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

InternVLA-A1 Team
Full author list in Appendix

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 when handling 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 hybrid synthetic-real datasets spanning InternData-A1 and Agibot-World, covering over 533M frames. This hybrid training strategy effectively harnesses the diversity of synthetic simulation data while minimizing the sim-to-real gap. We evaluated InternVLA-A1 across 12 real-world robotic tasks and simulation benchmark. It significantly outperforms leading models like π_0 and GR00T N1.5, achieving a 14.5% improvement in daily tasks and a 40%~73.3% boost in dynamic settings, such as conveyor belt sorting.

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

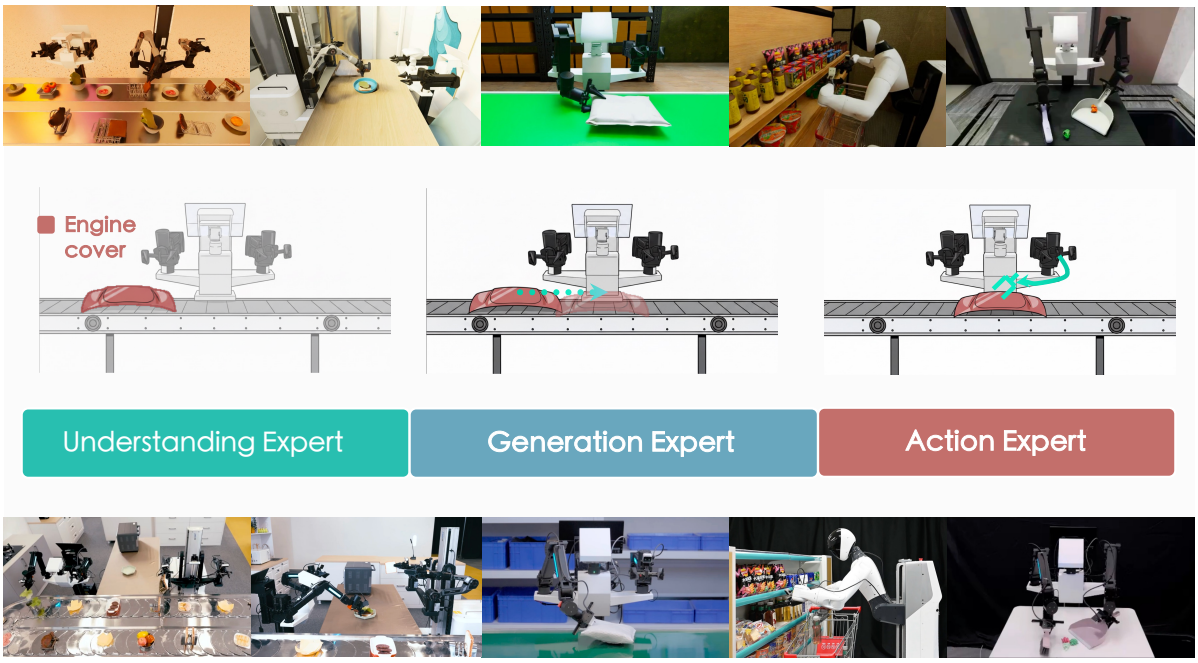

Figure 1. **InternVLA-A1** integrates understanding, generation, and action experts into a unified model, which synergizes semantic reasoning with dynamics prediction to guide action execution. By coupling this architectural design with pretraining on hybrid synthetic-real datasets, the model exhibits consistent robustness across diverse tasks, with 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). To realize generalist policies, the research 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). Built upon Multimodal Large Language Models (MLLMs) and trained on massive real-world robot demonstration datasets, 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, diverse robot datasets for skill learning. Consequently, achieving a generalist policy 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) and GR00T N1 (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 prediction errors. Consequently, developing a unified architecture that tightly couples semantic understanding with robust predictive dynamics remains crucial for scalable and 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 130,000 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 purely on real-world collection remains challenging: covering long-tail scene variations at scale is costly and inefficient, and high heterogeneity across embodiments further complicates joint training. For instance, even though π_0 was trained with 10,000 hours across 68 tasks from seven robot morphologies, achieving strong dexterous manipulation, its robustness against scene variations is still limited. Ultimately, end-to-end training at this scale incurs exceedingly high costs while delivering diminishing returns in long-tail scene generalization, making further scaling prohibitively expensive. 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, 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, imagination, and action. InternVLA-A1 combines the semantic reasoning of MLLMs with the prediction capability of a World Model-style imagination module. This design effectively bridges the gap between semantics and physical dynamics, facilitating foresight-aware action generation. Furthermore, to provide the diversity needed for robustness under scene variations, we build a *Data Pyramid* that composes complementary data sources: large-scale open-source robot data for broad scene understanding and physical priors, simulated synthetic data for scalable and controllable coverage of diverse scenes, and a small amount of specialized real-robot data for calibrating action execution in the real world. By incorporating simulation into training as a scalable source of variation, we can expand coverage over long-tail scene configurations at low cost. At the same time, we avoid the pitfalls of simulation-only learning by anchoring the policy with real-robot data, which helps correct transfer discrepancies arising from real-world sensing and physical interactions. We validate InternVLA-A1 through extensive experiments on real-world tasks and simulation benchmarks, and observe consistent improvements over strong VLA baselines. Specifically, InternVLA-A1 outperforms baselines like π_0 and GR00T N1.5 by a significant margin, showing a 14.5% increase in success rates for daily tasks and dominant performance in dynamic tasks, with improvements of 40% for Express Sorting and 73.3% for In-motion Ingredient Picking.

