Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

LDA-1B: Scaling Latent Dynamics Action Model via Universal Embodied Data Ingestion

Jiangran Lyu[*1,2], Kai Liu[*2,3,4], Xuheng Zhang[*1,2], Haoran Liao[2,6], Yusen Feng[1,2], Wenxuan Zhu[1], Tingrui Shen[1], Jiayi Chen[1,2], Jiazhao Zhang[1,2], Yifei Dong[1], Wenbo Cui[2,3,4], Senmao Qi[2], Shuo Wang[2], Yixin Zheng[2,3,4], Mi Yan[1,2], Xuesong Shi[2], Haoran Li[3], Dongbin Zhao[3], Ming-Yu Liu[7], Zhizheng Zhang[2,†], Li Yi[5,†], Yizhou Wang[1,†], He Wang[1,2,†]

[1]Peking University [2]Galbot [3]CASIA [4]BAAI [5]Tsinghua University [6]Sun Yat-sen University [7]NVIDIA

Code & Data: https://pku-epic.github.io/LDA * Equal contribution † Corresponding authors

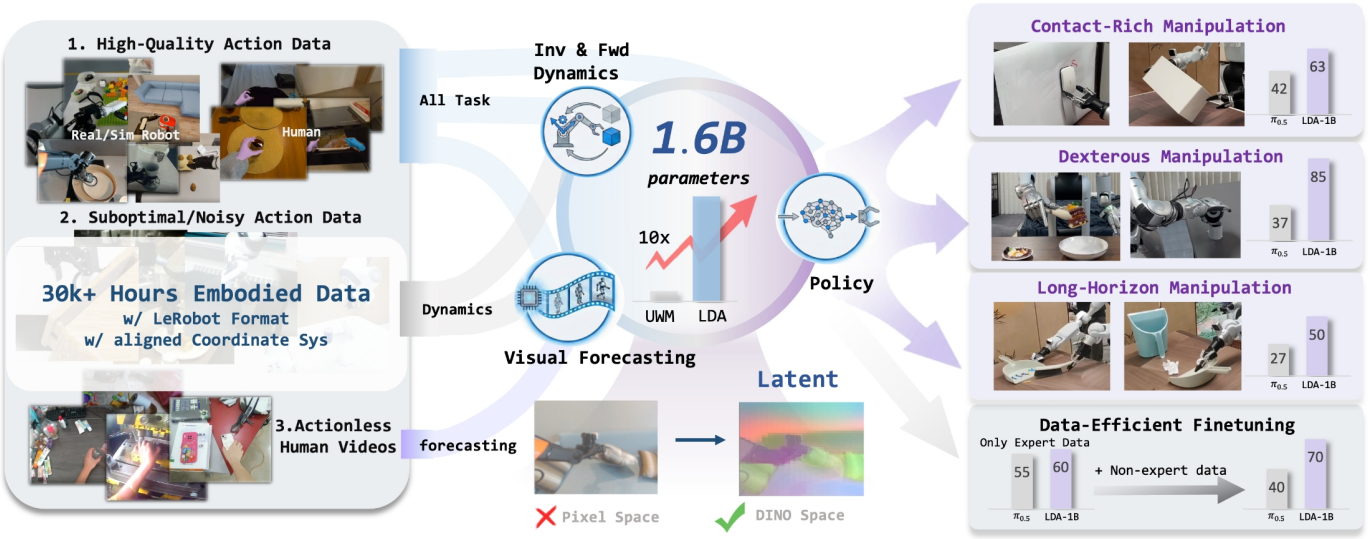

Fig. 1: We introduce LDA-1B, a 1.6 B-parameter robot foundation model scaled on over 30k hours of heterogeneous embodied data. LDA-1B unifies policy, dynamics, and visual forecasting in a structured DINO [34] latent space, allowing different data sources to play complementary roles. Beyond high-quality data alone, noisy data and actionless videos also provide valuable visual and physical priors for dynamics learning. This universal data ingestion paradigm enables stable scaling with data and model size, significantly outperforming strong baselines such as $\pi_{0.5}$ [15] across diverse manipulation tasks.

