al., 2022) to predict action chunks. This hybrid architecture enables decoupled yet synergistic processing of semantic understanding and continuous action generation. During training, we introduce Knowledge Insulation (Driess et al., 2025) to mitigate interference between continuous action-space learning and the VLM's semantic reasoning capabilities. Furthermore, we augment the VLM head with discrete action token prediction (Pertsch et al., 2025), which significantly accelerates pretraining convergence.

To enhance spatial reasoning capabilities, we incorporate RGB-D data during pretraining. Given an input tensor of shape $B \times H \times W \times 4$ (RGB + depth), we first normalize the input and extract visual features using SigLIP (Zhai et al., 2023). To adapt SigLIP to RGB-D inputs, we extend its first convolutional layer with zero-initialized kernels for the depth channel. This preserves pretrained RGB feature extraction while enabling
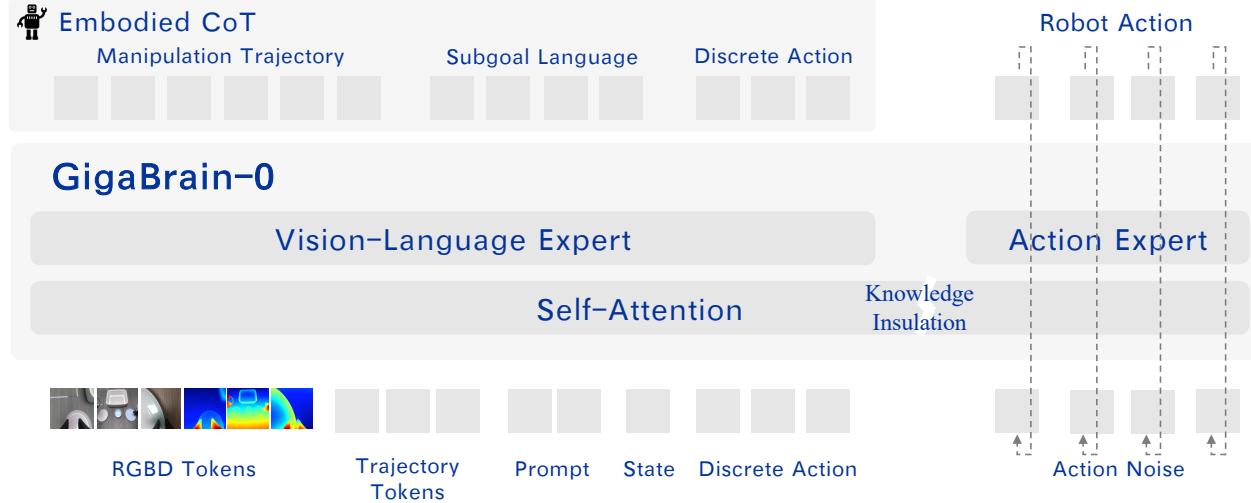
Figure 2: The framework of *GigaBrain-0*. *GigaBrain-0* takes RGB-D input to enhance spatial perception and outputs Embodied Chain-of-Thought (Embodied CoT) as an intermediate representation to strengthen embodied reasoning for manipulation. During training, *GigaBrain-0* employs Knowledge Insulation (Driess et al., 2025) to prevent interference between the optimization processes of action prediction and Embodied CoT generation.

depth-aware representation learning. Notably, SigLIP remains fully trainable throughout *GigaBrain-0*'s training, allowing adaptive fine-tuning to embodied RGB-D perception. During training, we randomly drop the depth channel (replacing it with zero padding) to ensure compatibility with RGB-only inputs during inference.

Inspired by Chain-of-Thought (CoT) reasoning in LLMs (Wei et al., 2022), we introduce Embodied CoT to improve *GigaBrain-0*'s reasoning in embodied environments. Unlike standard LLMs, *GigaBrain-0* explicitly generates intermediate reasoning tokens, including: (1) *manipulation trajectories*: 2D projections of end-effector paths onto the image plane, represented by 10 uniformly sampled keypoints; (2) *subgoal language*: natural language descriptions of intermediate objectives; and (3) *discrete action tokens*: discrete representations that accelerate training convergence of the subsequent DiT-based continuous action chunk prediction (Pertsch et al., 2025).

To balance model expressiveness with inference efficiency, we forgo autoregressive decoding for trajectory prediction. Instead, we introduce 10 learnable trajectory tokens as auxiliary inputs to the Vision-Language Model (VLM). During feature extraction, these tokens interact with the full visual context via bidirectional (non-causal) attention, enabling holistic spatial reasoning over the scene. The resulting output trajectory tokens are then passed through a lightweight GRU decoder to regress the 2D pixel-space coordinates of the end-effector's manipulation trajectory.

In contrast, subgoal language and discrete action tokens are generated autoregressively and supervised via standard next-token prediction. All components, including trajectory regression, language-based subgoals, discrete action tokens, and continuous action chunks predicted by the Diffusion Transformer (DiT) $f_\theta$, are jointly optimized under a unified objective:

$$\mathcal{L} = \mathbb{E}_{\mathcal{D},\tau,\epsilon} \left[ -\sum_{j=1}^{n-1} M_{\text{CoT},j} \log p_\theta \left( x_{j+1} \mid x_{1:j} \right) + \left\| \epsilon - a_{\text{chunk}} - f_\theta \left( a_{\text{chunk}}^{\tau,\epsilon} \right) \right\|^2 + \lambda \left\| \text{GRU}(\hat{\mathbf{t}}_{1:10}) - \mathbf{t}_{1:10} \right\|^2 \right], \quad (1)$$

where $\mathcal{D}$ denotes the training dataset. $\tau \in [0,1]$ is the flow-matching timestep, $\epsilon \sim \mathcal{N}(0,\mathbf{I})$ is Gaussian noise. $a_{\text{chunk}}^{\tau,\epsilon} = \tau \cdot a_{\text{chunk}} + (1-\tau) \cdot \epsilon$ is the noised action chunk used in flow matching. $M_{\text{CoT},j} \in \{0,1\}$ is a per-token mask indicating whether position $j$ belongs to the CoT reasoning stream (subgoal language or discrete actions). $\hat{\mathbf{t}}_{1:10}$ and $\mathbf{t}_{1:10}$ denote predicted and ground-truth 2D trajectory keypoints, respectively. $\lambda = 1$

Table 1: Comparison of training data usage across VLA models. *GigaBrain-0* leverages a diverse set of data sources to enhance generalization and reduce dependency on real-world robot data.

| Method | Public Data | Self-Collected Data | Egocentric Human Data | Video-Gen Data | Sim2Real Transfer Data | Real2Real Transfer Data | View Transfer Data |
|---|---|---|---|---|---|---|---|
| $\pi_0$ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| $\pi_{0.5}$ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| G0 | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| WALL-OSS | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| GR-3 | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| GR00T N1.5 | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| *GigaBrain-0* | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

is a hyperparameter balancing the trajectory regression loss. Notably, we do not manually assign loss weights to the language and action prediction terms, as Knowledge Insulation (Driess et al., 2025) inherently prevents interference between their optimization processes, allowing each stream to learn independently.

## 4. GigaBrain-0 Data

For Vision-Language-Action (VLA) models, training data diversity is paramount (Shi et al., 2025). While numerous public datasets (Bu et al., 2025; O'Neill et al., 2024; Wu et al., 2024) exist, they are often insufficient in scene variation, or task complexity to enable robust generalization. Real-world robot data collection suffers from high operational costs, low scaling efficiency, and limited environmental diversity, as most deployments repeatedly sample from the same narrow set of scenes.

Recent advances (Bjorck et al., 2025; Jang et al., 2025) have demonstrated that world models can effectively generate diverse and photorealistic training data to augment vision-language-action (VLA) capabilities. In *GigaBrain-0,* we further extend this paradigm by integrating a broad spectrum of world-model-generated data sources. As shown in Tab. 1, *GigaBrain-0* leverages a more diverse set of data sources compared to existing VLA counterparts (Bjorck et al., 2025; Black et al., 2024; Cheang et al., 2025; Intelligence et al., 2025; Jiang et al., 2025; Zhai et al., 2025). This expanded data diversity significantly reduces reliance on real-world robot-collected data while enhancing model generalization. We detail each data source in the following.

### 4.1. Real-World Data

**Data Source**. Our real-world data corpus integrates both publicly available datasets and proprietary data collected from our in-house robotic platforms. Publicly sourced datasets include AgiBotWorld (Bu et al., 2025), RoboMind (Wu et al., 2024), and Open X-Embodiment (O'Neill et al., 2024), which collectively provide foundational coverage of manipulation and locomotion tasks. In addition, we collect 1182 hours proprietary data using Agilex Cobot Magic platform (199 hours) and the AgiBot G1 platform (983 hours) across a total area of 3100m$^2$, spanning five broad environment categories: industrial, commercial, office, residential, and laboratory settings. These are further subdivided into 14 distinct real-world scenarios, including supermarkets, hotel lobbies, coffee shops, bubble tea stores, convenience stores, restaurants, warehouse material handling, industrial assembly lines, pantries, private residences, apartment interiors, meeting rooms, office workstations, and laboratories. The collected tasks range from basic pick-and-place operations to long-horizon sequential activities, mobile manipulation in dynamically changing layouts, and interactions with deformable objects, as illustrated in Fig. 3.

