InternVLA-A1: Unifying Understanding, Generation and Action for Robotic Manipulation

InternVLA-A1 Team

Full author list in Contributors section

Prevalent Vision-Language-Action (VLA) models are typically built upon Multimodal Large Language Models (MLLMs) and demonstrate exceptional proficiency in semantic understanding, but they inherently lack the capability to deduce physical world dynamics. Consequently, recent approaches have shifted toward World Models, typically formulated via video prediction; however, these methods often suffer from a lack of semantic grounding and exhibit brittleness in the presence of video prediction errors. To synergize semantic understanding with dynamic predictive capabilities, we present InternVLA-A1. This model employs a unified Mixture-of-Transformers architecture, coordinating three experts for scene understanding, visual foresight generation, and action execution. These components interact seamlessly through a unified masked self-attention mechanism. Building upon InternVL3 and Qwen3-VL, we instantiate InternVLA-A1 at 2B and 3B parameter scales. We pre-train these models on heterogeneous data sources over real-world robot data, synthetic simulation data, and human videos, covering over 692M frames. This hybrid training strategy effectively harnesses the diversity of synthetic simulation data while minimizing the sim-to-real gap. We evaluated InternVLA-A1 on 12 real-world robotic tasks and a simulation benchmark. The results show that InternVLA-A1 consistently outperforms prior leading models: compared with $\pi_{0.5}$, it achieves +4.4% on static manipulation tasks and +2.6% on the RoboTwin 2.0 simulation benchmark, and delivers a +26.7% boost on dynamic manipulation tasks.

Homepage | Code:InternVLA-A1 | Model:InternVLA-A1 | Data:InternData-A1

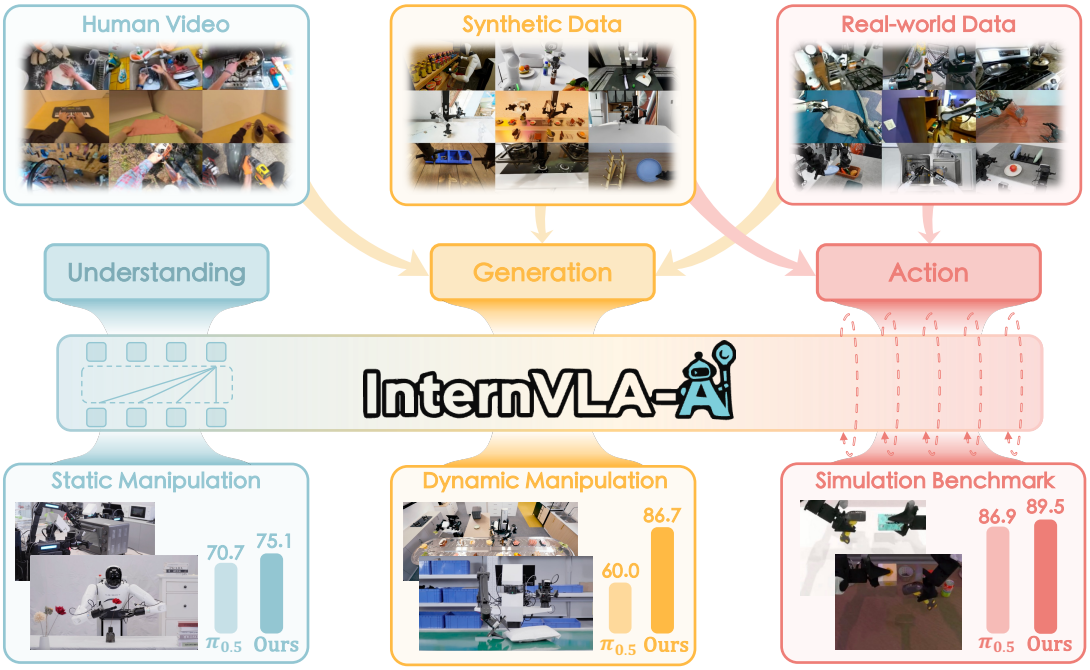

Figure 1. **InternVLA-A1** unifies scene understanding, visual foresight generation, and action execution into a single framework. This architecture couples semantic reasoning with dynamics prediction to guide action execution, and effectively enables joint training on heterogeneous data sources over human videos, synthetic data, and real-world demonstrations. The resulting model exhibits consistent robustness across static manipulation, dynamic manipulation, and simulation benchmarks, especially demonstrating remarkable superiority in dynamic scenarios.

1. Introduction

The pursuit of intelligent generalist robots remains a cornerstone of robotics research (Bu et al., 2024a; Clark et al., 2025; Cui et al., 2025; Fang et al., 2023; Huang et al., 2025b, 2023; Li et al., 2025b; Qu et al., 2025). Currently, the community favors the end-to-end learning paradigm and the Vision-Language-Action (VLA) architecture (Bjorck et al., 2025; Black et al., 2024, 2025; Cen et al., 2025; Cheang et al., 2025; Chen et al., 2025d; Li et al., 2025a; Lin et al., 2025; NVIDIA, 2025; Yang et al., 2025a,b; Zhai et al., 2025; Zhao et al., 2025; Zheng et al., 2025) to realize such generalist policies. Built upon Multimodal Large Language Models (MLLMs) and trained on massive real-world robot demonstrations, VLA models exhibit remarkable performance in daily tasks such as clothes folding and table bussing. Nevertheless, the generalizability of these models still falls short of practical application requirements. Their adaptability to scene variations remains inadequate, such as dynamic settings involving industrial conveyor belts.

The generalization bottlenecks in existing policies stem from two primary issues: deficient physical world cognition and a lack of adaptive manipulation capabilities. Addressing the first challenge requires integrating foundation models, such as Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Beyer et al., 2024; Chen et al., 2025a; Touvron et al., 2023) or world models (Assran et al., 2025; Blattmann et al., 2023b; HaCohen et al., 2024; Zheng et al., 2024), to enhance cognitive capacity, while addressing the latter necessitates large-scale and diverse robot action datasets for skill learning. Consequently, achieving generalist policies requires a synergistic strategy that advances both model architecture and training data.

Regarding *model architecture*, prevalent generalist policies, such as π_0 (Black et al., 2024), $\pi_{0.5}$ (Black et al., 2025), and GR00T N1/N1.5 (Bjorck et al., 2025), are built upon MLLMs that map visual data into a text-based feature space. Although this grants them strong semantic understanding, text tokens are ill-suited for modeling physical laws, resulting in a deficiency in physical dynamics reasoning. Consequently, these policies are optimized for reactive perception-to-action mapping, rather than reasoning about how states will evolve under motion and contact. This limitation becomes particularly evident in dynamic environments, such as industrial conveyor settings, where understanding momentum, inertia, and contact dynamics is critical. Recent efforts attempt to incorporate foresight via World Models, notably through video prediction paradigms like VPP (Hu et al., 2025) and Genie Envisioner (Liao et al., 2025). These methods generate anticipated observations to guide decision-making, yet they often suffer from weak semantic grounding and are sensitive to video prediction errors. Consequently, developing a unified architecture that tightly couples semantic understanding with robust predictive dynamics is crucial for reliable dynamics-aware manipulation.

Regarding *training data*, existing VLA models have gained adaptability by scaling up large real-robot datasets, and current generalist policies rely heavily on such real-robot data (Bu et al., 2025; Walke et al., 2023; Wu et al., 2024b). For example, pioneering works collected 130K episodes to train RT-1 (Brohan et al., 2022) and RT-2 (Brohan et al., 2023), and subsequent initiatives aggregated over one million demonstrations from 22 heterogeneous robots to build Open X-Embodiment (O'Neill et al., 2024). However, relying solely on real-world data collection remains challenging. Although π_0 was trained with 10,000 hours demonstrations, spanning 68 tasks from seven robot morphologies and achieving strong dexterous manipulation, it still struggles to adapt to scene variations, especially in highly dynamic environments. Further expanding real-robot datasets and covering long-tail scene variations at scale is costly and inefficient. In contrast, simulation presents a promising complementary approach. Its extensive libraries of scenes and objects enrich sample diversity, and domain randomization simulates scene variations to improve the policy's robustness in changing environments. Additionally, simulation ensures more controllable trajectories by eliminating noisy data. Our prior work, InternData-A1 (Tian et al., 2025b), validated that large-scale and high-fidelity

simulation data can effectively support the pre-training of VLA models. Nevertheless, simulation data suffers from the inevitable sim-to-real gap, especially in contact-rich dynamics. Therefore, synergizing the diversity of simulation with the physical fidelity of real-world data presents a promising avenue to overcome these respective limitations.

To address the above generalization challenges, particularly robustness against dynamic scene variations, we propose InternVLA-A1. As depicted in Figure 1, our model features a novel architecture integrating understanding, generation, and action. InternVLA-A1 combines the semantic reasoning of MLLMs with the prediction capability of a World Model-style imagination module, effectively bridging the gap between semantics and physical dynamics to facilitate foresight-aware action generation. Furthermore, we train InternVLA-A1 on three complementary data sources: large-scale simulated robot trajectories, real-robot demonstrations, and egocentric human videos. This joint training recipe enables scalable diversity from simulation, grounds action execution with real-robot interactions to reduce the sim-to-real gap, and enriches visual representations from human videos to better capture complex manipulations. By combining these three sources in a unified training pipeline, InternVLA-A1 benefits from scalable variation without sacrificing physical fidelity, leading to improved robustness and real-world transfer. We validate InternVLA-A1 through extensive experiments on 12 real-world tasks and a simulation benchmarks, and observe consistent improvements over strong VLA baselines. Specifically, InternVLA-A1 surpasses $\pi_{0.5}$ with a +4.4% improvement on real-world static tasks and +2.6% gain on the RoboTwin 2.0 simulation benchmark, alongside a substantial +26.7% boost on real-world dynamic tasks.