2. Related work

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

In **Vision-Language-Action models**, the field has shifted towards integrating the multimodal capabilities of foundation models with robotic control. RT-2 (Brohan et al., 2023) pioneered the co-finetuning of MLLMs on both robotic trajectories and Internet-scale data, enabling emergent reasoning capabilities. Diverging from RT-2's discrete token-based approach, π_0 (Black et al., 2024) combines a pre-trained MLLM with flow matching to leverage Internet-scale knowledge for controlling diverse robot embodiments. Building on this, $\pi_{0.5}$ (Black et al., 2025) integrates pre-training objectives with action prediction, thereby enhancing semantic inference for long-horizon tasks. To better synergize MLLM capabilities with real-time control, GR00T N1 (Bjorck et al., 2025) proposes a dual-system architecture that decouples reasoning from diffusion-based action generation. Its successor, GR00T N1.5, further improves instruction following and cross-embodiment compatibility. Distinct from these approaches, InternVLA-A1 unifies the semantic understanding of MLLMs with the dynamics-prediction capabilities, effectively bridging the semantic-dynamic gap prevalent in existing VLA architectures.

In **Video prediction and World models**, extensive research has focused on simulating real-world dynamics conditioned on actions 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. A growing body of video prediction policies—including CLOVER (Bu et al., 2024b), VPP (Hu et al., 2025), GR-1 (Wu et al., 2024a), GR-2 (Cheang et al., 2024) and $\mathcal{F}_1$ (Lv et al., 2025)—integrates

future state or video generation with action prediction, aiming to leverage the rich representations learned from large-scale video pre-training. To bridge the gap between action and visual spaces, Seer (Tian et al., 2025a) introduces an end-to-end inverse dynamics model that generates actions based on forecasted visual states. Meanwhile, the Genie Envisioner (Liao et al., 2025) enhances spatial understanding by utilizing a multi-view video diffusion model. 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 state prediction, thereby internalizing the causal relationship between actions and world dynamics.

In **Robotic manipulation data**, datasets fall into real-world and simulated categories. Real-world datasets (Bu et al., 2025; Khazatsky et al., 2024; Walke et al., 2023; Wu et al., 2024b) capture the intrinsic physical complexity of actual environments, serving as a cornerstone for current advancements by allowing models to learn behaviors end-to-end. For instance, the RT-1 and RT-2 training pipelines were built on approximately 130,000 teleoperated episodes (Brohan et al., 2022, 2023). The Open X-Embodiment dataset (O'Neill et al., 2024) aggregates one million episodes from multiple sources. Nevertheless, the significant heterogeneity inherent in this data poses new challenges for training generalizable VLA models. Simulated 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) leverages photorealistic rendering, procedural scene generation, and domain randomization to scale up data collection efficiently. For example, Robocasa (Nasiriany et al., 2024) produces demonstrations by combining human teleoperation with trajectory augmentation, yet the reliance on teleoperation still demands substantial manual effort. Alternatively, GraspVLA (Deng et al., 2025b) generates action data through an automated pipeline, though its scope is currently constrained to grasping tasks. The recently proposed InternData-A1 (Tian et al., 2025b) introduces a large-scale simulated demonstration dataset containing over 630k trajectories and 7,433 hours of data across diverse embodiments, skills, tasks, and scenes. Notably, models pre-trained on InternData-A1 have demonstrated performance comparable to π_0. In this work, we formulate a data recipe mixing InternData-A1 with real-world datasets, 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 representations then inform the generation expert, which simulates the task's evolution by predicting future states. Finally, the action expert combines these predictive dynamics with the semantic context, utilizing Flow Matching to produce precise robot control commands.