**Data Annotation & Processing**. For data annotation, if the captured RGB frames lack depth information, we employ MoGe (Wang et al., 2025) to generate metrically scaled depth maps. Regarding subgoal language annotation, we observe that VLMs struggle to accurately segment long-horizon tasks into meaningful subgoals,
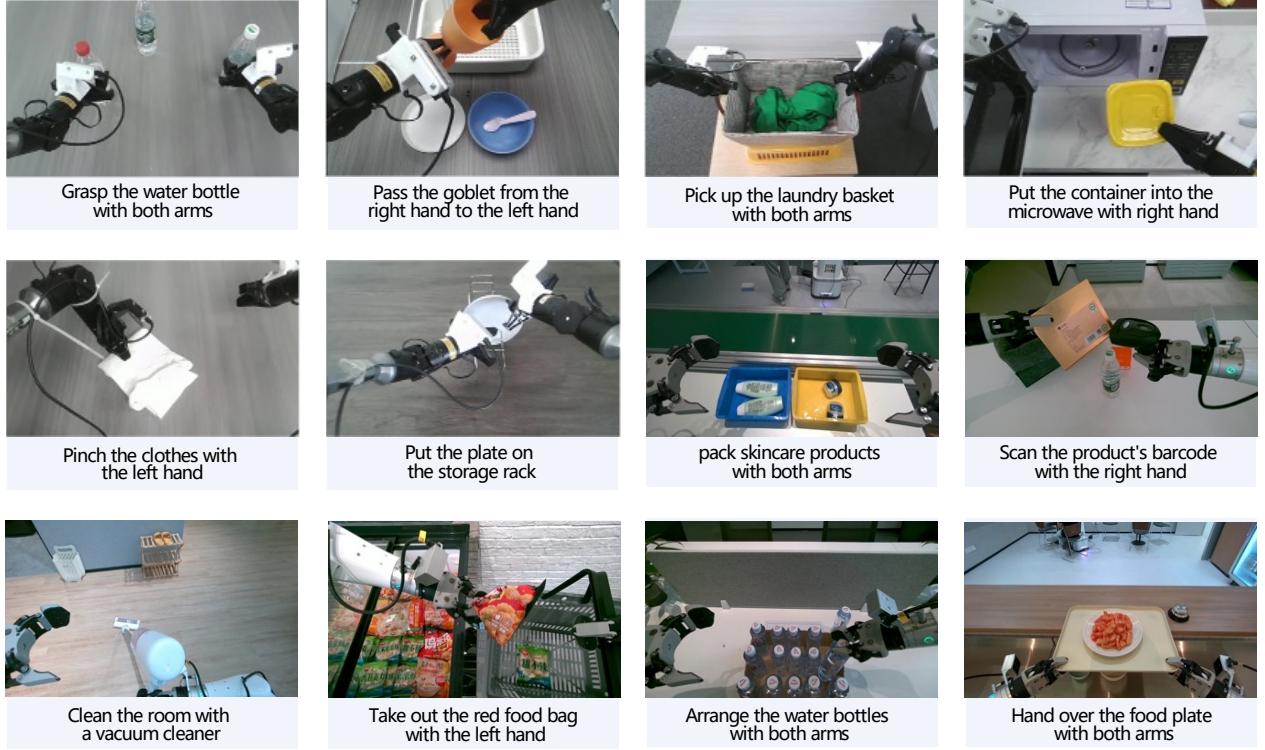
Figure 3: *GigaBrain-0*'s self-collected real-world robot data is gathered from PiPER arms and the AgiBot G1 platform, spanning diverse environments including homes, supermarkets, factories, and office settings.

while manual segmentation proves prohibitively time-consuming. To address this, we adopt an approach inspired by (James et al., 2020), leveraging gripper state transitions (e.g., open/close, grasp/release) to automatically segment trajectories into atomic subtasks. For each segmented subtask, we utilize Qwen-VL-2.5 (Bai et al., 2025) to generate subgoal language annotations. To mitigate hallucination and ensure consistency, we constrain the annotation process using a structured template and a predefined vocabulary of standardized action phrases (e.g., `pick [object]`, `place [object] into [container]`, `open [device]`), selected from a curated description library. For 2D manipulation trajectory annotation, we project the 3D end-effector coordinates into the head-mounted camera's image plane, yielding pixel-space motion traces aligned with visual observations. Notably, we annotate only a subset of the collected data and train our models using a mixture of fully annotated, partially annotated, and raw unannotated trajectories to maximize data utility while managing annotation costs. To further improve pretraining efficiency and reduce redundancy, we perform deduplication across the entire corpus. For each unique task, we retain at most 50 diverse demonstration trajectories, preserving behavioral variety while eliminating near-identical repetitions. This strategy enhances sample efficiency and promotes more robust and generalizable model learning.

## 4.2. World-Model-Generated Data

To overcome the limitations of physical data collection, we employ GigaWorld, our world-model framework, to generate diverse, physically plausible training sequences. GigaWorld synthesizes data through multiple complementary pipelines:

**Real2Real Transfer**. Real-world robot-collected videos are inherently constrained by the physical environments in which they are captured, including static backgrounds, fixed lighting conditions, and limited variation in object materials, textures, and colors. To overcome these limitations and significantly enhance visual and contextual diversity, we leverage GigaWorld to perform Real2Real Transfer: re-rendering real trajectories
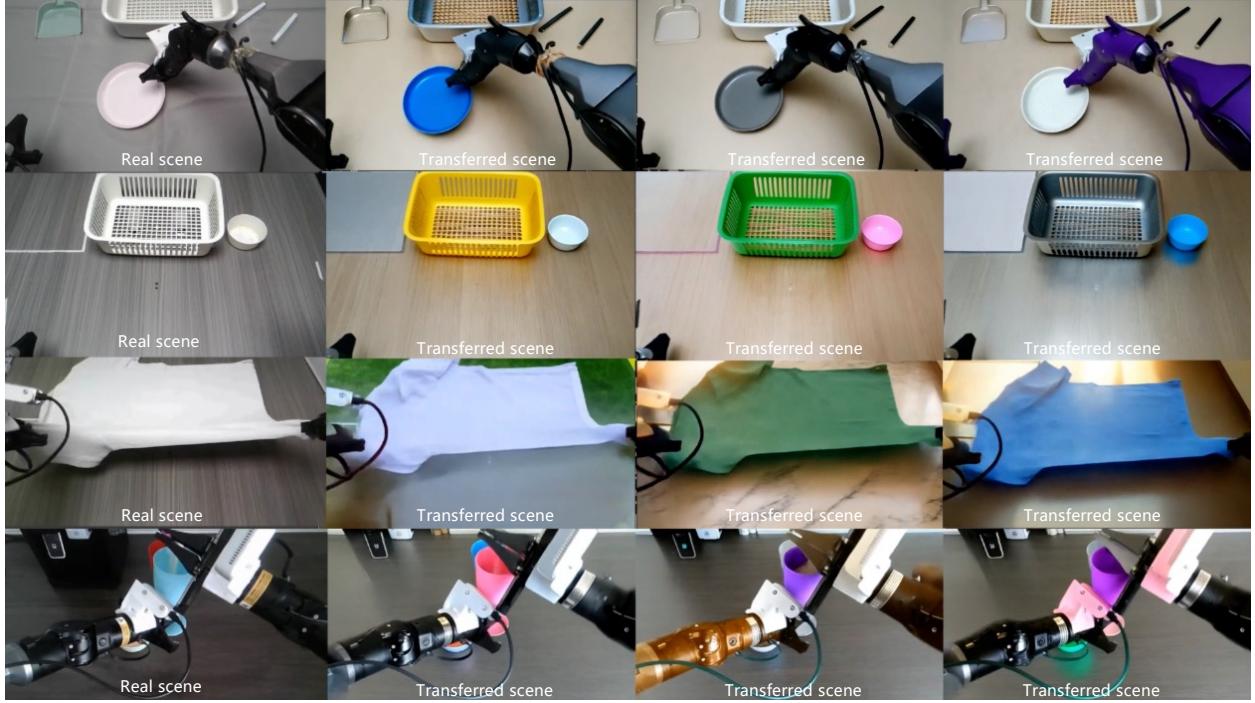
Figure 4: GigaWorld enables Real2Real apperance transfer by taking real-world captured data and generating generalized variations in texture, color, lighting, and material properties.

in synthetically altered but physically plausible visual contexts. Specifically, we train diffusion-based video generation model (Dong et al., 2025; Liu et al., 2025) conditioned on geometric and structural priors extracted from real footage. We employ VideoDepthAnything (Chen et al., 2025) to estimate per-frame depth maps and extract Canny edge maps to preserve object boundaries and scene structure. These signals serve as spatial control conditions via a ControlNet (Zhang et al., 2023) branch, enabling precise manipulation of appearance while preserving motion and layout consistency. During inference, for each real-world video clip, we generate approximately 10 visually distinct variants by textually prompting changes to foreground/background materials, surface textures, illumination conditions, and color palettes, all while maintaining the original action semantics and spatial dynamics, as shown in Fig. 4. This approach efficiently multiplies the effective diversity of real data without requiring additional physical collection.