2. Related Works

In this section, we compare InternVLA-A1 with current methods from the perspectives of model architecture and training data.

In **Vision-Language-Action models**, a common practice is to integrate the multimodal capabilities of foundation models with robotic control. RT-2 (Brohan et al., 2023) and OpenVLA (Kim et al., 2024) use a vocabulary-replacement technique that maps text tokens in an LLM to discrete action representations, thereby leveraging the LLM's general-purpose capabilities to facilitate emergent reasoning. Diverging from discrete token-based approaches, π_0 (Black et al., 2024) adopts a Mixture-of-Transformers architecture that combines a pre-trained MLLM with an action expert, and uses flow matching to output continuous actions, enabling more precise and dexterous control. Building on this, $\pi_{0.5}$ (Black et al., 2025) combines high-level sub-task prediction with low-level action prediction, thereby enhancing its capability for long-horizon tasks and cross-scene generalization. In addition, GR00T N1/N1.5 (Bjorck et al., 2025) adopts an approach that is closer to a dual-system paradigm, pairing a VLM for high-level reasoning with a DiT for action generation. Distinct from these approaches, InternVLA-A1 unifies the semantic understanding of MLLMs with the dynamics-prediction capabilities, effectively bridging the semantics-dynamics gap prevalent in existing VLA architectures.

In **Video prediction and world models**, extensive research leverages future video prediction to facilitate robotic control (Hung et al., 2025; Zhang et al., 2025; Zhu et al., 2025). UniPi (Du et al., 2023) trains a text-conditioned video generator paired with an inverse dynamics model to derive actions, whereas UniSim (Yang et al., 2024) employs generative modeling to construct a universal simulator for training both high-level and low-level policies. GR-1 (Wu et al., 2024a) and GR-2 (Cheang et al., 2024) use future image prediction as an auxiliary task to enhance the policy's visual representations. CLOVER (Bu et al., 2024b), Seer (Tian et al., 2025a), and $\mathcal{F}_1$ (Lv et al., 2025) leverage future image prediction to guide action generation in an inverse-dynamics-like manner. As video generation models have advanced, it has become increasingly popular to incorporate a

pretrained video generation model to guide action execution, as in VPP (Hu et al., 2025) and Genie Envisioner (Liao et al., 2025). Despite these advancements, these policies are often sensitive to video generation quality and lack the semantic reasoning capabilities inherent in MLLMs. In contrast, InternVLA-A1 integrates the MLLM with future visual state prediction, thereby providing semantically grounded guidance and enabling stronger adaptability in the presence of video prediction errors.

In **training data**, real-world robot data, simulated synthetic data, and human video data are widely used. Real-world datasets (Bu et al., 2025; Khazatsky et al., 2024; Walke et al., 2023; Wu et al., 2024b) capture realistic physical dynamics essential for end-to-end learning. RT-1 (Brohan et al., 2022) and RT-2 (Brohan et al., 2023) were trained on approximately 130k real-robot demonstrations. And the Open X-Embodiment dataset (O'Neill et al., 2024) aggregates one million episodes from multiple sources. π_0 (Black et al., 2024) further scales the amount of real-robot training data to 10,000 hours. However, further expanding real-robot datasets by another order of magnitude remains prohibitively expensive and operationally inefficient. Simulation-based synthetic data (Chen et al., 2025c; Gao et al., 2025; Gu et al., 2023; Hua et al., 2024; Huang et al., 2025a; James et al., 2020; Lian et al., 2025; Mandlekar et al., 2023; Mees et al., 2022) can fully exploit rich scene and object assets, and use domain randomization to enrich sample diversity and cover long-tail scene variations. And importantly, It provides a low-cost, efficient route for data collection. Robocasa (Nasiriany et al., 2024) collects data by teleoperating robots in simulation and then performing sample augmentation. Its initial data collection still requires manual effort. GraspVLA (Deng et al., 2025b) can automatically generate robot trajectories, but its skills are limited to grasping. More recently, InternData-A1 (Tian et al., 2025b) proposes an automated robot trajectory synthesis pipeline that covers the manipulation of rigid, articulated, deformable, and fluid objects, and constructs a large-scale simulated demonstration dataset with over 630k trajectories and 7,433 hours of data spanning diverse embodiments, skills, tasks, and scenes. There is a vast amount of video data on the internet that remains underexplored for its potential to benefit the pretraining of manipulation policies. Human videos (Damen et al., 2020; Goyal et al., 2017; Grauman et al., 2024; Hoque et al., 2025) provide rich visual priors without costly teleoperation, proving effective for human-to-robot transfer (Bi et al., 2025b; Kareer et al., 2025). For instance, Ego4D (Grauman et al., 2024) contains 3,670 hours of daily activities. Tailored for manipulation, EgoDex (Hoque et al., 2025) curates 829 hours of egocentric dexterous manipulation footage with paired 3D hand and finger tracking. Some studies (Bi et al., 2025a; Cheang et al., 2024) have incorporated human video data to enrich the diversity of pre-training datasets. In this work, we formulate a data recipe mixing real-world datasets, simulated data, and human videos, demonstrating that this combination enhances data diversity and effectively closes the sim-to-real gap.

3. InternVLA-A1: Unified Understanding-Generation-Action Framework

This section presents the design of InternVLA-A1. We first present an overview of the architecture, followed by a description of its components, optimization objectives, and implementation details.

3.1. Architecture Overview

The Mixture-of-Transformers (MoT) architecture, recently widely adopted in unified multimodal large language models (Deng et al., 2025a), demonstrates strong performance across both understanding and generation tasks. Drawing inspiration from these unified paradigms, InternVLA-A1 adopts the MoT architecture to seamlessly integrate scene understanding, visual foresight, and action execution within a single framework.

As illustrated in Figure 2, InternVLA-A1 employs three experts in a unified pipeline. The understanding expert first processes multimodal inputs to capture the environmental context. These

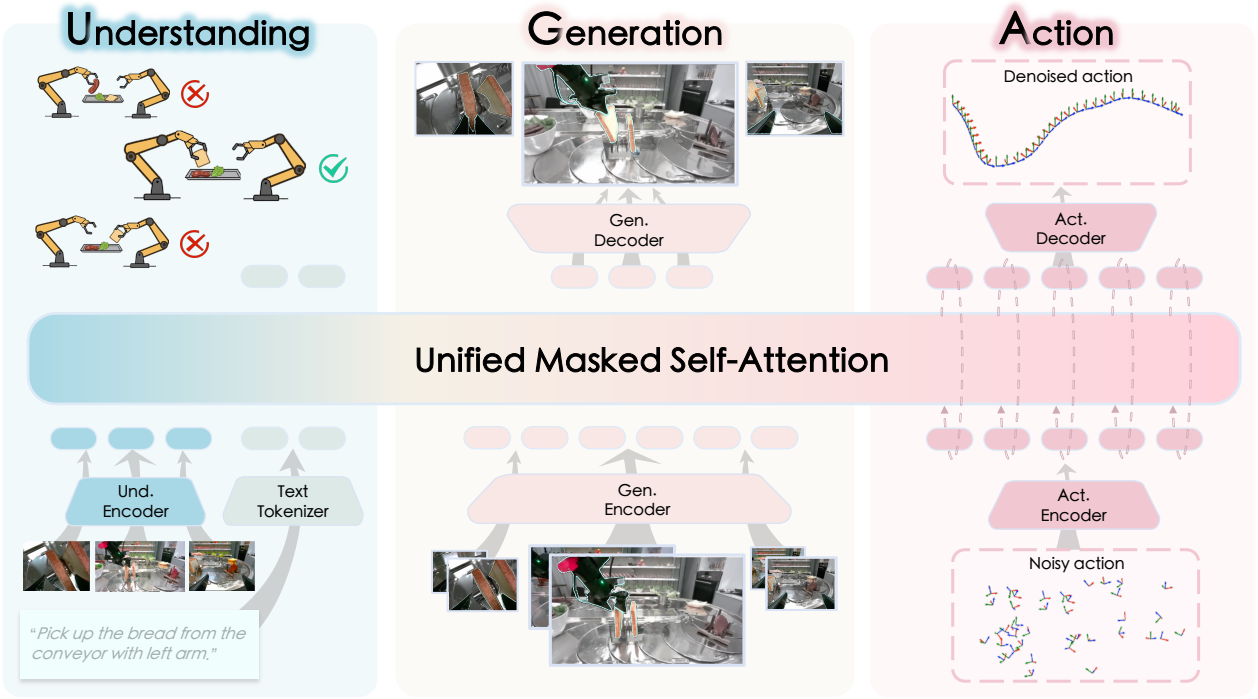

Figure 2. **Framework of InternVLA-A1**. The architecture comprises three experts: (1) an **understanding expert** that encodes scene context from image and text inputs; (2) a **generation expert** that predicts future visual states and task dynamics; and (3) an **action expert** that integrates the encoded scene context with these predictive dynamics to synthesize control commands via Flow Matching. This tripartite design enables adaptive manipulation under scene variations.

representations then inform the generation expert, which simulates the task's evolution by predicting future visual states. Finally, the action expert combines these predictive dynamics with the semantic context, utilizing Flow Matching to produce precise robot control commands.

3.2. Core Components

Inspired by the success of multimodal large language models, all three experts in InternVLA-A1 adopt a decoder-only transformer architecture.