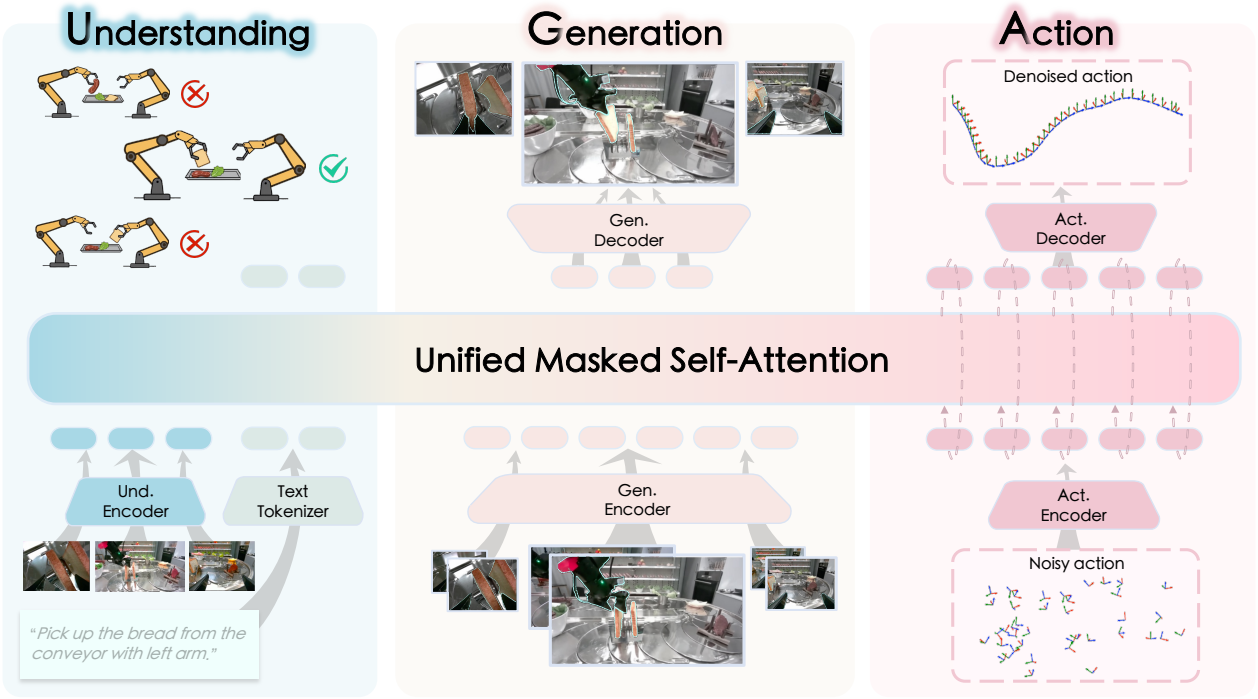

Figure 2. **Framework of InternVLA-A1**. The architecture comprises three experts: (1) an **understanding expert** that encodes scene context from image and text inputs; (2) a **generation expert** that predicts future visual states and task dynamics; and (3) an **action expert** that integrates the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

3.2. Unified Understanding-Generation-Action Transformer

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 concatenated to form prefix tokens $h_{\text{und}} = f_{\text{und}}(l, o_t)$ that condition downstream experts.

Generation Expert. 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. While understanding tasks demand high-level semantic abstraction typically captured by ViT-based encoders, generation tasks require preserving fine-grained spatial structure and pixel-level fidelity. To bridge this gap, 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 $\times$ 8 continuous VAE tokenizer (Agarwal et al., 2025) to encode input images into continuous latent features, providing a dense and structurally rich representation for the subsequent generative processes. The tokenizer comprises an encoder and a decoder; we denote its encoder as ϕ_{cosmos} and use the decoder to

reconstruct images from latents. Let $z_t = \phi_{\text{cosmos}}(o_t)$ denote the Cosmos latent at timestamp t; conditioned on h_{und} and $\{z_{t-m}, z_t\}$, the generation expert predicts $\hat{z}_{t+m}$.

While generative models have advanced significantly, applying them to the VLA domain presents a significant challenge: the need for high-frequency real-time inference. Traditional image and video generation models—whether based on diffusion (Blattmann et al., 2023a; Rombach et al., 2022) or next-token prediction (Sun et al., 2024)—are typically too slow for this purpose. Even optimized solutions like SANA-Sprint (Chen et al., 2025b) require 0.16 seconds per generation, restricting control to less than 6Hz. This limitation hinders fluid robotic movement, highlighting the urgent need for faster visual generation mechanisms in embodied AI. 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. Consequently, we have pursued a more aggressive design strategy to optimize performance.

We process images from three views across two timestamps (the current frame and one historical frame) using the Cosmos CI8 $\times$ 8 continuous VAE image tokenizer. With input images resized to 256×256, each is initially encoded into a 32×32 latent feature map. Directly feeding these latent tokens into the generation expert would result in an excessive sequence length, hindering both inference efficiency and training convergence. To mitigate this, we apply a convolutional layer to compress the spatial dimensions of each latent feature to 4×4 (representing each image with just 16 tokens) and use a projector to align them with the transformer's hidden dimension. These compressed tokens are then input into the generation expert, where they attend to the prefix tokens h_{und} (cached as K/V at inference) via multiple layers of masked self-attention. Their hidden states are pooled over time to form the predicted future latent $\hat{z}_{t+m}$, supervised against the Cosmos encoding z_{t+m} of frame $t+m$. Finally, the predicted future latent are upsampled back to the original VAE latent dimensions using a deconvolutional layer and a projector. These restored features are then fed into the Cosmos decoder to reconstruct the predicted images.

Action Expert. Conditioned on the language goal l, current observation o_t (via prefix tokens h_{und}), proprioception q_t, and the generation expert's predicted latent tokens $\hat{z}_{t+m}$, 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 $\rightarrow$ generation $\rightarrow$ action: tokens in a later block can attend to all earlier blocks, while earlier blocks cannot attend forward. The prefix block (vision + language) is fully bidirectional. The generation block is also fully bidirectional and only receives Cosmos latent tokens from frames $t-m$ and t. The action block is split into a state token and action tokens: the state token attends only to itself and earlier blocks, while action tokens attend to the state and to 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.

(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 $z_t = \phi_{\text{cosmos}}(o_t)$ denote the Cosmos latent. Given the current and historical observations $\{o_{t-m}, o_t\}$

and language instruction l, the generation expert predicts $\hat{z}_{t+m}$ conditioned on the understanding prefix h_{und} and is supervised by the frozen Cosmos latent z_{t+m}. This objective is formulated as:

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

where f_{gen} denotes the generation expert, ϕ_{cosmos} is the frozen Cosmos encoder used in z_t, and $\text{sg}[\cdot]$ indicates the stop-gradient operation. This objective compels the model to internalize physical dynamics, creating a robust prior that enhances action prediction.