**View Transfer**. Beyond texture and illumination variation, generalization across observation viewpoints is equally critical for robust embodied perception (Xing et al., 2025). To this end, we exploit GigaWorld's viewpoint generalization capability to augment real robot-collected data with novel camera perspectives while preserving 3D scene consistency. Specifically, to ensure geometric coherence under viewpoint changes, we project the original RGB frames into novel views using collected depth maps. If depth is unavailable in the source data, we annotate metric-scale depth using MoGe (Wang et al., 2025). The reprojected views inevitably contain occluded or incomplete regions, we inpaint these using a DiT-based video completion model (Xu et al., 2025) conditioned on the reprojected views. Notably, when the camera viewpoint shifts, the robot's end-effector must remain functionally consistent with the task, even though its joint configuration changes. We compute the new joint angles via IK based on the updated ego-pose and end-effector pose. The resulting articulated robot geometry is then rendered using its URDF model in a physics-aware simulation engine, and provided as a structural condition (Xu et al., 2025) to the generative model. To mitigate potential discrepancies between simulated and real robot dynamics, we optionally employ a differentiable physics engine (Wang et al., 2025) to fine-tune motion plausibility and close the sim-to-real physics gap. As shown in Fig. 5, from a single real-world trajectory, our pipeline generates multiple viewpoint-consistent renderings of the scene, complete
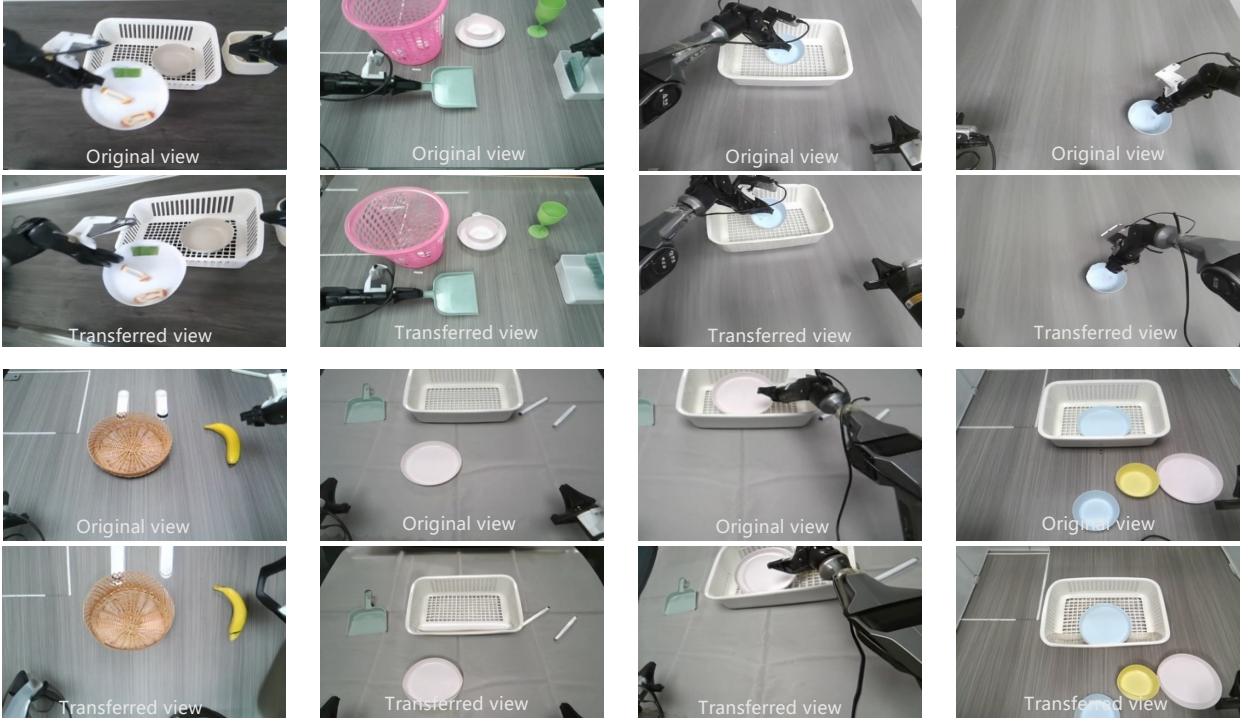
Figure 5: GigaWorld supports view transfer by re-rendering real-world captured data from diverse viewpoints, thereby enriching the dataset with varied perspective changes.

with dynamically adjusted robot poses that preserve task semantics and physical feasibility.

**Sim2Real Transfer**. While the methods described above augment real-world data, we further scale our training corpus by leveraging simulation assets to synthesize diverse embodied interaction sequences. Specifically, we compose manipulation scenes in Isaac Sim (NVIDIA) using either procedurally generated assets from EmbodiedGen (Wang et al., 2025) or curated objects from open-source repositories such as ArtVIP (Jin et al., 2025). Robot morphology is defined via URDF files, and end-effector trajectories are computed using IK to ensure physically plausible motion. To bridge the sim-to-real domain gap, particularly in visual appearance, we apply GigaWorld's Sim2Real Transfer pipeline. Similar in architecture to Real2Real Transfer, this method conditions a diffusion-based video generator (Dong et al., 2025; Liu et al., 2025) on depth maps exported from the simulation environment. Using text prompts, we dynamically alter surface textures, material reflectance, lighting conditions, and environmental clutter to produce photorealistic renderings that retain the original scene geometry and action semantics, as illustrated in Fig. 6. Crucially, unlike real-world data, simulation affords us full control over scene parameters: we can systematically vary object initial positions, camera viewpoints, background layouts, and even physics properties (e.g., friction, mass) to maximize combinatorial diversity.

**Human Video Transfer**. Human demonstration videos have emerged as a promising source for training embodied agents (Bu et al., 2025; Cheang et al., 2025; Kareer et al., 2025; Wang et al., 2023; Yang et al., 2025), offering diversity in tasks, environments, and interaction styles, far exceeding the scale and variety achievable through robotic data collection alone. However, raw human videos exhibit significant gaps when directly applied to robot learning, as egocentric footage often suffers from motion blur, unstable viewpoints, and a visual and kinematic mismatch between human hands and robotic end-effectors. To bridge this gap, we leverage GigaWorld's video inpainting capabilities to transform large-scale first-person human videos into stable, robot-centric demonstrations. Specifically, we convert videos from the EgoDex dataset (Hoque et al., 2025) into stabilized, robot-executable sequences with articulated mechanical arms replacing human hands. Specifically, we use SAM2 (Ravi et al., 2024) to segment and mask out human hands in each frame. The annotated 3D hand

Figure 6: GigaWorld enables Sim2Real transfer by generalizing simulation-collected data in terms of texture, color, lighting, and material properties to better bridge the domain gap and enhance realism.

wrist positions (provided in EgoDex) are treated as target end-effector poses for a simulated robot arm. We solve for corresponding joint angles via IK, then render the robot's URDF model using a physics-aware simulation engine. This rendered arm geometry serves as a structural condition for our diffusion-based generator (Li et al., 2025), ensuring kinematically plausible and visually consistent robot appearances. As shown in Fig. 7, the output is a stabilized, robot-embodied version of the original human demonstration, preserving task intent and spatial relationships while eliminating visual and kinematic domain gaps.

**Video Generation with Inverse Dynamics Modeling**. Given a single input image, GigaWorld can generate diverse embodied robotic operation videos conditioned on different textual prompts, as illustrated in Fig. 8. Furthermore, we leverage Inverse Dynamics Models (IDM) (Jang et al., 2025) to infer corresponding robot action sequences from these generated videos, which are then used as synthetic training data for embodied manipulation tasks.

**Multiview Video Generation**. In embodied manipulation scenarios, multiple cameras (e.g., head-mounted and wrist-mounted cameras) are often deployed to capture the operating environment from different viewpoints, creating a demand for generating temporally and spatially consistent multiview videos. To address this, GigaWorld adopts the multiview modeling approach from (Dong et al., 2025; Liu et al., 2025; Zhao et al., 2025), which concatenates noise maps from multiple views as input to the diffusion model. This design preserves the original diffusion architecture without modification and enables consistent multiview video generation with only minimal fine-tuning data. As shown in Fig. 9, GigaWorld is capable of generating diverse yet geometrically consistent multiview videos, demonstrating strong cross-view coherence and scene fidelity.