Understanding Expert. The understanding expert directly adopts the architecture of existing MLLMs. In this implementation, we employ InternVL3 or Qwen3-VL, distinguished by their native multimodal capabilities and strong alignment between language and vision. We adhere to the processing pipeline of the base MLLMs: the multi-view observation at time t, denoted o_t, is encoded into visual tokens via the integrated vision encoder, while language instructions l are converted into text tokens using the text tokenizer. These visual and text tokens are subsequently processed by the transformer blocks of understanding expert, to form contextual embeddings $h_{\text{und}} = f_{\text{und}}(l, o_t)$. These embeddings serve as a shared context memory that is made accessible to downstream experts through masked self-attention, enabling the generation and action experts to attend to semantic scene context when predicting future latents and control commands.

Generation Expert. While video generation models have seen substantial progress, applying them to manipulation policies presents a significant challenge due to the requirements for high-frequency real-time inference. Mainstream image or video generation architectures, whether based on diffu-

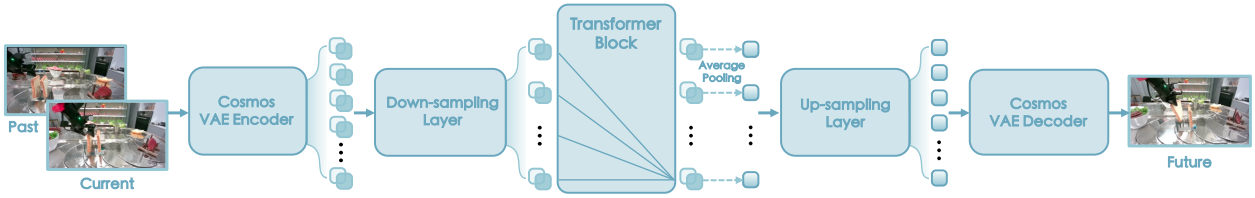

Figure 3. **The internal architecture specifics of the Generation Expert**.

sion (Blattmann et al., 2023a; Rombach et al., 2022) or next-token prediction (Sun et al., 2024), are typically too computationally intensive for end-to-end control. Even optimized solutions such as SANA-Sprint (Chen et al., 2025b) require 0.16 seconds per generation on one RTX 4090 GPU, restricting control frequencies to no more than 6Hz. Recent attempts (Bi et al., 2025a; GEAR, 2026) leverage large-scale pretrained video foundation models and tailor them for action control; however, these methods lack sufficient real-time performance for high-dynamic scenarios. A notable example is DreamZero (GEAR, 2026): even after a 38× speedup from extensive engineering optimization, it attains only 7Hz on a GB200 GPU, presenting a significant barrier to robotic edge deployment and real-time inference. Instead, we choose to implement a lightweight generation module. In our preceding work $\mathcal{F}_1$ (Lv et al., 2025), we adopted a next-resolution paradigm for generating visual foresight. Although effective, this approach required iterative forward passes, which compromised the real-time inference capabilities essential for VLA tasks. In this work, we have pursued a more efficient and effective design strategy.

The detailed architectural design of the Generation Expert is illustrated in Figure 3. Specifically, we expand upon three key specifics: Input Tokenization, Token Compression, and Parallel Decoding.

Input Tokenization: the Generation Expert takes as input images from three perspectives, including a head view and two wrist views. To capture temporal dynamics, we sample frames from two timestamps: the current timestep t and a historical timestep $t - 15$. This results in a total of 6 input images, each resized to a resolution of 256×256. Inspired by the unified multimodal model Janus Pro (Chen et al., 2025e), we adopt a decoupled visual encoding strategy to address the divergent requirements of understanding and generation. Unlike understanding tasks demanding high-level semantic abstraction typically captured by ViT-based encoders, generation tasks require preserving fine-grained spatial structure and pixel-level fidelity. Accordingly, our generation expert employs a VAE-based tokenizer, widely employed in image and video generation for its ability to compress visual data into a latent space optimized for high-quality reconstruction. Specifically, we utilize the COSMOS CI8×8 continuous VAE image tokenizer (Agarwal et al., 2025) to encode these input images. In its raw form, each image is encoded into a 32×32 latent grid, resulting in 1,024 tokens per image. Directly feeding these $1,024 \times 6$ latent tokens into the generation expert would result in an excessive sequence length, hindering both inference efficiency and training convergence. Therefore, token compression is required.

Token Compression: to address the sequence length challenge, we implement a token compression mechanism. We apply a convolution layer with an 8×8 kernel to downsample the latent grid, reducing the representation of each image to 4×4 (16 tokens). Consequently, the input sequence for the 6 images (3 views × 2 timestamps) is compacted to just 96 tokens. These tokens are then passed through a projector to align with the transformer blocks' hidden dimension. Within the transformer blocks, tokens from the Generation Expert attend to both themselves and the Understanding Expert's tokens via a unified masked self-attention mechanism, utilizing the preceding Understanding Expert tokens as the KV cache. Subsequently, the processed tokens undergo parallel decoding to generate future images.

Parallel Decoding: following the transformer processing, the output remains a sequence of 96 tokens. We apply temporal average pooling along the time axis to aggregate information from the two timestamps, resulting in 48 tokens representing the three views (16 tokens per view). These tokens are processed by a projector and subsequently upsampled via a deconvolution layer back to a 32×32 grid. Finally, the COSMOS VAE decoder reconstructs the predicted future frames (corresponding to $t + 15$). Notably, our approach avoids auto-regressive next-token prediction. Instead, we employ a single-forward parallel decoding strategy, generating all tokens for the future frames simultaneously. We find that this non-autoregressive approach is not only computationally efficient but also sufficient to provide effective visual guidance for the subsequent action execution.

Action Expert. Conditioned on the latent features produced by the understanding and generation experts, together with the proprioceptive state q_t, the action expert predicts a target action chunk $\hat{a}_{t:t+k}$. We adopt a flow matching objective to train the VLA model.

Attention Mechanism. We implement a blockwise attention mask over the concatenated token streams of the understanding, generation, and action experts. A cumulative segment mask enforces a strict information flow understanding → generation → action: tokens in a later block can attend to all earlier blocks, while earlier blocks cannot attend forward. Within the transformer blocks of both the understanding and generation experts, the tokens are fully bidirectionally attended. The tokens processed by the action expert's transformer blocks are split into a state token and action tokens. The state token attends only to itself and the tokens from earlier blocks, while action tokens attend to the state token, previous block tokens, and each other.

3.3. Optimization Objectives

Our training process consists of two sequential stages: **Pre-training** and **Post-training**. Although these two stages utilize distinct data sources and training hyperparameters (detailed in Table 2), they share a unified optimization framework and identical objectives. Throughout both stages, we jointly optimize the model for two key objectives: visual foresight generation and action prediction. Let $\mathcal{D}_1$ denote the training corpus for visual foresight (including human video data without action labels), and $\mathcal{D}_2$ denote the subset with action annotations. We define $\xi_1 = (o_{t-m}, o_t, o_{t+m}, l)$ and $\xi_2 = (a_{t:t+k}, o_{t-m}, o_t, q_t, l)$ as the corresponding training tuples.

(1) Visual Foresight Generation. To endow the model with predictive capabilities about future visual states, we train the generation expert to forecast the latent representation of a future frame. Let ϕ_{cosmos} denotes the COSMOS VAE encoder and $z_t = \phi_{\text{cosmos}}(o_t)$ denotes the COSMOS latent feature. Conditioned on current and historical observations $\{o_{t-m}, o_t\}$, as well as the understanding prefix h_{und}, the generation expert predicts the latent feature $\hat{z}_{t+m}$ for the future timestamp $t + m$. The prediction is supervised by the ground truth COSMOS latent feature z_{t+m}. We minimize the following objective:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\xi_1}\left[\left\| f_{\text{gen}}(z_{t-m}, z_t; h_{\text{und}}) - \text{sg}[z_{t+m}] \right\|^2\right] \tag{1}$$

where f_{gen} denotes the generation expert, and $\text{sg}[\cdot]$ indicates the stop-gradient operation. This objective compels the model to internalize physical dynamics, creating a robust prior for action.

(2) Flow Matching-based Action Prediction. We employ a flow matching framework for action learning. This approach formulates the action generation process as learning continuous transformation pathways from noise to expert demonstrations, offering superior handling of multi-modal action distributions compared to direct regression. Formally, given $\xi_2 \sim \mathcal{D}_2$, we sample time steps $\tau \sim \text{Beta}(1.5, 1.0)$ and construct interpolated action chunks $a^{\tau}_{t:t+k} = (1-\tau)\epsilon + \tau a_{t:t+k}$, where $\epsilon \sim \mathcal{N}(0, I)$ represents Gaussian noise. The model learns a velocity field v_θ that transports noisy samples toward target actions:

$$\mathcal{L}_{\text{act}} = \mathbb{E}_{\xi_2}\left[\left\|v_\theta(q_t, a^\tau_{t:t+k}; h_{\text{und}}, h_{\text{gen}}) - (a_{t:t+k} - \epsilon)\right\|^2\right], \tag{2}$$

where q_t denotes the proprioception state at time t, and h_{und} and h_{gen} are the contextual conditioning features produced by the understanding expert and the generation expert, respectively. During inference, sampling from the learned policy distribution is achieved by solving an ODE: starting from Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, we iteratively apply the Euler update:

$$a^{\tau+\Delta\tau}_{t:t+k} = a^\tau_{t:t+k} + \Delta\tau \cdot v_\theta(q_t, a^\tau_{t:t+k}; h_{\text{und}}, h_{\text{gen}}), \tag{3}$$

where τ progresses from 0 to 1 over K steps with step size $\Delta\tau = 1/K$.

Loss Function. The total training objective is a weighted summation of two loss components:

$$\mathcal{L}_{\text{total}} = \lambda \cdot \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{act}} \tag{4}$$

where λ is a hyperparameter balancing the two objectives. This joint optimization enforces representational consistency across modalities, enables implicit causal modeling of action-environment dynamics, and facilitates cross-modal knowledge transfer for enhanced generalization.