(2) Flow Matching-based Action Prediction. To establish a principled connection between visual imagination and motor control, we employ a flow matching framework for action learning. This approach parameterizes the action generation process as learning continuous transformation pathways from noise to expert demonstrations, offering superior handling of multimodal action distributions compared to direct regression. Formally, 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:

$$\mathcal{L}_{\text{action}} = \mathbb{E}_{\{a_{t:t+k}, o_i, q_t, l\}\sim\mathcal{D}} \left[\left\| v_\theta(l, \{o_i\}^t_{i=t-m}, q_t, a^{\tau}_{t:t+k}) - (a_{t:t+k} - \epsilon) \right\|^2 \right], \tag{2}$$

where q_t denotes the proprioception state at time t. 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:

$$a^{\tau+\Delta\tau}_{t:t+k} = a^{\tau}_{t:t+k} + \Delta\tau \cdot v_\theta(l, \{o_i\}^t_{i=t-m}, q_t, a^{\tau}_{t:t+k}), \tag{3}$$

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:

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

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`. 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 Hyperparameters. We train InternVLA-A1 in two stages: a large-scale pre-training stage on heterogeneous robotic 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. All experiments are conducted using bfloat16 precision. The hyperparameter λ, which balances the two loss components, is 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.

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

Load-balanced Parallel Training. InternVLA-A1 is trained on a mixture of heterogeneous robot datasets. At this scale, naively instantiating the full dataset mixture on every worker can trigger out-of-memory failures 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

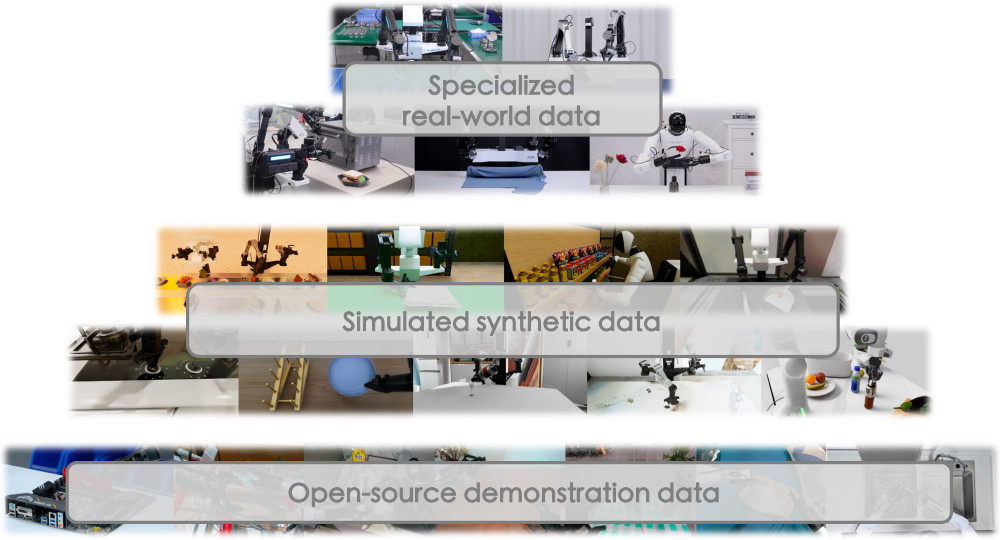

Figure 3. **The hierarchical data pyramid**.

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. Hierarchical data pyramid

This study proposes a hierarchical data pyramid framework, as presented in Figure 3. This data pyramid strategically integrates diverse data sources across training stages. The foundation is built during pre-training, where we utilize large-scale open-source robot demonstrations alongside high-fidelity synthetic data to establish broad generalizability. This is followed by a post-training phase that leverages specialized real-world data, refining the model's capabilities for precise physical deployment.

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. 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.

4.3. Open-source robot demonstration

While synthetic data excels in scalability, real-world demonstrations remain essential for capturing authentic 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). This dataset contains over 1 million trajectories covering 217 tasks across five domains, including domestic, retail, industrial, restaurant, and office scenarios. By incorporating this large-scale real-world demonstration dataset 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. Specialized real-world data

At the apex of our data pyramid, we utilize specialized real-world data to fine-tune the model for specific downstream applications. Collected via teleoperation in targeted environments, these high-quality, task-centric episodes—though smaller in scale—are highly aligned with the target domain. Post-training on this data effectively adapts the model's broad pre-trained knowledge to the precise nuances of the deployment environment.

4.5. Pre-training data recipe

We pretrain the model on a mixture of large-scale simulated synthetic data and open-source real-world demonstration data. The synthetic portion is drawn from InternData-A1, which covers four robot embodiments (Genie-1, Franka, ARX Lift-2, and AgileX) with diverse task families and long-horizon compositions. We additionally include AgiBot-World as real-world data to complement synthetic trajectories with physical interaction variability. During pretraining, we interleave trajectories from different sources using configurable sampling weights (Table 3).

Table 3. Data mixture used for pretraining.