**Generation Efficiency**. Video diffusion models (Alhaija et al., 2025; Kong et al., 2024; Wang et al., 2025) suffer from low generation efficiency, often requiring tens of minutes to synthesize hundreds of frames resolution. To accelerate inference, GigaWorld employs NATTEN (Hassani et al., 2023) as a computationally efficient
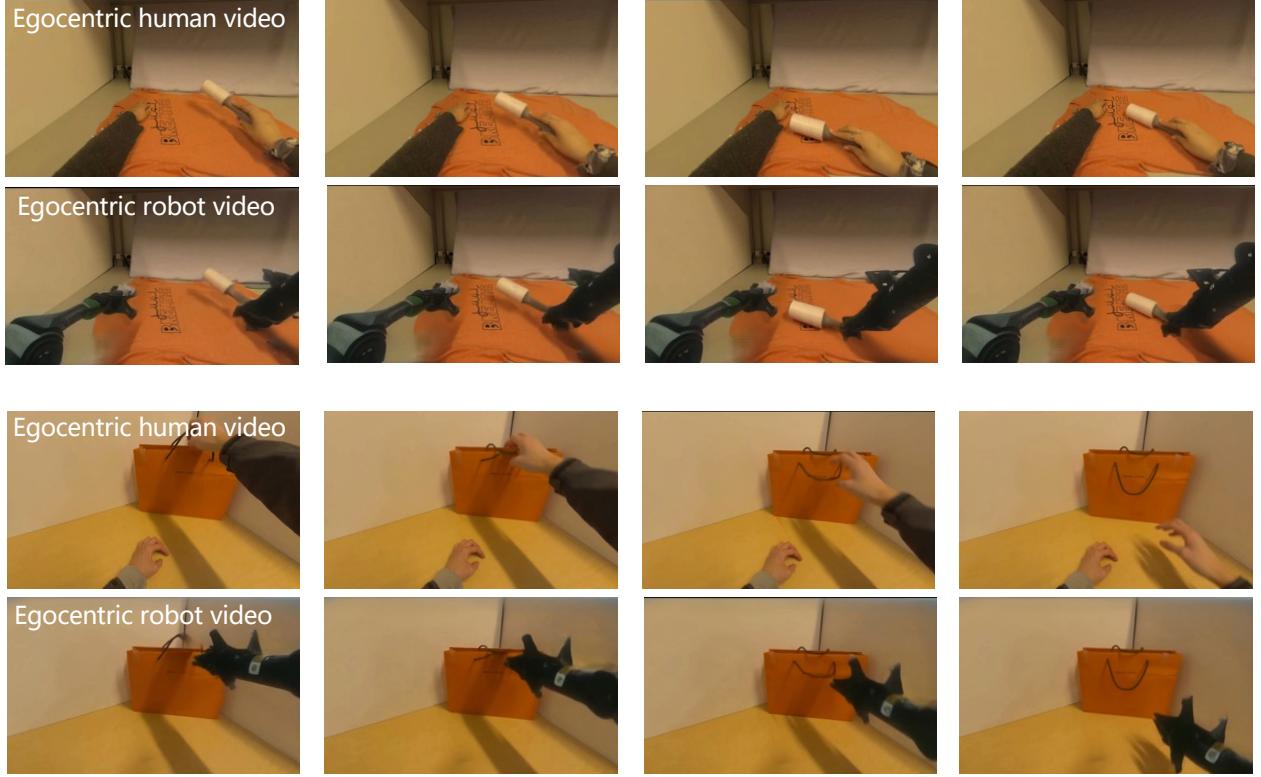
Figure 7: GigaWorld supports egocentric human video transfer by transforming first-person human hand actions into robotic manipulation scenarios, effectively mapping human demonstrations to robot-executable tasks.

alternative to standard self-attention. Furthermore, GigaWorld leverages step distillation (Yin et al., 2024) to reduce the denoising process from dozens of steps to a single-step generation. Combined with FP8-precision inference, these optimizations collectively yield over a $50\times$ speedup in video generation compared to baseline diffusion models.

**Generated Data Quality Inspection**. Generated videos inevitably contain hallucinations or artifacts that may degrade downstream training performance. To mitigate this, GigaWorld introduces a comprehensive quality assessment pipeline that evaluates generated videos across multiple dimensions: geometric consistency (Liu et al., 2025), multiview consistency (Liu et al., 2025), text-description alignment (Azzolini et al., 2025), and physical plausibility (Azzolini et al., 2025). Each video is assigned a composite quality score, which determines whether it is suitable for pre-training, fine-tuning, or should be discarded. All model and training details mentioned above will be fully elaborated in the upcoming GigaWorld technical report.

## 5. Experiment

To evaluate the real-world manipulation performance, we conducted robot experiments on two robotic platforms: the dual-arm PiPER robot platform and the AgiBot G1 platform. These experiments covered a diverse range of tasks, including dexterous manipulation (e.g., `laundry folding` and `paper towel preparation`), long-horizon tasks (e.g., `table bussing` and `juice preparation`), and mobile manipulation (e.g., `boxes moving` and `laundry basket moving`). Furthermore, we validated *GigaBrain-0*'s generalization capabilities, specifically in appearance, placement, and viewpoint generalization. We also performed on-device experiments to verify the real-time inference performance of the lightweight variant, *GigaBrain-0*-Small. In the following sections, we will elaborate on the experimental setup and present detailed performance comparisons.

Figure 8: GigaWorld can generate diverse future trajectories from the same initial frame under different text prompts, thereby augmenting the dataset with novel manipulation sequences.

## 5.1. Dexterous Manipulation Experiment

**Experiment Setup.** We evaluate *GigaBrain-0* on two dexterous manipulation tasks, `laundry folding` and `paper towel preparation`. For `laundry folding`, we deploy the model on the G1 robot platform and fine-tune it using 300 human-collected demonstration trajectories, the training is conducted with a batch size of 128 for 40K steps. For `paper towel preparation`, we use 100 demonstrations from PiPER arms, and the training is conducted with a batch size of 128 for 20K steps. We compare *GigaBrain-0* against $\pi_0$ (Black et al., 2024), which is implemented using the official open-source code, and fine-tuned with the same training config for fair comparison.

**Experiment Results.** As shown in Fig. 10(a–b), *GigaBrain-0* consistently achieves the highest task success rate across both tasks, surpassing $\pi_0$ by 30% and 10% in each case. Its integration of depth sensing enhances spatial awareness, enabling precise coordination in contact-rich scenarios. Qualitatively, as illustrated in Fig. 11 and Fig. 12, *GigaBrain-0* robustly executes complex, dexterous workflows: for laundry, it performs synchronized dual-gripper grasping, accurate folding, and final alignment; for paper towels, it tears off excess material, rolls the towel with controlled tension, and precisely applies adhesive labels. These results highlight *GigaBrain-0*'s superior dexterous manipulation capability and generalization in real-world, unstructured settings.

## 5.2. Long-horizon Experiment

**Experiment Setup.** We evaluate *GigaBrain-0* on two long-horizon manipulation tasks: `table bussing` and `juice preparation`. For `table bussing`, we deploy the model on the dual-arm PiPER robot platform and
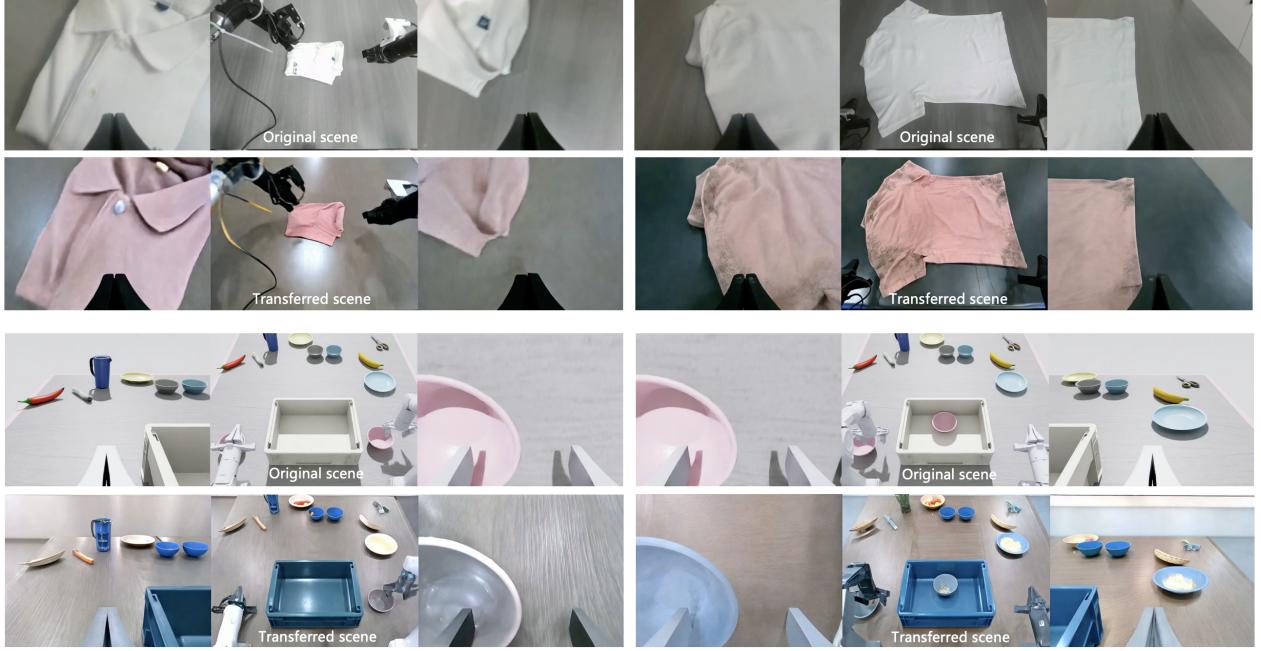
Figure 9: GigaWorld can generate multi-view consistent videos, thereby enabling 3D-aware training and improving spatial reasoning in downstream tasks.