3.4. Implementation Details

Model configurations and parameters. We instantiate our model with two scales: InternVLA-A1 (2B) and InternVLA-A1 (3B). Both are built upon MLLM backbones and expanded into a unified system via the Mixture-of-Transformers (MoT) architecture. Specifically, InternVLA-A1 (2B) utilizes InternVL3-1B as the understanding expert. Its generative and action experts are derived from the transformer blocks of Qwen2.5—the underlying LLM of InternVL3. InternVLA-A1 (3B) employs Qwen3-VL-2B as the foundation, with its generative and action experts derived from the Qwen3 transformer blocks. Detailed configurations and parameters are provided in Table 1.

Regarding inference efficiency, both InternVLA-A1 (2B) and InternVLA-A1 (3B) run at approximately 13 Hz with `torch.compile` on one RTX 4090 GPU. Notably, InternVLA-A1 (2B) does not exhibit lower latency than InternVLA-A1 (3B) despite having fewer total parameters. This is because the InternVL3 backbone in InternVLA-A1 (2B) requires a higher input resolution 448×448 compared to the 224×224 input used by the Qwen3-VL backbone in InternVLA-A1 (3B). Consequently, the computational cost of processing longer visual token sequences in InternVLA-A1 (2B) offsets its parameter advantage, leading to comparable overall speeds.

Table 1. Model configurations and parameters for InternVLA-A1. The inference speed in frame per second (FPS) is evaluated on NVIDIA RTX 4090 GPU.

Model Variant	#Param.	Und. Expert	Gen. Expert	Act. Expert	FPS
InternVLA-A1 (2B)	1.8B	InternVL3 (0.94B)	Qwen2.5 (0.36B)	Qwen2.5 (0.36B)	~13 Hz
InternVLA-A1 (3B)	3.2B	Qwen3-VL (2.13B)	Qwen3 (0.44B)	Qwen3 (0.44B)	~13 Hz

Training Protocol and Hyperparameters. We train InternVLA-A1 in two stages: a large-scale pre-training stage on heterogeneous datasets, followed by a task-specific post-training stage. During pre-training, we optimize the model using AdamW with a constant learning rate schedule for 700K steps. For post-training, we adopt a lower learning rate with a warmup and decay schedule to stabilize adaptation to downstream tasks. Detailed hyperparameter settings for pre-training and post-training are provided Table 2. In addition, the hyperparameter λ, which balances the two loss components, is

Table 2. Training hyperparameters for InternVLA-A1.

Configuration	Pre-training	Post-training
Optimizer	AdamW	AdamW
Batch size	512	128
Learning rate	5×10^{-5} (constant)	$5 \times 10^{-5} \rightarrow 5 \times 10^{-6}$
Warmup steps	–	2,000
Decay steps	–	60,000
Training steps	700K	60,000
Optimizer betas	$\beta_1 = 0.9,\ \beta_2 = 0.95$	$\beta_1 = 0.9,\ \beta_2 = 0.95$
Optimizer epsilon	1×10^{-8}	1×10^{-8}
Weight decay	0.01	0.01
Gradient clipping	1.0	1.0
Model precision	bfloat16	bfloat16

set to 0.01. The interval m between the historical frame and the current frame, as well as between the future frame and the current frame, is set to 15.

Load-balanced Parallel Training. InternVLA-A1 is trained on a mixture of heterogeneous datasets over 692 frames. At this scale, naively instantiating the complete datasets on every worker like LeRobot codebase (Cadene et al., 2024) can trigger out-of-memory issues and exacerbate I/O contention. We therefore adopt Load-balanced Parallel Training (LPT), a distributed data-loading strategy that assigns datasets to workers to achieve both scalability and statistically well-behaved sampling.

Let $\{D_i\}_{i=1}^{n}$ denote the set of training datasets,and let s_i be a lightweight proxy of the size of D_i (e.g., the number of frames). LPT computes an assignment $\pi : \{1, \ldots, n\} \rightarrow \{1, \ldots, K\}$ that maps each dataset to one of K workers, subject to two desiderata: coverage (each worker receives at least one dataset) and balance (the total assigned size per worker is approximately uniform). In practice, we employ a simple greedy load-balancing rule that iteratively assigns the next dataset to the currently least-loaded worker:

$$\pi(i) \;=\; \operatorname{argmin}_{k \in \{1,\ldots,K\}} \sum_{j:\, \pi(j)=k} s_j \,,$$

with datasets processed in descending order of s_i. This procedure is efficient, deterministic, and empirically yields near-uniform per-worker loads.

LPT improves robustness in large-scale training by (i) reducing per-worker memory pressure since each worker materializes only a subset of datasets, and (ii) mitigating implicit re-weighting effects that would otherwise arise when workers traverse datasets with highly heterogeneous sizes at different rates. When the number of datasets is smaller than the number of workers, we allow controlled replication to avoid idle workers. Replicated datasets are assigned to different workers with independent random seeds and load-aware placement, such that no worker is disproportionately dominated by a small dataset. While replicas may sample from overlapping episode pools, this strategy empirically approximates uniform effective sampling by equalizing per-worker data throughput, thereby mitigating the implicit re-weighting effects introduced by dataset size heterogeneity.

4. Data Corpus

4.1. Pre-training data recipe

We pretrain the model on a mixture of heterogeneous data sources including synthetic simulation data, real-world robot data, and human videos. The pre-training data recipe is shown in Table 3.

During pretraining, we interleave trajectories from different sources using configurable sampling weights.

Table 3. Data mixture used for pretraining.

Data source	Type	Num. frame	Sampling weight
InternData-A1	Sim.	396M	0.64
RoboTwin	Sim.	17M	0.08
AgiBot-World (Beta)	Real	206M	0.18
RoboMind	Real	5M	0.02
EgoDex	Human	68M	0.08

4.2. Simulated synthetic data

We incorporate our prior work, InternData-A1 (Tian et al., 2025b), a large-scale synthetic robot dataset that is among the most diverse and comprehensive to date. The dataset contains over 630k trajectories and 7,433 hours of data spanning 4 embodiments, 18 skills, 70 tasks, and 227 scenes, covering manipulation of rigid, articulated, deformable, and fluid objects. It is generated via a highly autonomous, fully decoupled, and compositional simulation pipeline, enabling long-horizon skill composition, flexible task assembly, and support for heterogeneous embodiments with minimal manual tuning. InternData-A1 is the first to demonstrate that synthetic-only data can match the performance of large-scale real-world datasets when pre-training VLA models, achieving comparable results to the strongest closed-source real-world π-dataset (Black et al., 2024). Moreover, models trained on InternData-A1 exhibit strong zero-shot sim-to-real transfer on several challenging tasks. With InternData-A1's synthesis pipeline and system optimization enabled by the Nimbus (He et al., 2026), it generates 209.7 hours of simulation data per day on 8 RTX 4090 GPUs. As the foundation of our data corpus, InternData-A1 provides rich diversity in trajectories, objects, and environments. We select InternData-A1 as the foundation of our pre-training corpus due to its exceptional sample diversity and proven efficacy in pre-training VLA models. Additionally, we incorporate the simulation dataset proposed in RoboTwin (Chen et al., 2025c).

4.3. Real-world robot data

While synthetic data excels in scalability, real-world demonstrations remain essential for capturing nuanced physical dynamics and bridging the sim-to-real gap. When selecting real-world demonstration data for our pre-training corpus, we prioritize datasets that exhibit large-scale trajectory coverage, diverse task distributions, and high-quality teleoperation. Based on these criteria, we incorporate the open-source AgiBot-World dataset (Bu et al., 2025) and RoboMind (Wu et al., 2024b). By incorporating these large-scale real-world demonstration datasets into our pre-training corpus, we benefit from its high-quality and diverse demonstrations, which complement our synthetic data and help bridge the sim-to-real gap.

4.4. Egocentric human video

We additionally incorporate human video data into our data corpus. To align with robot manipulation scenarios, we prioritize egocentric videos that share a similar viewpoint and feature diverse interactions in robot-like environments. Specifically, we utilize the EgoDex (Hoque et al., 2025) dataset, a large-scale collection focused on egocentric dexterous manipulation. It comprises 829 hours of footage spanning over 200 tasks. Notably, we exclude human action labels during pre-training. Rich in

human-hand-object interactions, this data is instrumental for the generation expert, enabling it to capture the nuanced dynamics and diverse manipulation skills inherent in the real world. The human video data can also be further expanded to incorporate large-scale datasets like Ego4D (Grauman et al., 2024) and EPIC-KITCHENS (Damen et al., 2020).

5. Experiments

To evaluate the effectiveness of our proposed InternVLA-A1, we conduct extensive experiments on 12 real-world tasks and a simulation benchmark. We first compare InternVLA-A1 with existing leading VLA models. Following this, we conduct a series of ablation studies.

5.1. Evaluation Configuration

Hardware & platform. We benchmark the policies on three physical robot embodiments: **Agibot Genie-1**, **ARX Lift-2**, and **ARX AC One**. These platforms cover diverse bimanual manipulation capabilities and execution characteristics, enabling us to assess real-robot performance across both long-horizon and contact-rich behaviors under consistent sensing and control setups. All evaluations are performed with the same deployment pipeline across the three robots.

Task setup. We design a real-world task suite comprising **ten** static manipulation tasks and **two** dynamic task families:

- **Static manipulation tasks:** *Zip Bag*, *Unscrew Cap*, *Make Sandwich*, *Operate Oven*, *Sort Parts*, *Sort Rubbish*, *Sweep Trash*, *Wipe Stain*, *Place Markpen*, *Place Flower*.
- **Dynamic manipulation tasks:** *Express Sorting tasks* and *In-motion Ingredient Picking tasks*.