Data source	Num. frame	Frame proportion
InternData-A1 (ARX Lift-2)	96M	0.18
InternData-A1 (AgileX)	122.5M	0.23
InternData-A1 (Franka)	90.5M	0.17
InternData-A1 (Genie-1)	16M	0.03
AgiBot-World (Beta)	208M	0.39

5. Experiments

To evaluate the effectiveness of our proposed InternVLA-A1, we conduct extensive experiments on twelve real-world tasks and a simulation benchmark. We first compare InternVLA-A1 with existing leading vision-language-action models π_0 and GR00T N1.5. Then, we conduct an ablation study to verify the effectiveness of pretraining and the proposed generation expert.

5.1. Evaluation Configuration

Hardware & platform. We benchmark our policy on three physical robot embodiments: **Genie-1**, **ARX Lift-2**, and **ARX AC One**. These platforms cover diverse bimanual manipulation capabilities and

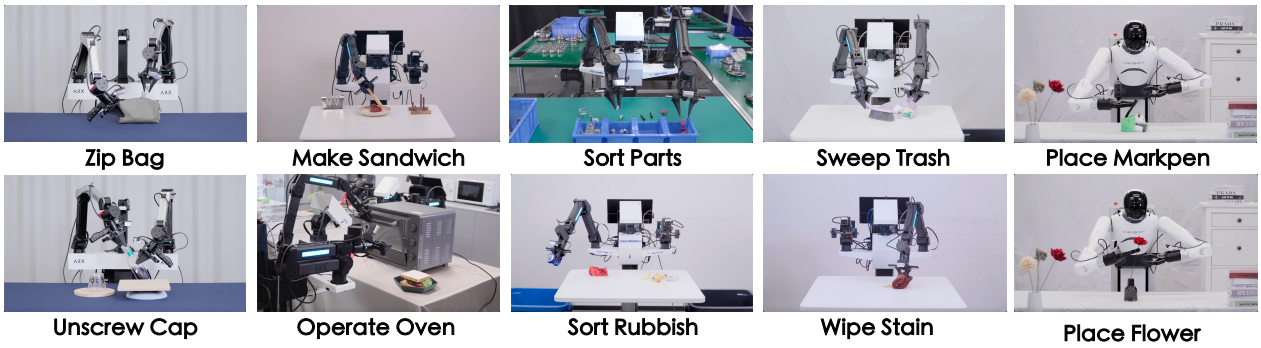

Figure 4. **The experimental setting of general-purpose tasks**.

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 on real hardware with the same deployment pipeline across the three robots, and we report task success rates based on repeated rollouts.

Task setup. We design a real-world task suite that covers both **general-purpose** manipulation and **specialized** high-dynamics scenarios. The suite includes **ten** general-purpose tasks and **two** specialized task families:

- **General-purpose tasks:**
  - **Dexterous manipulation:** *Unscrew Cap*, *Zip Bag*, *Sort Parts*.
  - **Regular manipulation:** *Operate Oven*, *Make Sandwich*, *Sort Rubbish*, *Sweep Trash*, *Wipe Stain*, *Place Markpen*, *Place Flower*.
- **Specialized tasks (high dynamics):** *In-motion Ingredient Picking tasks* and *Express Sorting tasks*.

Overall, the tasks span articulated-object interaction and contact-rich manipulation (e.g., cap removal and zipper closure), long-horizon bimanual routines (e.g., sweeping/disposing and multi-object sorting), and, importantly, high-dynamics settings where the scene evolves 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 (and grasp success when applicable) 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 initial conditions.

5.2. Evaluation on the General-purpose Tasks

To evaluate the performance of InternVLA-A1 against leading VLA models, including π_0 and GR00T N1.5, we conducted a comprehensive evaluation across 10 diverse real-world tasks. These tasks, as displayed in Figure 4, range from dexterous manipulation (e.g., Unscrew Cap, Zip Bag, Sort Parts) to regular manipulation (e.g., Make Sandwich, Operate Oven, Sort Rubbish).

As shown in Table 4, InternVLA-A1 demonstrates superior performance compared to the prior state-of-the-art models, π_0 (3.3B) and GR00T N1.5 (3B). Notably, our smaller InternVLA-A1 (2B) model 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 efficiency of our architecture and the quality of our data pyramid, enabling a lighter model to outperform a larger parameter counterpart. When scaling to

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
π_0	40.0	**66.7**	**53.3**	66.7	43.3
GR00T N1.5	33.3	0.0	6.7	46.7	16.7
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
73.3	96.0	73.3	53.3	40.0	60.6
46.7	66.7	40.0	40.0	33.3	33.0
53.3	**97.3**	80.0	**66.7**	**66.7**	64.7
86.7	**97.3**	**86.7**	**66.7**	60.0	**75.1**

a comparable size, InternVLA-A1 (3B) yields a substantial performance improvement, reaching an average success rate of 75.1%—a 14.5% absolute improvement over π_0. In regular manipulation tasks, InternVLA-A1 (3B) exhibits exceptional robustness; it achieves dominant leads in complex tasks like Make Sandwich (93.3% vs. 66.7%) and Operate Oven (86.7% vs. 73.3%), while maintaining top-tier performance on Sort Rubbish (97.3%), matching the strong baseline set by π_0. Furthermore, in challenging dexterous manipulation tasks, InternVLA-A1 (3B) equals the strongest baseline in Unscrew Cap and Sort Parts while significantly surpassing it in Zip Bag (73.3% vs. 40.0%). These results confirm that InternVLA-A1 not only excels in high-level task planning but also possesses superior fine-grained control capabilities.