fine-tune it using 100 human-collected demonstration trajectories, and the training is conducted with a batch size of 128 for 20K steps. For `juice preparation`, we utilize the AgiBot G1 robot with 489 demonstrations, and the training is conducted with a batch size of 128 for 35K steps. We compare *GigaBrain-0* against $\pi_0$ (Black et al., 2024), which is implemented using the official open-source code, and fine-tuned with the same training config to ensure a fair comparison.

**Experiment Results.** As shown in Fig. 10(c–d), *GigaBrain-0* consistently achieves the highest task success rate across both long-horizon tasks. Its integration of embodied Chain-of-Thought (CoT) reasoning enables fine-grained, temporally ordered planning—critical for complex multi-step execution. Qualitative results in Fig. 13 and Fig. 14 further demonstrate *GigaBrain-0*'s robustness: in `table bussing`, it first detaches bowls from plates, then sequentially places plates followed by bowls into the collection bin. In `juice preparation`, it scoops powdered mix with a spoon, dispenses water from the dispenser, and finally stirs the mixture to completion. These results highlight *GigaBrain-0*'s superior capacity for structured reasoning and reliable execution in long-horizon, real-world manipulation scenarios.

## 5.3. Mobile Manipulation Experiment

**Experiment Setup.** We evaluate *GigaBrain-0* on two mobile manipulation tasks: `boxes moving` and `laundry baskets moving`. For `boxes moving`, we deploy the model on the AgiBot G1 robot platform and fine-tune it using 300 human-collected demonstration trajectories, and the training is conducted with a batch size of 128 for 30K steps. For `laundry baskets moving`, we utilize the dual-arm PiPER robot platform with 378 demonstrations, and the training is conducted with a batch size of 192 for 30K steps. We compare *GigaBrain-0* against $\pi_0$ (Black et al., 2024), which is implemented using the official open-source code, and fine-tuned with the same training config to ensure a fair comparison.

**Experiment Results.** As shown in Fig. 10(e–f), *GigaBrain-0* achieves the highest success rate in both mobile manipulation tasks, surpassing $\pi_0$ by 10% in each task. Its unified architecture, combining global navigation priors with local manipulation policies, enables seamless transitions between mobility and interaction. In `boxes moving`, *GigaBrain-0* navigates around obstacles, identifies target boxes using semantic segmentation,
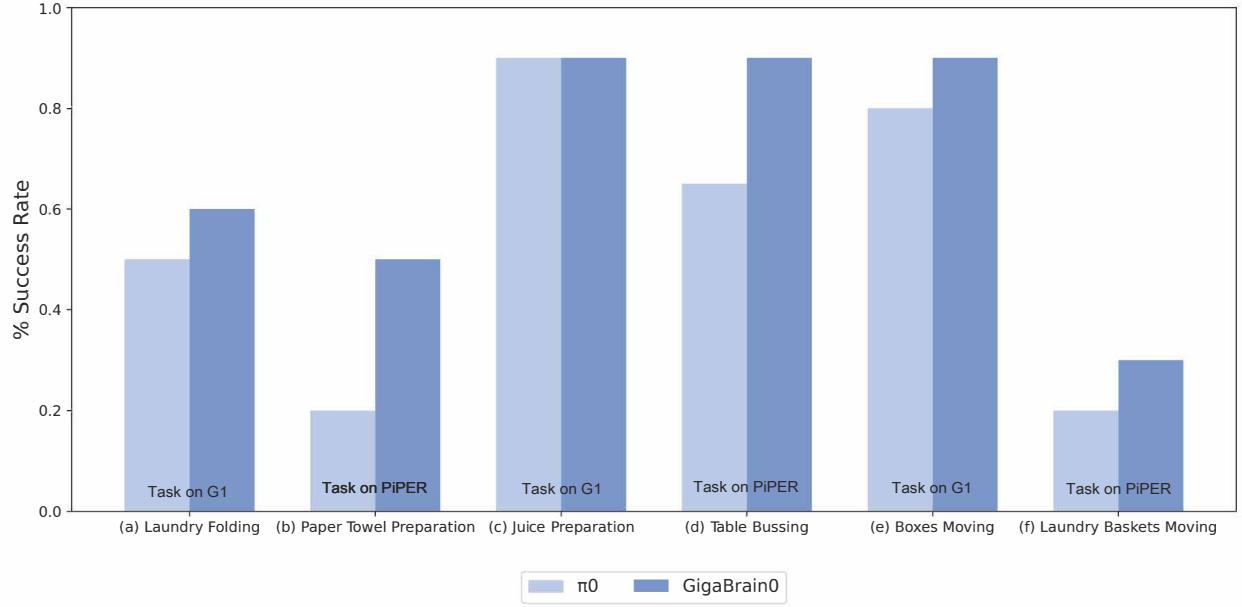
Figure 10: Performance comparison between *GigaBrain-0* and $\pi_0$ across six tasks on G1 and PiPER robot platforms, where (a) `Laundry Folding` and (b) `Paper Towel Preparation` are dexterous manipulation tasks; (c) `Juice Preparation` and (d) `Table Bussing` are long-horizon tasks; (e) `Boxes Moving` and (f) `Laundry Baskets Moving` are mobile manipulation tasks.
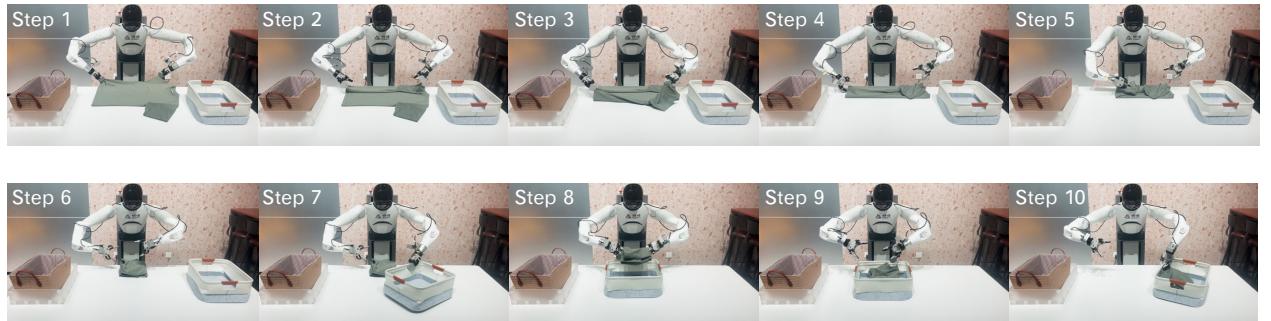


Figure 11: Deployment of *GigaBrain-0* on the G1 humanoid robot for real-world `laundry folding`.
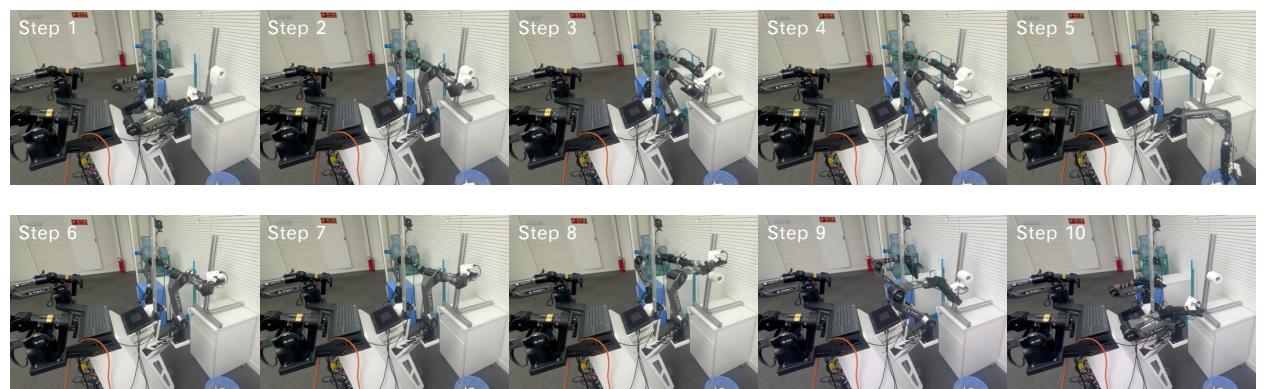