Overall, the tasks span articulated-object interaction and contact-rich manipulation (e.g., Operate Oven and Unscrew Cap), long-horizon bimanual manipulation (e.g., Zip Bag and Make Sandwich), and, importantly, high-dynamics settings where the scene evolves and the target is moving during execution. The latter requires the policy to reason about near-future scene changes, so that actions can be generated with upcoming dynamics in mind rather than solely based on the current observation.

Evaluation protocol. We report the average success rate over 30 rollouts per task. Specifically, each task is evaluated under 30 predefined settings (e.g., object placements and scene initializations within bounded ranges), with one trial per setting. Results are averaged across all rollouts to summarize performance across diverse conditions.

5.2. Evaluation on the Static Manipulation Tasks

To evaluate the general manipulation capabilities of InternVLA-A1 against leading VLA models, we conducted a comprehensive evaluation across 10 diverse real-world static tasks (Figure 4).

As shown in Table 4, InternVLA-A1 clearly outperforms prior state-of-the-art models on real-world static manipulation tasks. Notably, our smaller InternVLA-A1 (2B) achieves an average success rate of 64.7%, surpassing the 60.6% average of the larger π_0 (3.3B) baseline. This result highlights the effectiveness of our architecture and joint training of simulated trajectories, real-robot demonstrations, and egocentric human videos, enabling a lighter model to outperform a larger counterpart. When scaling to a comparable size, InternVLA-A1 (3B) delivers a clear performance gain, reaching an average success rate of 75.1%, which is +14.5% over π_0 (60.6%) and +4.4% over $\pi_{0.5}$ (70.7%). It shows strong advantages on long-horizon tasks such as Make Sandwich (93.3% vs. 73.3% for $\pi_{0.5}$) and Operate Oven (86.7% vs. 80.0%), and remains top-tier on Sort Rubbish (97.3%), matching the

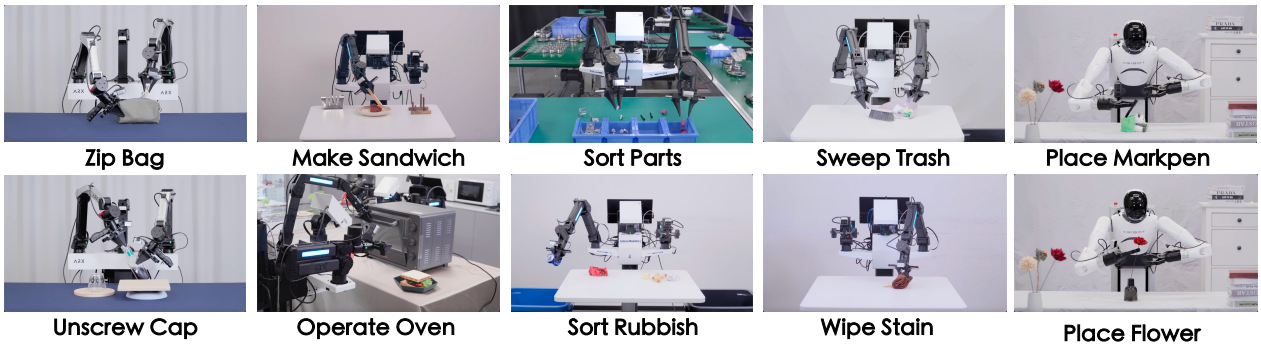

Figure 4. **The experimental setting of real-world static tasks**.

best baseline. On challenging manipulation tasks, InternVLA-A1 (3B) matches the strongest baseline on Unscrew Cap (66.7%) and Sort Parts (53.3%), while achieving substantial gains on Zip Bag (73.3% vs. 60.0% for $\pi_{0.5}$, and 40.0% for π_0). These results confirm that InternVLA-A1 not only excels in high-level task planning but also possesses superior fine-grained control capabilities.

Table 4. **Experimental results on real-world general-purpose tasks.** The best results are **bolded**.

Method	Zip Bag	Unscrew Cap	Sort Parts	Make Sandwich	Sweep Trash
GR00T N1.5	33.3	0.0	6.7	46.7	16.7
π_0	40.0	**66.7**	**53.3**	66.7	43.3
$\pi_{0.5}$	60.0	**66.7**	**53.3**	73.3	50.0
InternVLA-A1(2B)	66.7	33.3	46.7	73.3	63.3
InternVLA-A1(3B)	**73.3**	**66.7**	**53.3**	**93.3**	**66.7**

Operate Oven	Sort Rubbish	Wipe Stain	Place Markpen	Place Flower	Average
46.7	66.7	40.0	40.0	33.3	33.0
73.3	96.0	73.3	53.3	40.0	60.6
80.0	**97.3**	**93.3**	**66.7**	**66.7**	70.7
53.3	**97.3**	80.0	**66.7**	**66.7**	64.7
86.7	**97.3**	86.7	**66.7**	60.0	**75.1**

5.3. Evaluation on Dynamic Manipulation Tasks

To assess robustness under environmental dynamics, we evaluate the policies on two challenging real-world dynamic manipulation tasks: *Express Sorting* and *In-motion Ingredient Picking*, where targets are moving during execution. The experimental settings for the Express Sorting and In-motion Ingredient Picking tasks are depicted in Figure 5. Both involve long-horizon operations within dynamic environments. In the Express Sorting Task, the robot must first determine if the package label is facing upwards. If the label is facing downwards, a four-step sequence is initiated: the right arm first executes a "chasing" grasp along the direction of the conveyor movement, followed by flipping the package, after which the left arm performs a "waiting" grasp. Finally, the package is lifted to present the label to the head-mounted camera. Conversely, if the label is facing upwards, the robot skips the flipping sequence and directly proceeds to the final two steps. In the In-motion Ingredient Picking tasks, two robots coordinate to grasp the ingredients required to assemble a beef sandwich, consisting of two slices of bread, a steak, and a piece of lettuce.

As shown in Figure 6, InternVLA-A1 establishes a clear lead over all baselines, surpassing even the strongest prior VLA model $\pi_{0.5}$. Notably, the InternVLA-A1 (3B) variant demonstrates the most

Figure 5. **The experimental setting of real-world dynamic tasks: (a) Express Sorting task, (b) In-motion Ingredient Picking task.**

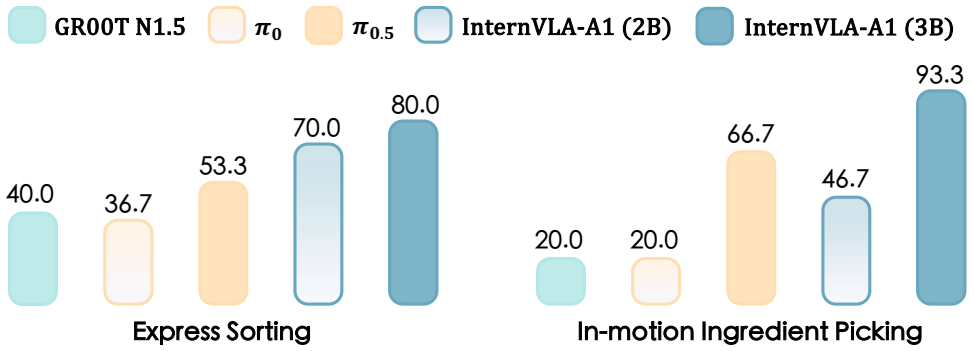

Figure 6. **The experimental results of Express Sorting and In-motion Ingredient Picking tasks**.

pronounced advantage. While InternVLA-A1 (2B) already performs strongly on Express Sorting (70.0%), surpassing $\pi_{0.5}$ (53.3%) by +16.7% and substantially outperforming GR00T N1.5 (40.0%) and π_0 (36.7%), scaling to 3B pushes performance to 80.0% on Express Sorting and 93.3% on Ingredient Picking. Notably, on Ingredient Picking, a highly dynamic and timing-sensitive setting, InternVLA-A1 (3B) delivers a substantial gain over $\pi_{0.5}$ (93.3% vs. 66.7%, +26.6%), while π_0 and GR00T N1.5 remain at 20.0%. Similarly, on Express Sorting it improves over $\pi_{0.5}$ by +26.7% (80.0% vs. 53.3%). Overall, InternVLA-A1 (3B) achieves an average success rate of 86.7% across these two dynamic tasks. These results highlight that InternVLA-A1 is markedly more robust to environmental dynamics and motion-induced distribution shifts, translating foresight-guided decision making into reliable closed-loop control in dynamic scenarios.

5.4. Evaluation on Simulation Benchmark

We also evaluate InternVLA-A1 on the RoboTwin 2.0 (Chen et al., 2025c) benchmark, covering 50 bimanual tasks under both Easy (clean) and Hard (domain randomized) settings. All models are fine-tuned on a total of 27,500 demonstrations, consisting of 2,500 clean and 25,000 randomized episodes (i.e., 50 and 500 per task), with results averaged over 100 evaluation trials. As shown in Figure 7, InternVLA-A1(3B) significantly outperforms π_0 by margins of 9.4% and 10.1%, and surpasses the strong $\pi_{0.5}$ baseline by 2.6% in both the Easy and Hard settings.

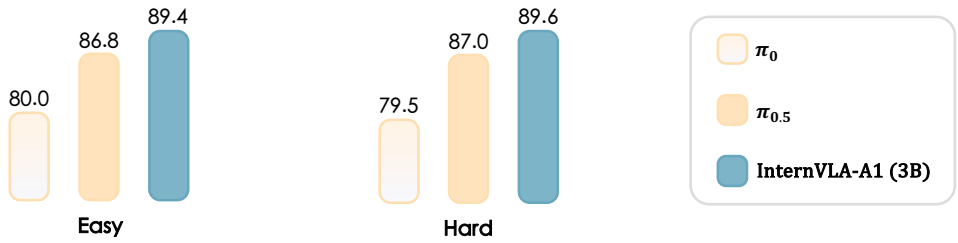

Figure 7. **Evaluation on RoboTwin 2.0 Simulation Benchmark**.