5.3. Evaluation on the Specialized Tasks

To further validate the performance of our proposed InternVLA-A1, we designed two tasks involving manipulation in dynamic scenes: Express Sorting and In-motion Ingredient Picking. The experimental settings for the Express Sorting and In-motion Ingredient Picking tasks are illustrated 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 wherein 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. 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, the experimental results reveal that state-of-the-art baselines, π_0 and GR00T N1.5, struggle significantly in dynamic manipulation scenarios, yielding consistently low success rates. Specifically, these models fail to exceed 40.0% on the Express Sorting task and achieve only 20.0% on the In-motion Ingredient Picking task. In stark contrast, InternVLA-A1 (2B) and InternVLA-A1 (3B) demonstrate superior stability and robustness across both settings. The 2B model achieves success rates of 70.0% and 46.7%, respectively. Most notably, InternVLA-A1 (3B) exhibits exceptionally strong performance, attaining 80.0% in Express Sorting and a remarkable 93.3% in In-motion Ingredient Picking. This represents a substantial margin over the baselines, particularly in the multi-robot coordination task, where the 3B model outperforms π_0 and GR00T N1.5 by over 70 percentage points.

5.4. Evaluation on Simulation Benchmark

We evaluate our method on the RoboTwin 2.0 benchmark (Chen et al., 2025c), which comprises 50 bimanual tasks. The evaluation includes two protocols: the Easy setting, identical to the demonstration

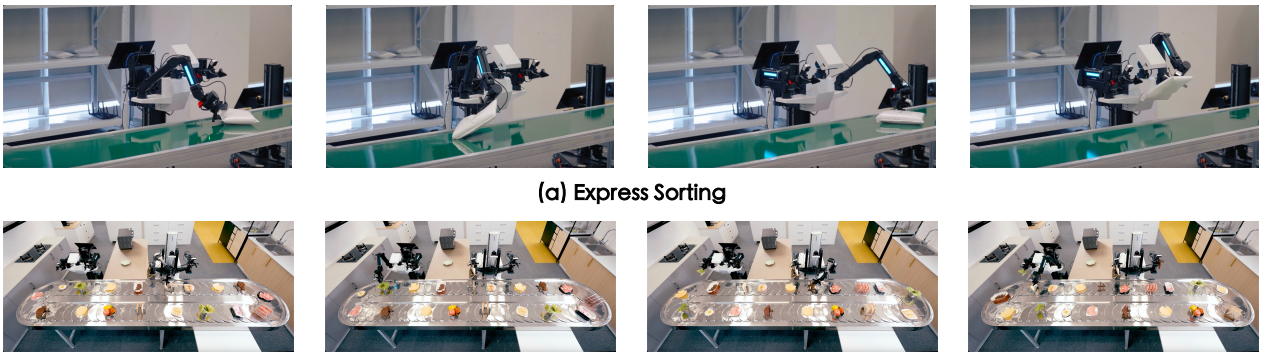

Figure 5. **The experimental setting of specialized tasks: (a) Express Sorting task, (b) In-motion Ingredient Picking task.**

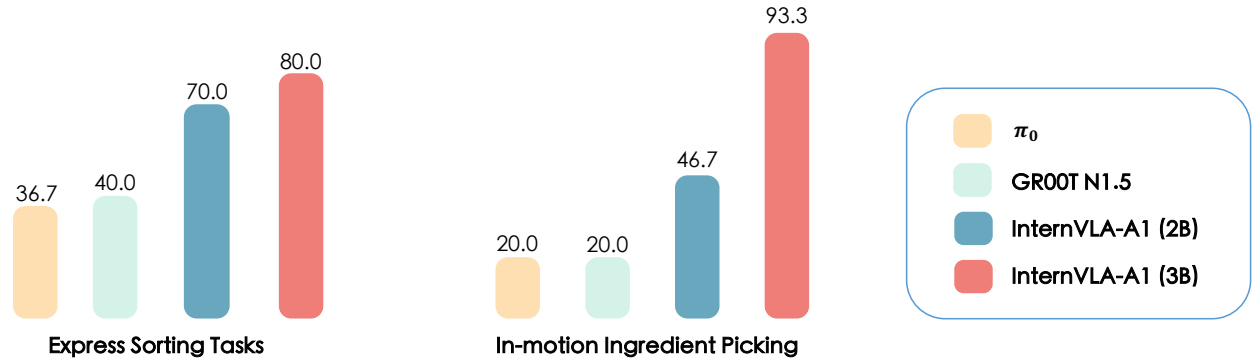

Figure 6. **The experimental results of Express Sorting Tasks and In-motion Ingredient Picking tasks.**

Table 5. **Evaluation on RoboTwin Simulation Benchmark.** The best results are **bolded**.