Figure 12: Deployment of *GigaBrain-0* on the PiPER arms for real-world `paper towel preparation`.
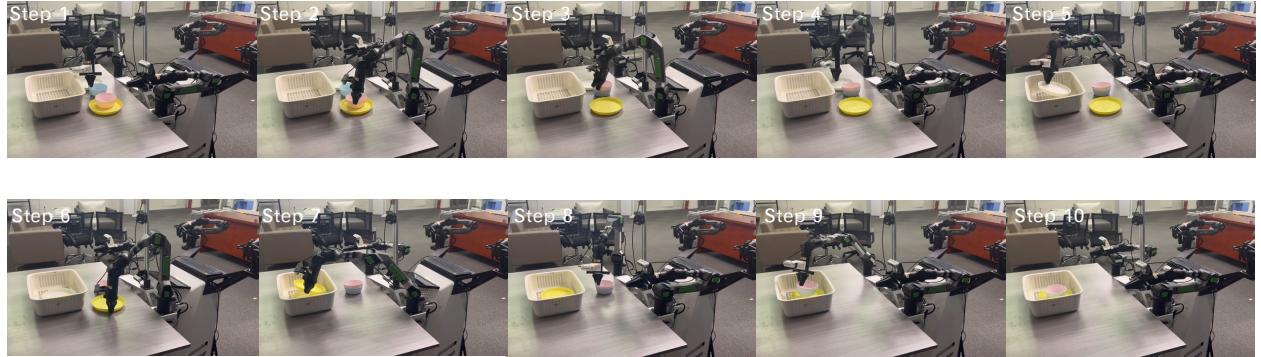
Figure 13: Deployment of *GigaBrain-0* on PiPER arms for real-world `table bussing`.
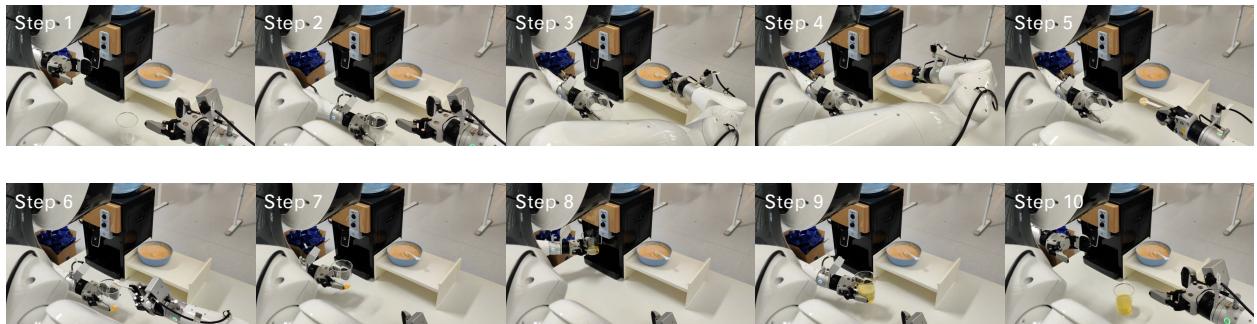


Figure 14: Deployment of *GigaBrain-0* on G1 humanoid robot for real-world `juice preparation`.



Figure 15: Deployment of *GigaBrain-0* on the G1 humanoid robot for real-world `paper towel preparation`.
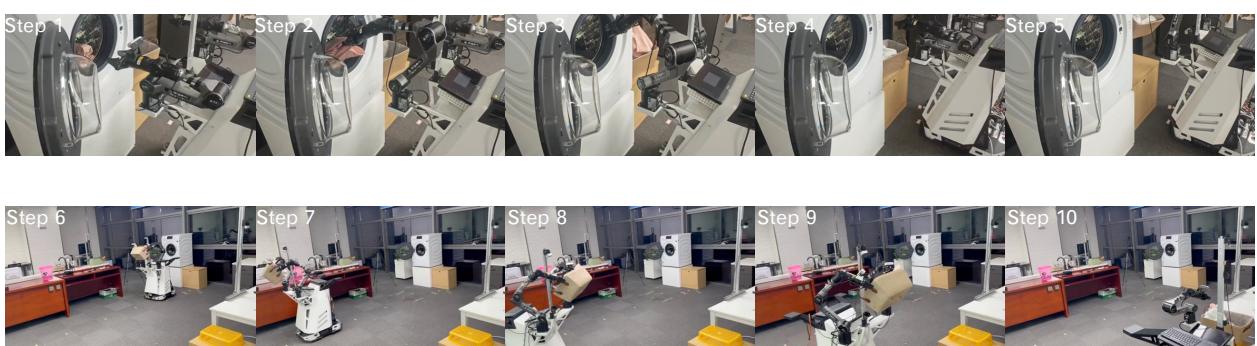


Figure 16: Deployment of *GigaBrain-0* on the PiPER arms for real-world `laundry baskets moving`.
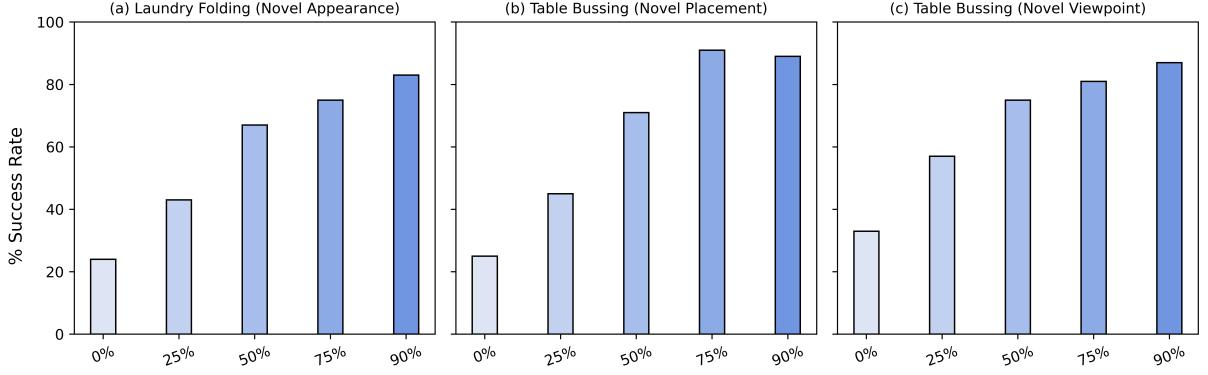
Figure 17: Generalization performance of *GigaBrain-0* under appearance, placement, and viewpoint shifts. The horizontal axis denotes the sampling probability $\alpha$ of world model-generated data used during training.

grasps them with adaptive force control, and delivers them to designated zones. In `laundry baskets moving`, it locates partially occluded baskets, adjusts its base position for optimal reach, lifts baskets using compliant gripper control, and transports them across uneven terrain without spilling contents. Qualitative results in Fig. 15 and Fig. 16 further illustrate *GigaBrain-0*'s robustness to environmental variability and its ability to recover from minor localization or grasping errors through real-time replanning.

## 5.4. Manipulation Generalization Experiment

Diverse training data is crucial for the generalization of VLA models (Shi et al., 2025). *GigaBrain-0* leverages GigaWorld to generate diverse training data, which significantly enhances its robustness and generalization under variations in appearance, object placement, and camera viewpoint. Below, we present our experimental setup and analysis in detail.



Figure 18: Illustration of the appearance generalization experiment on `laundry folding`: one white garment and nine items with varied colors and textures.

**Appearance Generalization.** We conduct appearance generalization experiments on the `laundry folding` task. Specifically, we perform post-training on *GigaBrain-0* using a small set of real-world folding demonstrations, limited to only 50 trajectories of white garments collected on the physical robot. To enhance visual diversity, we combine this real data with trajectories generated by the GigaWorld Real2Real transfer model, which effectively transfers appearance attributes while preserving geometric and dynamic consistency (see Fig. 4 for illustration). We then evaluate how the mixing ratio $\alpha$ (i.e., the probability of sampling generated data in each training batch) affects real-world success rates. The model is fine-tuned for 20k steps with a batch size of 128. For evaluation, we test on 10 distinct garments with diverse appearances, including the original white garments and 9 additional items with varied colors and textures (examples shown in Fig. 18). Each garment is folded five times, and success is measured per trial. As shown in Fig. 17(a), for a complex deformable manipulation

task like garment folding, training exclusively on white garments leads to poor generalization to real-world garments with diverse textures and colors. However, incorporating appearance-transferred data significantly improves texture generalization. Notably, as the sampling probability $\alpha$ of generated data increases from 0 to 50%, the success rate rises substantially, reaching nearly 70%. Further increasing $\alpha$ to 75% and 90% yields additional gains, pushing the success rate above 80%.