5.5. Ablation Studies

Impact of Pre-training. As shown in Figure 8, we assess the impact of pre-training by comparing our model with a version trained from scratch. The removal of the pre-training stage resulted in an overall performance drop of 51.6%, with the average success rate falling from 77.0% to 25.4%. In severe cases, the baseline failed completely, while the pre-trained model retained high proficiency. This finding suggests that pre-training acts as a crucial inductive prior.

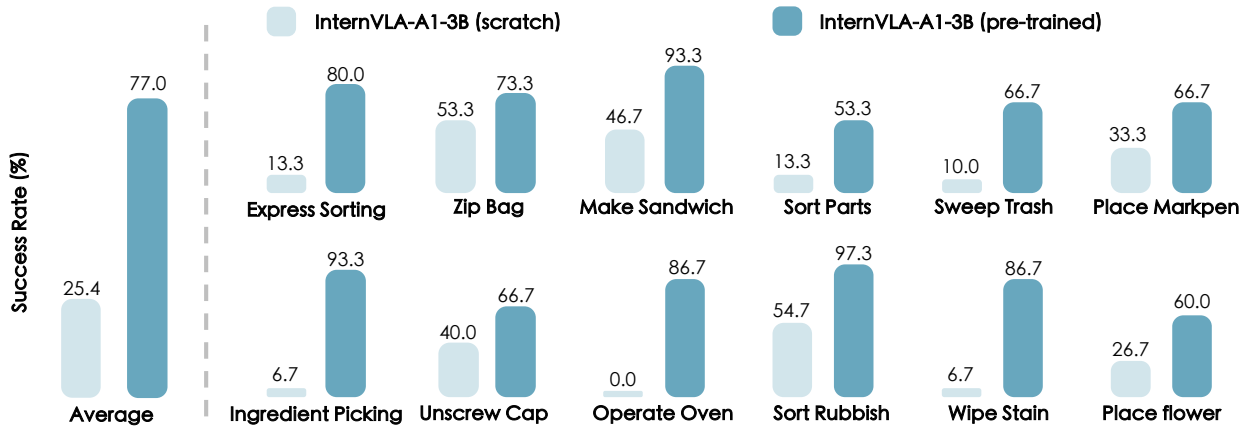

Figure 8. **The ablation studies on pre-training**.

Impact of Pre-training Datasets. As shown in Table 5, we evaluate the efficacy of different pre-training data sources. Training on simulation data alone already yields strong results on RoboTwin 2.0 (88.3%/88.5% on Easy/Hard), but generalizes less well to real-world tasks (53.3% on Place flower and 33.3% on Sort parts). Incorporating human videos improves simulation performance (up to 89.4%/89.3%) and slightly benefits real-world generalization (specifically improving Sort parts to 40.0%). Most notably, jointly pre-training on heterogeneous data sources (human videos, synthetic data, and real-world demonstrations) achieves the best overall performance, and substantially enhances real-world manipulation performance (reaching 60.0% on Place flower and 53.3% on Sort parts). This demonstrates the effectiveness of our joint training strategy.

Table 5. **Ablation studies on the pre-training dataset.**

Pre-training dataset	RoboTwin 2.0		Real-world tasks	
	Easy	Hard	Place flower	Sort parts
Sim. only	88.3	88.5	53.3	33.3
Sim. + Human	**89.4**	89.3	53.3	40.0
Sim. + Real + Human	**89.4**	**89.6**	**60.0**	**53.3**

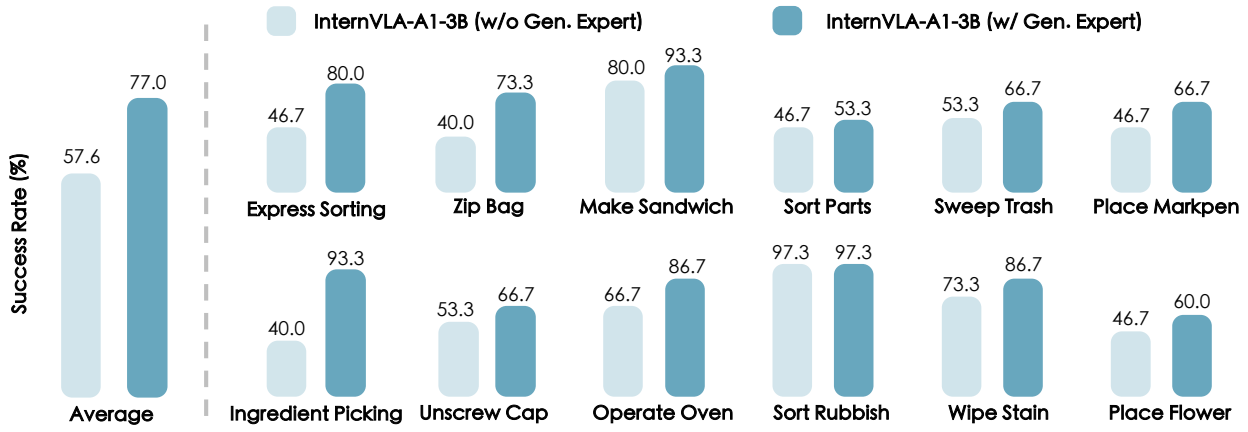

Figure 9. **The ablation studies on generation expert**.

Impact of Generation Expert. To evaluate the contribution of the generation expert, we conduct a key ablation study by comparing InternVLA-A1 (3B) with the variant without the generation expert. The experimental results are presented in Figure 9. The results demonstrate that InternVLA-A1 (3B) outperforms the ablated version (without the generation expert) in 11 out of 12 real-world tasks, achieving performance gains ranging from 6.7% to 53.3%. Notably, removing the generation expert significantly reduces the average success rate from 77.0% to 57.6%, with the most substantial decline observed in dynamic manipulation tasks (Express Sorting and In-motion Ingredient Picking). This ablation study validates the superiority of the proposed generation expert and the unified architecture integrating understanding, generation, and action.

5.6. Visualization

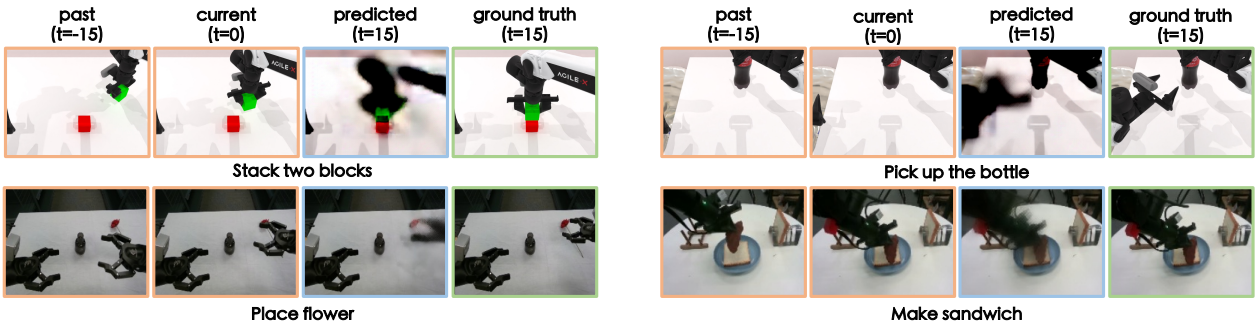

Figure 10. **Visualization of future image prediction**.

The head-view future predictions generated by InternVLA-A1 are illustrated in Figure 10. As shown in the samples, the predictions sacrifice some high-frequency visual details. This is a deliberate design choice to prioritize inference efficiency over pixel-level granularity. Despite this, the model accurately captures the essential motion trends and dynamics. We argue that when integrating visual foresight generation module into manipulation policies, high-frequency detail or visual clarity is secondary; what matters most is that the latent features encapsulate sufficient instructive information to guide action execution.

6. Conclusion and Limitations

In this work, we presented InternVLA-A1, a unified framework that integrates scene understanding, visual-foresight generation, and action execution through a Mixture-of-Transformers (MoT) design. This architecture couples semantic reasoning with dynamics prediction, and enables effective joint training on heterogeneous data sources (human videos, synthetic data, and real-world demonstrations). As a result, InternVLA-A1 achieves consistent robustness across static manipulation, dynamic manipulation, and simulation benchmarks, with particularly strong gains in highly dynamic scenarios.

Limitations. Despite these advancements, two primary limitations remain. First, the understanding expert is not jointly trained with large-scale multimodal VQA data, which weakens general semantic reasoning and complex instruction following. Second, to ensure efficient inference for the visual foresight generation module, we compromised the fidelity of image prediction, limiting the granularity of generated future frames. We will address the limitations in future work.

References

N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding, D. Dworakowski, J. Fan, M. Fenzi, F. Ferroni, S. Fidler, D. Fox, S. Ge, Y. Ge, J. Gu, S. Gururani, E. He, J. Huang, J. Huffman, P. Jannaty, J. Jin, S. W. Kim, G. Klár, G. Lam, S. Lan, L. Leal-Taixe, A. Li, Z. Li, C.-H. Lin, T.-Y. Lin, H. Ling, M.-Y. Liu, X. Liu, A. Luo, Q. Ma, H. Mao, K. Mo, A. Mousavian, S. Nah, S. Niverty, D. Page, D. Paschalidou, Z. Patel, L. Pavao, M. Ramezanali, F. Reda, X. Ren, V. R. N. Sabavat, E. Schmerling, S. Shi, B. Stefaniak, S. Tang, L. Tchapmi, P. Tredak, W.-C. Tseng, J. Varghese, H. Wang, H. Wang, H. Wang, T.-C. Wang, F. Wei, X. Wei, J. Z. Wu, J. Xu, W. Yang, L. Yen-Chen, X. Zeng, Y. Zeng, J. Zhang, Q. Zhang, Y. Zhang, Q. Zhao, and A. Zolkowski. Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575, 2025.