Method	Hanging Mug		Move Pillbottle Pad		Pick Dual Bottles		Place Object Stand	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
π_0	9.0%	5.0%	30.0%	0.0%	58.0%	16.0%	47.0%	14.0%
InternVLA-A1(2B)	17.0%	**13.0%**	54.0%	**8.0%**	**70.0%**	**25.0%**	67.0%	25.0%
InternVLA-A1(3B)	**24.0%**	**13.0%**	**69.0%**	**8.0%**	45.0%	23.0%	**70.0%**	**38.0%**

Blocks Ranking Size		Shake Bottle		Stamp Seal		Place Burger Fries		Avg. (50 Tasks)	
Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
19.0%	1.0%	96.0%	53.0%	16.0%	3.0%	66.0%	34.0%	54.5%	19.8%
27.0%	7.0%	**99.0%**	64.0%	**42.0%**	6.0%	**99.0%**	10.0%	58.6%	20.5%
26.0%	**8.0%**	**99.0%**	**79.0%**	39.0%	**12.0%**	96.0%	**52.0%**	**65.0%**	**25.4%**

environment, and the Hard setting, featuring domain randomization. For each task, we fine-tuned the pre-trained model using 100 demonstrations and conducted 100 evaluation trials for both settings. Table 5 demonstrates that InternVLA-A1 surpasses the π_0 model across both the 2B and 3B variants. Specifically, the InternVLA-A1(3B) model yields performance gains of 10.5% in the easy setting and 5.6% in the hard setting.

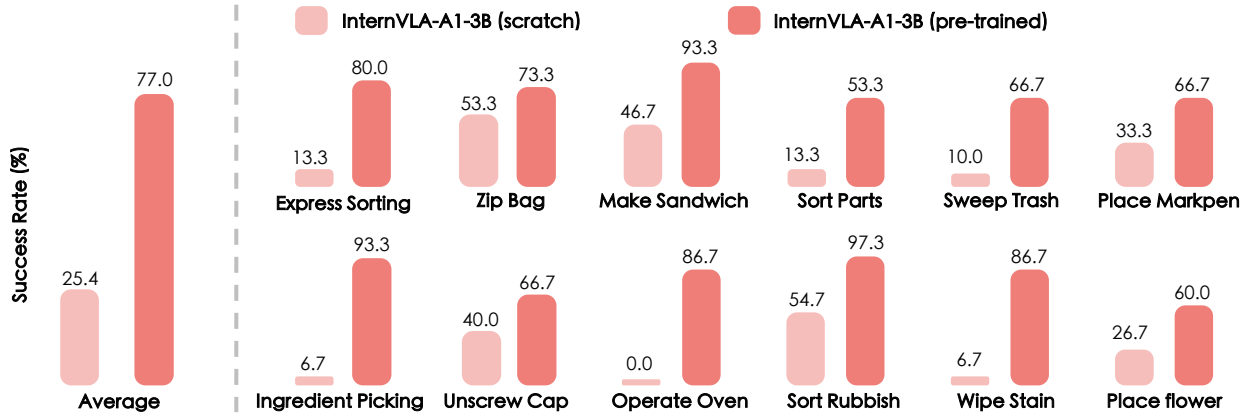

Figure 7. **The ablation studies on pre-training**.

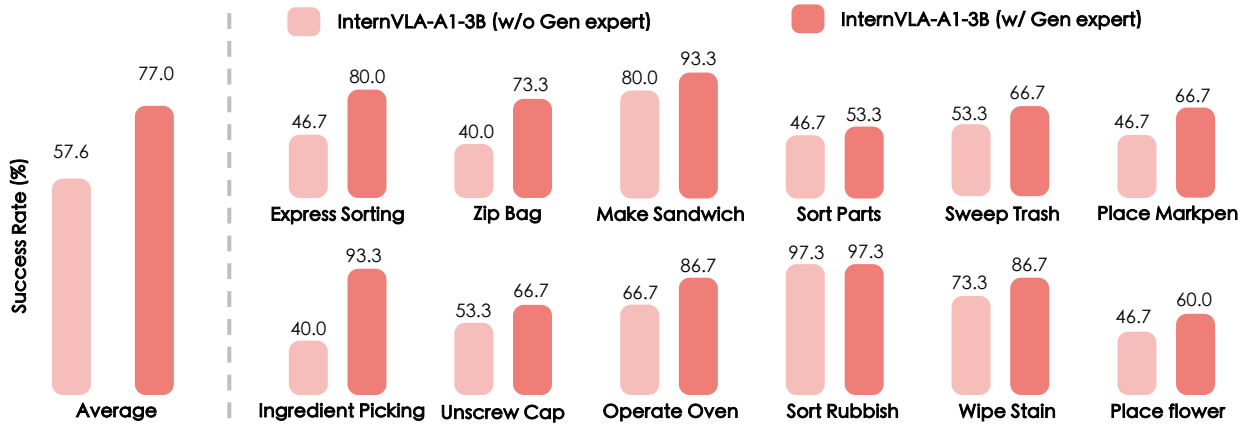

Figure 8. **The ablation studies on generation expert**.

5.5. Ablation Studies

Effectiveness of Pre-training. As shown in Figure 7, 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 our model retained high proficiency. This finding suggests that pre-training acts as a crucial inductive prior, effectively stabilizing the optimization process and allowing the model to inherit foundational manipulation skills. As a result, our model achieves a significantly higher success rate on downstream tasks by building upon the robust capabilities acquired during pre-training.

Effectiveness of Generation Expert. To evaluate the contribution of the generation expert, we conduct a key ablation study by comparing InternVLA-A1 (3B) with that without the generation expert. The experimental results are shown in Figure 8. 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%. This ablation study validates the superiority of the proposed generation expert and the unified architecture integrating understanding, generation, and action.