Figure 19: Illustration of the object placement generalization experiment on `table bussing`: one collected placement and nine novel object placements.

**Placement Generalization.** We investigate generalization to novel object placements in the table bussing task. Our approach begins with post-training *GigaBrain-0* on a small set of real-world demonstrations, only 50 trajectories collected under a single, fixed bowl configuration. To broaden the distribution of object positions during training, we supplement this limited real data with trajectories produced by the GigaWorld Sim2Real transfer model. This model not only introduces a wide variety of object placements but also enhances visual realism by adapting simulated appearances to better match real-world conditions (see Fig. 6). We study the impact of the mixing ratio $\alpha$ on downstream performance. The policy is fine-tuned for 20k steps with a batch size of 128. During evaluation, we assess performance across 10 distinct placement layouts: the original fixed setup with 9 new configurations with diverse object placements (examples in Fig. 19). Each layout is tested five times, and success is recorded per trial. Results in Fig. 17(b) reveal that, for this long-horizon bussing task, models trained solely on fixed-placement data struggle to handle real-world variations in object layout. In contrast, blending in Sim2Real transferred data dramatically boosts generalization to unseen placements. As $\alpha$ increases from 0 to 75%, the success rate climbs sharply, surpassing 90%. This gain arises because the simulation provides abundant novel placement scenarios that teach the policy manipulation behaviors, while the Sim2Real transfer simultaneously mitigates the visual domain gap between simulation and real images.

**Viewpoint Generalization.** We investigate generalization to novel camera viewpoints in the table bussing task. Begining by post-training *GigaBrain-0* on a small set of 50 real-world demonstrations, all captured from a single, fixed camera viewpoint. To enrich viewpoint diversity during training, we augment this data with trajectories generated by the GigaWorld view transfer model, which effectively re-renders the originally collected demonstrations from a broad range of virtual camera angles while preserving task-relevant robot actions (see Fig. 5). We analyze how the mixing ratio $\alpha$ influences real-world performance. The policy is fine-tuned for 20k steps with a batch size of 128. For evaluation, we test on 9 distinct camera viewpoints: the original fixed view plus 8 additional, previously unseen viewpoints with varied positions and rotations (examples shown in Fig. 20). Each viewpoint is evaluated over five independent trials, with success judged per trial. As shown in Fig. 17(c), when trained only on data from a single viewpoint, the policy exhibits significant performance degradation under novel viewing conditions. In contrast, incorporating view-transferred data substantially improves robustness to viewpoint shifts. As $\alpha$ increases from 0 to 90%, the success rate rises sharply, exceeding 80%. This improvement is attributed to two factors: (1) exposure to multi-view observations enables the policy to learn viewpoint-invariant visual representations, and (2) the GigaWorld view transfer model provides photorealistic, geometrically consistent renderings that closely approximate real-world visual input from arbitrary angles.
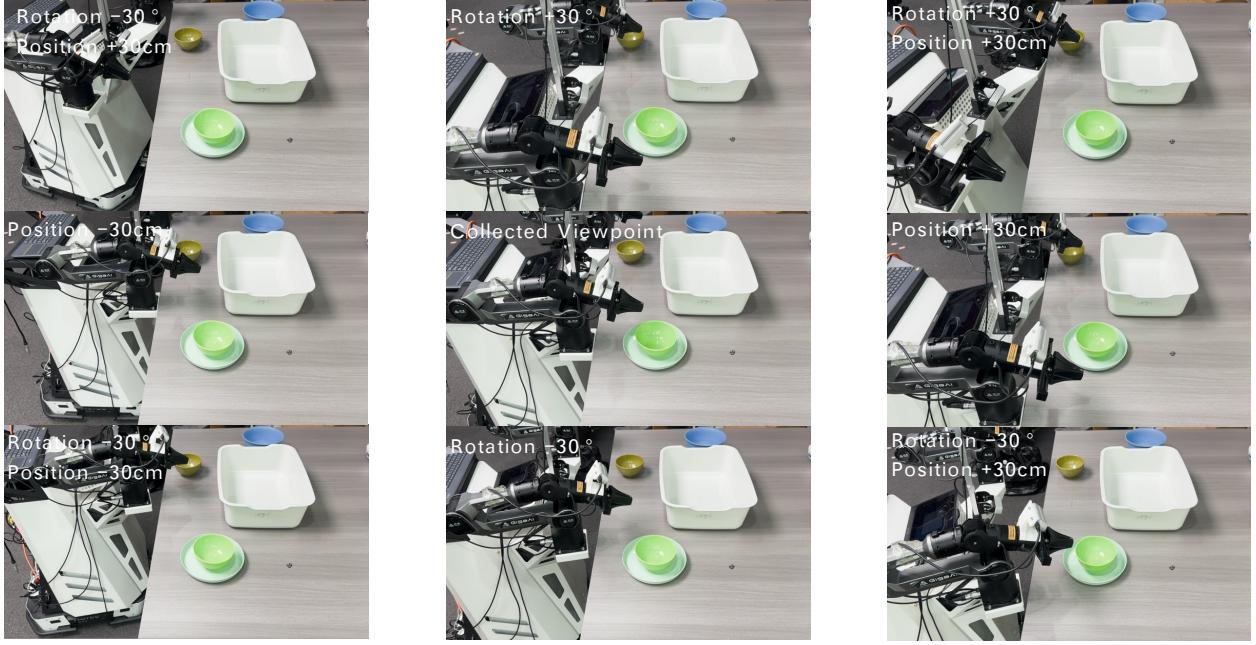
Figure 20: Illustration of the camera viewpoint generalization experiment on `table bussing`: one collected viewpoint and eight novel camera viewpoints.


## 5.5. On-device Experiment

Recent VLA models (Bjorck et al., 2025; Black et al., 2024; Cheang et al., 2025; Intelligence et al., 2025; Jiang et al., 2025; Zhai et al., 2025), while powerful, are often hindered by their large parameter counts and high computational complexity, making them impractical for deployment on resource-constrained edge devices. However, efficient on-device execution is crucial for real-time robotic applications that require low latency, privacy, and autonomy. To address this challenge, we present *GigaBrain-0-Small*, an optimized lightweight variant specifically designed for efficient inference on edge platforms such as the NVIDIA Jetson AGX Orin.

Compared to *GigaBrain-0*, *GigaBrain-0-Small* adopts the compact vision-language model SmolVLM2 (Marafioti et al., 2025) and reduces the action expert parameters to approximately 100M. Beyond architectural simplification, we implement a series of system-level optimizations: (1) we eliminate redundant CPU–GPU memory transfers and unnecessary data type (dtype) conversions; (2) we enable automatic mixed-precision inference via `torch.autocast`; (3) we optimize the Rotary Position Embedding (RoPE) (Su et al., 2024) computation by precomputing and caching the sine and cosine lookup tables; and (4) we apply `torch.compile` to key components, including the denoising step and VLM forward pass, to convert dynamic PyTorch code into optimized static graphs.

These optimizations collectively enable *GigaBrain-0-Small* to achieve significantly lower latency and memory footprint compared to $\pi_0$ model on the Orin platform, as summarized in Tab. 2. We collected 1K episodes of `table bussing` data using the G1 humanoid robot and fine-tuned both models. Despite having 12.5% parameters, *GigaBrain-0-Small* achieves a comparable success rate to $\pi_0$.

Table 2: Model comparison between *GigaBrain-0-Small* and $\pi_0$ on the Orin platform.

| Model | FLOPs (GFLOPs) | Parameters (B/M) | Inference VRAM (GB) | Inference Latency (seconds) | Success Rate |
|---|---|---|---|---|---|
| $\pi_0$ | 4400 | 3.2 B | 17.5 | 1.28 | 80% |
| *GigaBrain-0-Small* | 840 | 402 M | 1.9 | 0.13 | 80% |

# 6. Conclusion and Future Work

In this work, we presented *GigaBrain-0*, a vision-language-action model that leverages data generated by world models to overcome the scalability and diversity limitations of real-world robot data collection. By training on rich, photorealistic trajectories that span diverse scene appearances, object placements, and camera viewpoints, *GigaBrain-0* achieves strong generalization across a wide spectrum of real-world robotic tasks, from dexterous manipulation to long-horizon mobile operations. Key architectural innovations, including RGBD input modeling and embodied Chain-of-Thought supervision, further enhance its spatial reasoning and sequential decision-making capabilities. Moreover, we introduced *GigaBrain-0-Small*, a lightweight variant optimized for edge deployment on platforms such as the NVIDIA Jetson AGX Orin, demonstrating that VLA model can be made practical for real-time, on-device robotic control.