J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. In Advances in Neural Information Processing Systems (NeurIPS), pages 23716–23736, 2022.

M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, Mojtaba, Komeili, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus, S. Arnaud, A. Gejji, A. Martin, F. R. Hogan, D. Dugas, P. Bojanowski, V. Khalidov, P. Labatut, F. Massa, M. Szafraniec, K. Krishnakumar, Y. Li, X. Ma, S. Chandar, F. Meier, Y. LeCun, M. Rabbat, and N. Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985, 2025.

L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, T. Unterthiner, D. Keysers, S. Koppula, F. Liu, A. Grycner, A. Gritsenko, N. Houlsby, M. Kumar, K. Rong, J. Eisenschlos, R. Kabra, M. Bauer, M. Bošnjak, X. Chen, M. Minderer, P. Voigtlaender, I. Bica, I. Balazevic, J. Puigcerver, P. Papalampidi, O. Henaff, X. Xiong, R. Soricut, J. Harmsen, and X. Zhai. Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726, 2024.

H. Bi, H. Tan, S. Xie, Z. Wang, S. Huang, H. Liu, R. Zhao, Y. Feng, C. Xiang, Y. Rong, H. Zhao, H. Liu, Z. Su, L. Ma, H. Su, and J. Zhu. Motus: A unified latent action world model. arXiv preprint arXiv:2512.13030, 2025a.

H. Bi, L. Wu, T. Lin, H. Tan, Z. Su, H. Su, and J. Zhu. H-rdt: Human manipulation enhanced bimanual robotic manipulation. arXiv preprint arXiv:2507.23523, 2025b.

J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. J. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, J. Jang, Z. Jiang, J. Kautz, K. Kundalia, L. Lao, Z. Li, Z. Lin, K. Lin, G. Liu, E. Llontop, L. Magne, A. Mandlekar, A. Narayan, S. Nasiriany, S. Reed, Y. L. Tan, G. Wang, Z. Wang, J. Wang, Q. Wang, J. Xiang, Y. Xie, Y. Xu, Z. Xu, S. Ye, Z. Yu, A. Zhang, H. Zhang, Y. Zhao, R. Zheng, and Y. Zhu. Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734, 2025.

K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky. π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024.

K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine,

A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054, 2025.

A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127, 2023a.

A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127, 2023b.

A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817, 2022.

A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In arXiv preprint arXiv:2307.15818, 2023.

Q. Bu, H. Li, L. Chen, J. Cai, J. Zeng, H. Cui, M. Yao, and Y. Qiao. Towards synergistic, generalized, and efficient dual-system for robotic manipulation. arXiv preprint arXiv:2410.08001, 2024a.

Q. Bu, J. Zeng, L. Chen, Y. Yang, G. Zhou, J. Yan, P. Luo, H. Cui, Y. Ma, and H. Li. Closed-loop visuomotor control with generative expectation for robotic manipulation. In Advances in Neural Information Processing Systems (NeurIPS), 2024b.

Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang, S. Jiang, Y. Jiang, C. Jing, H. Li, J. Li, C. Liu, Y. Liu, Y. Lu, J. Luo, P. Luo, Y. Mu, Y. Niu, Y. Pan, J. Pang, Y. Qiao, G. Ren, C. Ruan, J. Shan, Y. Shen, C. Shi, M. Shi, M. Shi, C. Sima, J. Song, H. Wang, W. Wang, D. Wei, C. Xie, G. Xu, J. Yan, C. Yang, L. Yang, S. Yang, M. Yao, J. Zeng, C. Zhang, Q. Zhang, B. Zhao, C. Zhao, J. Zhao, and J. Zhu. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint arXiv:2503.06669, 2025.

R. Cadene, S. Alibert, A. Soare, Q. Gallouedec, A. Zouitine, S. Palma, P. Kooijmans, M. Aractingi, M. Shukor, D. Aubakirova, M. Russi, F. Capuano, C. Pascal, J. Choghari, J. Moss, and T. Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. https://github.com/huggingface/lerobot, 2024.

J. Cen, C. Yu, H. Yuan, Y. Jiang, S. Huang, J. Guo, X. Li, Y. Song, H. Luo, F. Wang, D. Zhao, and H. Chen. Worldvla: Towards autoregressive action world model. arXiv preprint arXiv:2506.21539, 2025.

C. Cheang, S. Chen, Z. Cui, Y. Hu, L. Huang, T. Kong, H. Li, Y. Li, Y. Liu, X. Ma, H. Niu, W. Ou, W. Peng, Z. Ren, H. Shi, J. Tian, H. Wu, X. Xiao, Y. Xiao, J. Xu, and Y. Yang. Gr-3 technical report. arXiv preprint arXiv:2507.15493, 2025.

C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, H. Wu, J. Xu, Y. Yang, H. Zhang, and M. Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. arXiv preprint arXiv:2410.06158, 2024.

G. Chen, Z. Li, S. Wang, J. Jiang, Y. Liu, L. Lu, D.-A. Huang, W. Byeon, M. Le, T. Rintamaki, T. Poon, M. Ehrlich, T. Rintamaki, T. Poon, T. Lu, L. Wang, B. Catanzaro, J. Kautz, A. Tao, Z. Yu, and G. Liu. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. arXiv preprint arXiv:2504.15271, 2025a.

J. Chen, S. Xue, Y. Zhao, J. Yu, S. Paul, J. Chen, H. Cai, S. Han, and E. Xie. Sana-sprint: One-step diffusion with continuous-time consistency distillation. arXiv preprint arXiv:2503.09641, 2025b.

T. Chen, Z. Chen, B. Chen, Z. Cai, Y. Liu, Z. Li, Q. Liang, X. Lin, Y. Ge, Z. Gu, W. Deng, Y. Guo, T. Nian, X. Xie, Q. Chen, K. Su, T. Xu, G. Liu, M. Hu, H.-a. Gao, K. Wang, Z. Liang, Y. Qin, X. Yang, P. Luo, and Y. Mu. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. arXiv preprint arXiv:2506.18088, 2025c.

X. Chen, Y. Chen, Y. Fu, N. Gao, J. Jia, W. Jin, H. Li, Y. Mu, J. Pang, Y. Qiao, Y. Tian, B. Wang, B. Wang, F. Wang, H. Wang, T. Wang, Z. Wang, X. Wei, C. Wu, S. Yang, J. Ye, J. Yu, J. Zeng, J. Zhang, J. Zhang, S. Zhang, F. Zheng, B. Zhou, and Y. Zhu. Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy. arXiv preprint arXiv:2510.13778, 2025d.

X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, and C. Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811, 2025e.

J. Clark, S. Mirchandani, D. Sadigh, and S. Belkhale. Action-free reasoning for policy generalization. arXiv preprint arXiv:2502.03729, 2025.

C. Cui, P. Ding, W. Song, S. Bai, X. Tong, Z. Ge, R. Suo, W. Zhou, Y. Liu, B. Jia, H. Zhao, S. Huang, and D. Wang. Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation. arXiv preprint arXiv:2505.03912, 2025.

D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. The epic-kitchens dataset: Collection, challenges and baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11):4125–4141, 2020.

C. Deng, D. Zhu, K. Li, C. Gou, F. Li, Z. Wang, S. Zhong, W. Yu, X. Nie, Z. Song, G. Shi, and H. Fan. Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683, 2025a.

S. Deng, M. Yan, S. Wei, H. Ma, Y. Yang, J. Chen, Z. Zhang, T. Yang, X. Zhang, H. Cui, Z. Zhang, and H. Wang. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. arXiv preprint arXiv:2505.03233, 2025b.

Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel. Learning universal policies via text-guided video generation. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. IEEE Transactions on Robotics, 39(5): 3929–3945, 2023.

N. Gao, Y. Chen, S. Yang, X. Chen, Y. Tian, H. Li, H. Huang, H. Wang, T. Wang, and J. Pang. Genmanip: Llm-driven simulation for generalizable instruction-following manipulation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 12187–12198, 2025.

N. GEAR. Dreamzero: World action models are zero-shot policies, 2026.

R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In International Conference on Computer Vision (ICCV), pages 5842–5850, 2017.

K. Grauman, A. Westbury, E. Byrne, V. Cartillier, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, D. Kukreja, et al. Ego4d: Around the world in 3,000 hours of egocentric video. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–32, 2024.

J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, X. Yuan, P. Xie, Z. Huang, R. Chen, and H. Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In International Conference on Learning Representations (ICLR), 2023.

Y. HaCohen, N. Chiprut, B. Brazowski, D. Shalem, D. Moshe, E. Richardson, E. Levin, G. Shiran, N. Zabari, O. Gordon, P. Panet, S. Weissbuch, V. Kulikov, Y. Bitterman, Z. Melumian, and O. Bibi. Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103, 2024.

Z. He, Y. Zhang, Y. Zhou, M. Tao, H. Li, Y. Tian, J. Zeng, T. Wang, W. Cai, Y. Chen, N. Gao, and J. Pang. Nimbus: A unified embodied synthetic data generation framework. arXiv preprint arXiv:2601.21449, 2026.

R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. arXiv preprint arXiv:2505.11709, 2025.