Impact of Pre-training Datasets. As shown in Figure 9, we evaluate the efficacy of different pre-

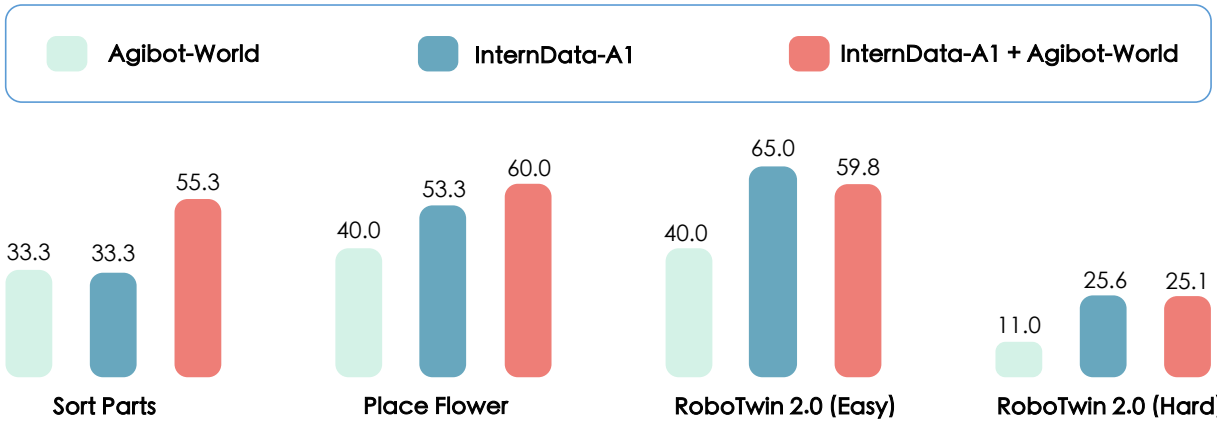

Figure 9. **The ablation studies on the pre-training dataset**.

training data sources: the large-scale real-world dataset Agibot-World, our high-fidelity synthetic dataset InternData-A1, and a mixed strategy combining both. First, InternData-A1 demonstrates exceptional transfer capabilities. Despite being synthetic, the model pre-trained solely on InternData-A1 matches or surpasses the real-data-only baseline, confirming InternData-A1's high fidelity for supporting pre-training. Second, InternData-A1 dominates in simulation benchmarks. On the RoboTwin simulation evaluation, the pure InternData-A1 model achieves the highest performance, outperforming even the mixed pre-training setting. We attribute this to two factors: (1) the minimal visual domain gap between the training and evaluation environments (Sim-to-Sim), and (2) the diverse range of heterogeneous embodiments within InternData-A1, which provides better adaptability to the target robot compared to the specific embodiment found in Agibot-World. Finally, mixed pre-training synergizes complementary strengths. While adding real-world data causes a slight performance decline in simulation compared to pure synthetic training, the mixed strategy yields substantial gains in real-world scenarios, particularly in fine-grained manipulation tasks such as Sort Parts, where we observe a 20% improvement. This confirms that our hybrid approach effectively synergizes the scale and diversity of simulation with the physical realism of real-world data.

6. Conclusion and Limitations

In this work, we presented InternVLA-A1, a VLA model with enhanced robustness against environmental dynamics. We achieved this through dual advancements in model architecture and data paradigms. Architecturally, InternVLA-A1 employs a Mixture-of-Transformers (MoT) design to unify semantic understanding, visual foresight, and action prediction, effectively synergizing high-level reasoning with low-level dynamics. Complementing this, our hybrid synthetic-real pre-training strategy combines the scene diversity of simulation with the physical fidelity of real-world data. Extensive evaluations demonstrate that InternVLA-A1 significantly outperforms leading baselines such as π_0 and GR00T, exhibiting exceptional robustness, particularly in highly dynamic scenarios. Furthermore, ablation studies validate the critical contributions of the proposed generation expert, the unified architecture, and the complementary efficacy of our hybrid data strategy.

Limitations. Despite these advancements, two primary limitations remain. First, the understanding expert currently lacks joint training with multimodal VQA datasets, leading to a degradation in general semantic reasoning and complex instruction-following capabilities. Second, to ensure efficient inference for the visual foresight module, we compromised the fidelity of image prediction, which limits the granularity of the generated future frames. Future work will address these two limitations.

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. Advances in neural information processing systems, 35:23716–23736, 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.

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.

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.

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 Proceedings of the Computer Vision and Pattern Recognition Conference, pages 12187–12198, 2025.

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 The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=b_CQDy9vrD1.

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.

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 Proceedings of the Computer Vision and Pattern Recognition Conference, 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 Forty-second International Conference on Machine Learning, 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.

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. arXiv preprint arXiv:2403.12945, 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 The Thirteenth International Conference on Learning Representations, 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 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 Proceedings of the Computer Vision and Pattern Recognition Conference, 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. Appendix

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

Core Contributors

Junhao Cai[1], Yang Li[1], Jiangmiao Pang[1†], Zherui Qiu[1], Yang Tian[1], Jia Zeng[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], Haoxiang Ma[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], Hongrui Zhu[1], Yangkun Zhu[1], Yuchen Zhu[2]

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