Looking ahead, our work opens several promising directions for future research. First, while we currently employ world models as scalable *data engines*, a natural next step is to integrate them as interactive *policy environments* for reinforcement learning. By enabling VLA agents to roll out trajectories and receive reward signals within the world model, we could drastically reduce the need for real-world trial-and-error while supporting policy refinement through simulated experience. Second, world models may learn universal representations of physical dynamics and task structure. Such representations could allow world models to evolve beyond passive simulators into active *policy generators*, capable of proposing feasible action sequences or subgoals directly. Finally, closing the loop between VLA policies and world models through self-improvement cycles—where real-world rollouts continuously refine the world model, which in turn generates better training data, could pave the way toward truly autonomous, lifelong-learning robotic systems.

# References

[1] 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. 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 2

[3] Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025. 3, 9

[4] 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. 3

[5] Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025. 10

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6

[7] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 2

[8] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 2, 3, 5, 17

[9] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi_0$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 2, 3, 5, 11, 12, 17

[10] 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. 2

[11] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Xindong He, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *IROS*, 2025. 5

[12] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025. 8

[13] Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, et al. Gr-3 technical report. *arXiv preprint arXiv:2507.15493*, 2025. 2, 3, 5, 8, 17

[14] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *CVPR*, 2025. 7

[15] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019. 3

[16] Zhehao Dong, Xiaofeng Wang, Zheng Zhu, Yirui Wang, Yang Wang, Yukun Zhou, Boyuan Wang, Chaojun Ni, Runqi Ouyang, Wenkang Qin, et al. Emma: Generalizing real-world robot manipulation via generative visual transfer. *arXiv preprint arXiv:2509.22407*, 2025. 3, 7, 8, 9

[17] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. *arXiv preprint arXiv:2408.11812*, 2024. 2

[18] Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *arXiv preprint arXiv:2505.23705*, 2025. 3, 4, 5

[19] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *NeurIPS*, 2023. 3

[20] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021. 3

[21] Yu Fang, Yue Yang, Xinghao Zhu, Kaiyuan Zheng, Gedas Bertasius, Daniel Szafir, and Mingyu Ding. Rebot: Scaling robot learning with real-to-sim-to-real robotic video synthesis. *arXiv preprint arXiv:2503.14526*, 2025. 3

[22] Yao Feng, Hengkai Tan, Xinyi Mao, Guodong Liu, Shuhe Huang, Chendong Xiang, Hang Su, and Jun Zhu. Generalist bimanual manipulation via foundation video diffusion models. *arXiv preprint arXiv:2507.12898*, 2025. 3

[23] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 3

[24] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024. 3

[25] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *CVPR*, 2023. 9

[26] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 8

[27] 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. 3

[28] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 2, 3, 5, 17

[29] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *RAL*, 2020. 6

[30] 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 video world models. *arXiv preprint arXiv:2505.12705*, 2025. 3, 5, 9

[31] Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025. 2, 5, 17

[32] 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. 3

[33] Zhao Jin, Zhengping Che, Zhen Zhao, Kun Wu, Yuheng Zhang, Yinuo Zhao, Zehui Liu, Qiang Zhang, Xiaozhu Ju, Jing Tian, et al. Artvip: Articulated digital assets of visual realism, modular interaction, and physical fidelity for robot learning. *arXiv preprint arXiv:2506.04941*, 2025. 8

[34] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023. 2

[35] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *ICRA*, 2025. 8

[36] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 3

[37] 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. 2

[38] 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. 3, 9

[39] Haoyun Li, Ivan Zhang, Runqi Ouyang, Xiaofeng Wang, Zheng Zhu, Zhiqin Yang, Zhentao Zhang, Boyuan Wang, Chaojun Ni, Wenkang Qin, et al. Mimicdreamer: Aligning human and robot demonstrations for scalable vla training. *arXiv preprint arXiv:2509.22199*, 2025. 3, 9

[40] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024. 2

[41] Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024. 3

[42] 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

[43] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 2

[45] Liu Liu, Xiaofeng Wang, Guosheng Zhao, Keyu Li, Wenkang Qin, Jiaxiong Qiu, Zheng Zhu, Guan Huang, and Zhizhong Su. Robotransfer: Geometry-consistent video diffusion for robotic visual policy transfer. *arXiv preprint arXiv:2505.23171*, 2025. 3, 7, 8, 9, 10

[46] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022. 3

[47] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 2

[48] Zeyi Liu, Shuang Li, Eric Cousineau, Siyuan Feng, Benjamin Burchfiel, and Shuran Song. Geometry-aware 4d video generation for robot manipulation. *arXiv preprint arXiv:2507.01099*, 2025. 3

[49] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 17

[50] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 2

[51] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. *arXiv preprint arXiv:2411.19548*, 2024. 3

[52] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Xinze Chen, Guanghong Jia, Guan Huang, and Wenjun Mei. Recondreamer-rl: Enhancing reinforcement learning via diffusion-based scene reconstruction. *arXiv preprint arXiv:2508.08170*, 2025. 3

[53] NVIDIA. Isaac Sim. URL https://github.com/isaac-sim/IsaacSim. 8

[54] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *ICRA*, 2024. 2, 3

[55] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *ICRA*, 2024. 5

[56] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 3

[57] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2

[58] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025. 2, 3, 4

[59] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025. 2

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[61] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 8

[62] Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025. 3

[63] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025. 3

[64] Modi Shi, Li Chen, Jin Chen, Yuxiang Lu, Chiming Liu, Guanghui Ren, Ping Luo, Di Huang, Maoqing Yao, and Hongyang Li. Is diversity all you need for scalable robotic manipulation? *arXiv preprint arXiv:2507.06219*, 2025. 5, 15

[65] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024. 3

[66] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 17

[67] Octo Model Team, Dibya Ghosh, Homer Walke, 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. 2

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

[69] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 9

[70] Boyuan Wang, Xinpan Meng, Xiaofeng Wang, Zheng Zhu, Angen Ye, Yang Wang, Zhiqin Yang, Chaojun Ni, Guan Huang, and Xingang Wang. Embodiedreamer: Advancing real2sim2real transfer for policy training via embodied world modeling. *arXiv preprint arXiv:2507.05198*, 2025. 3, 7

[71] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023. 8

[72] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *NeurIPS*, 2024. 2

[73] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR*, 2025. 5, 7

[74] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *ECCV*, 2024. 3

[75] Xinjie Wang, Liu Liu, Yu Cao, Ruiqi Wu, Wenkang Qin, Dehui Wang, Wei Sui, and Zhizhong Su. Embodiedgen: Towards a generative 3d world engine for embodied intelligence. *arXiv preprint arXiv:2506.10600*, 2025. 8

[76] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022. 4

[77] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024. 5

[78] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022. 2

[79] Youguang Xing, Xu Luo, Junlin Xie, Lianli Gao, Hengtao Shen, and Jingkuan Song. Shortcut learning in generalist robot policies: The role of dataset diversity and fragmentation. *arXiv preprint arXiv:2508.06426*, 2025. 7

[80] Yuan Xu, Jiabing Yang, Xiaofeng Wang, Yixiang Chen, Zheng Zhu, Bowen Fang, Guan Huang, Xinze Chen, Yun Ye, Qiang Zhang, et al. Egodemogen: Novel egocentric demonstration generation enables viewpoint-robust manipulation. *arXiv preprint arXiv:2509.22578*, 2025. 3, 7

[81] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 3

[82] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025. 8

[83] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *NeurIPS*, 2024. 10

[84] Chengbo Yuan, Suraj Joshi, Shaoting Zhu, Hang Su, Hang Zhao, and Yang Gao. Roboengine: Plug-and-play robot data augmentation with semantic robot segmentation and background generation. *arXiv preprint arXiv:2505.18738*, 2025. 3

[85] Andy Zhai, Brae Liu, Bruno Fang, Chalse Cai, Ellie Ma, Ethan Yin, Hao Wang, Hugo Zhou, James Wang, Lights Shi, et al. Igniting vlms toward the embodied space. *arXiv preprint arXiv:2509.11766*, 2025. 2, 5, 17

[86] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 3

[87] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 7

[88] Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, et al. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. *arXiv preprint arXiv:2410.13571*, 2024. 3

[89] Guosheng Zhao, Xiaofeng Wang, Chaojun Ni, Zheng Zhu, Wenkang Qin, Guan Huang, and Xingang Wang. Recondreamer++: Harmonizing generative and reconstructive models for driving scene representation. *arXiv preprint arXiv:2503.18438*, 2025. 3

[90] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *AAAI*, 2025. 3, 9

[91] 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. 3

[92] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024. 3

[93] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024. 3