Y. Hu, Y. Guo, P. Wang, X. Chen, Y.-J. Wang, J. Zhang, K. Sreenath, C. Lu, and J. Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In International Conference on Machine Learning (ICML), 2025.

P. Hua, M. Liu, A. Macaluso, Y. Lin, W. Zhang, H. Xu, and L. Wang. Gensim2: Scaling robot data generation with multi-modal and reasoning llms. arXiv preprint arXiv:2410.03645, 2024.

H. Huang, X. Chen, Y. Chen, H. Li, X. Han, Z. Wang, T. Wang, J. Pang, and Z. Zhao. Roboground: Robotic manipulation with grounded vision-language priors. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 22540–22550, 2025a.

H. Huang, F. Liu, L. Fu, T. Wu, M. Mukadam, J. Malik, K. Goldberg, and P. Abbeel. Otter: A vision-language-action model with text-aware visual feature extraction. In International Conference on Machine Learning (ICML), 2025b.

W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. arXiv preprint arXiv:2307.05973, 2023.

C.-Y. Hung, N. Majumder, H. Deng, L. Renhang, Y. Ang, A. Zadeh, C. Li, D. Herremans, Z. Wang, and S. Poria. Nora-1.5: A vision-language-action model trained using world model-and action-based preference rewards. arXiv preprint arXiv:2511.14659, 2025.

S. James, Z. Ma, D. R. Arrojo, and A. J. Davison. Rlbench: The robot learning benchmark & learning environment. IEEE Robotics and Automation Letters, 5(2):3019–3026, 2020.

S. Kareer, K. Pertsch, J. Darpinian, J. Hoffman, D. Xu, S. Levine, C. Finn, and S. Nair. Emergence of human to robot transfer in vision-language-action models. arXiv preprint arXiv:2512.22414, 2025.

A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, V. Guizilini, D. A. Herrera, M. Heo, K. Hsu, J. Hu, M. Z. Irshad, D. Jackson, C. Le, Y. Li, K. Lin, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O'Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, and C. Finn. Droid: A large-scale in-the-wild robot manipulation dataset. In Robotics: Science and Systems Conference (RSS), 2024.

M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246, 2024.

H. Li, S. Yang, Y. Chen, X. Chen, X. Yang, Y. Tian, H. Wang, T. Wang, D. Lin, F. Zhao, and J. Pang. Cronusvla: Transferring latent motion across time for multi-frame prediction in manipulation. arXiv preprint arXiv:2506.19816, 2025a.

Y. Li, Y. Deng, J. Zhang, J. Jang, M. Memmel, R. Yu, C. R. Garrett, F. Ramos, D. Fox, A. Li, A. Gupta, and A. Goyal. Hamster: Hierarchical action models for open-world robot manipulation. In International Conference on Learning Representations (ICLR), 2025b.

X. Lian, Z. Yu, R. Liang, Y. Wang, L. R. Luo, K. Chen, Y. Zhou, Q. Tang, X. Xu, Z. Lyu, B. Dai, and J. Pang. Infinite mobility: Scalable high-fidelity synthesis of articulated objects via procedural generation. arXiv preprint arXiv:2503.13424, 2025.

Y. Liao, P. Zhou, S. Huang, D. Yang, S. Chen, Y. Jiang, Y. Hu, J. Cai, S. Liu, J. Luo, L. Chen, S. Yan, M. Yao, and G. Ren. Genie envisioner: A unified world foundation platform for robotic manipulation. arXiv preprint arXiv:2508.05635, 2025.

F. Lin, R. Nai, Y. Hu, J. You, J. Zhao, and Y. Gao. Onetwovla: A unified vision-language-action model with adaptive reasoning. arXiv preprint arXiv:2505.11917, 2025.

Q. Lv, W. Kong, H. Li, J. Zeng, Z. Qiu, D. Qu, H. Song, Q. Chen, X. Deng, and J. Pang. F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951, 2025.

A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. arXiv preprint arXiv:2310.17596, 2023.

O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters, 7(3):7327–7334, 2022.

S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In Robotics: Science and Systems Conference (RSS), 2024.

NVIDIA. Gr00t-n1.5. https://huggingface.co/nvidia/GR00T-N1.5-3B, 2025.

A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Kolobov, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. V. Frujeri, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Yang, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. Ben Amor, H. I. Christensen, H. Furuta, H. Bharadhwaj, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Vakil, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, F.-F. Li, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, M. Z. Irshad, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. Di Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Tulsiani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Kumar, V. Vanhoucke, V. Guizilini, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, L. Xu, X. Li, Y. Pang, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Dou, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, Z. Fu, and Z. Lin. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In IEEE International Conference on Robotics and Automation (ICRA), 2024.

D. Qu, H. Song, Q. Chen, Z. Chen, X. Gao, X. Ye, Q. Lv, M. Shi, G. Ren, C. Ruan, M. Yao, H. Yang, J. Bao, B. Zhao, and D. Wang. Eo-1: Interleaved vision-text-action pretraining for general robot control. arXiv preprint arXiv:2508.21112, 2025.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 10684–10695, 2022.

P. Sun, Y. Jiang, S. Chen, S. Zhang, B. Peng, P. Luo, and Z. Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525, 2024.

Y. Tian, S. Yang, J. Zeng, P. Wang, D. Lin, H. Dong, and J. Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In International Conference on Learning Representations (ICLR), 2025a.

Y. Tian, Y. Yang, Y. Xie, Z. Cai, X. Shi, N. Gao, H. Liu, X. Jiang, Z. Qiu, F. Yuan, Y. Li, P. Wang, J. Cai, J. Zeng, H. Dong, and J. Pang. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. arXiv preprint arXiv:2511.16651, 2025b.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine. Bridgedata v2: A dataset for robot learning at scale. In Conference on Robot Learning (CoRL), 2023.

H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In International Conference on Learning Representations (ICLR), 2024a.

K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang, S. Fan, X. Wang, F. Liao, Z. Zhao, G. Li, Z. Jin, L. Wang, J. Mao, N. Liu, P. Ren, Q. Zhang, Y. Lyu, M. Liu, J. He, Y. Luo, Z. Gao, C. Li, C. Gu, Y. Fu, D. Wu, X. Wang, S. Chen, Z. Wang, P. An, S. Qian, S. Zhang, and J. Tang. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. arXiv preprint arXiv:2412.13877, 2024b.

S. Yang, Y. Du, S. K. S. Ghasemipour, J. Tompson, L. P. Kaelbling, D. Schuurmans, and P. Abbeel. Learning interactive real-world simulators. In International Conference on Learning Representations (ICLR), 2024.

S. Yang, H. Li, Y. Chen, B. Wang, Y. Tian, T. Wang, H. Wang, F. Zhao, Y. Liao, and J. Pang. Instructvla: Vision-language-action instruction tuning from understanding to manipulation. arXiv preprint arXiv:2507.17520, 2025a.

Y. Yang, Z. Cai, Y. Tian, J. Zeng, and J. Pang. Gripper keypose and object pointflow as interfaces for bimanual robotic manipulation. arXiv preprint arXiv:2504.17784, 2025b.

A. Zhai, B. Liu, B. Fang, C. Cai, E. Ma, E. Yin, H. Wang, H. Zhou, J. Wang, L. Shi, L. Liang, M. Wang, Q. Wang, R. Gan, R. Yu, S. Li, S. Liu, S. Chen, V. Chen, and Z. Xu. Igniting vlms toward the embodied space. arXiv preprint arXiv:2509.11766, 2025.

W. Zhang, H. Liu, Z. Qi, Y. Wang, X. Yu, J. Zhang, R. Dong, J. He, F. Lu, H. Wang, Z. Zhang, L. Yi, W. Zeng, and X. Jin. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. arXiv preprint arXiv:2507.04447, 2025.

Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, A. Handa, M.-Y. Liu, D. Xiang, G. Wetzstein, and T.-Y. Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1702–1713, 2025.

J. Zheng, J. Li, Z. Wang, D. Liu, X. Kang, Y. Feng, Y. Zheng, J. Zou, Y. Chen, J. Zeng, Y.-Q. Zhang, J. Pang, J. Liu, T. Wang, and X. Zhan. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. arXiv preprint arXiv:2510.10274, 2025.

Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You. Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404, 2024.

F. Zhu, Z. Yan, Z. Hong, Q. Shou, X. Ma, and S. Guo. Wmpo: World model-based policy optimization for vision-language-action models. arXiv preprint arXiv:2511.09515, 2025.

A. Contributors

All contributors are listed in alphabetical order by their last names.

Core Contributors

Junhao Cai[1], Yang Li[1], Haoxiang Ma[1], Jiangmiao Pang[1↯], Zherui Qiu[1], Yang Tian[1], Jia Zeng[1†], Hongrui Zhu[1]

Contributors

Zetao Cai[1], Jiafei Cao[1], Yilun Chen[1], Zeyu He[1], Lei Jiang[2], Hang Li[1], Hengjie Li[1], Yufei Liu[2], Yanan Lu[1], Qi Lv[1], Yu Qiao[1], Yanqing Shen[1], Xu Shi[1], Bolun Wang[1], Hanqing Wang[1], Jiaheng Wang[1], Tai Wang[1], Xueyuan Wei[1], Chao Wu[1], Yiman Xie[1], Boyang Xing[2], Yuqiang Yang[1], Yuyin Yang[1], Qiaojun Yu[1], Feng Yuan[1], Jingjing Zhang[1], Shenghan Zhang[1], Shi Zhang[1], Zhuoma Zhaxi[1], Bowen Zhou[1], Yuanzhen Zhou[1], Yunsong Zhou[1], Yangkun Zhu[1], Yuchen Zhu[2]

[1]Shanghai AI Laboratory [2]Humanoid Robot (Shanghai) Co., Ltd. [†]Project lead [↯]Corresponding author