Abstract*—Recent robot foundation models largely rely on large-scale behavior cloning, which imitates expert actions but discards transferable dynamics knowledge embedded in heterogeneous embodied data. While the Unified World Model (UWM) formulation has the potential to leverage such diverse data, existing instantiations struggle to scale to foundation-level due to coarse data usage and fragmented datasets. We introduce LDA-1B, a robot foundation model that scales through universal embodied data ingestion by jointly learning dynamics, policy, and visual forecasting, assigning distinct roles to data of varying quality. To support this regime at scale, we assemble and standardize EI-30k, an embodied interaction dataset comprising over 30k hours of human and robot trajectories in a unified format. Scalable dynamics learning over such heterogeneous data is enabled by prediction in a structured DINO latent space, which avoids redundant pixel-space appearance modeling. Complementing this representation, LDA-1B employs a multimodal diffusion transformer to handle asynchronous vision and action streams, enabling stable training at the 1B-parameter scale. Experiments in simulation and the real world show LDA-1B outperforms prior methods (e.g., $\pi_{0.5}$) by up to 21%, 48%, and 23% on contact-rich, dexterous, and long-horizon tasks, respectively. Notably, LDA-1B enables data-efficient fine-tuning, gaining 10% by leveraging 30% low-quality trajectories typically harmful and discarded.

I. Introduction

Inspired by the success of Large Language Models (LLMs) and Vision-Language Models (VLMs), the robotics community has increasingly pursued general-purpose robot foundation models through large-scale pretraining [4, 30]. Most existing approaches center on scaling behavior cloning (BC), which imitates expert actions but fundamentally restricts learning to high-quality demonstrations. Consequently, a large portion of heterogeneous embodied data [31] is discarded or only weakly utilized, despite containing rich physical interaction dynamics [18].

Unified World Model (UWM) formulation [20, 41] provides an alternative by jointly optimizing dynamics, policy, and video generation within a single model, which can leverage not only expert data. Despite the potential value, existing UWM instantiations remain far from scaling to foundation-level. A major limitation lies in coarse data usage: heterogeneous embodied data are often treated uniformly, without differentiating their roles by quality or supervision, which underutilizes transferable dynamics knowledge. In addition, the community lacks ready-to-use large-scale datasets that

unify varying-quality data with consistent formats and aligned action representations. Furthermore, UWM represents future state in pixel space, entangling dynamics learning with redundant appearance modeling. Subtle variations in illumination, texture, background clutter, or camera viewpoint can dominate the training objective, making large-scale training inefficient and hindering the learning of interaction-relevant dynamics.

To overcome these limitations, we introduce **LDA-1B**, a robot foundation model that scales via *universal embodied data ingestion*. In this framework, heterogeneous data play distinct yet complementary roles: actionless human videos supervise visual forecasting [26, 25, 17], lower-quality trajectories primarily inform dynamics learning, and high-quality trajectories support both policy and dynamics. To realize this approach at scale, we assemble **EI-30k**, a large-scale embodied interaction dataset with over 30k hours of human and robot trajectories across real and simulated environments, standardized in format and aligned in action representation. Scalable learning on such diverse data is facilitated by a structured **DINO latent space** [34, 40, 14], which reduces redundant appearance modeling [26, 17], and a multimodal diffusion transformer that aligns asynchronous visual and action prediction. By combining this ingestion strategy, dataset, latent representation, and model architecture, LDA-1B achieves stable training at the 1B-parameter scale while maximizing data utilization.

We evaluate LDA-1B on challenging RoboCasa-GR1 benchmark and a diverse set of real-world tasks involving both grippers and high-DoF dexterous hands [38]. LDA-1B consistently outperforms $\pi_{0.5}$, achieving **21%** gains on contact-rich manipulation, benefiting from improved dynamics understanding and **48%** gains on dexterous manipulation, benefiting from effective utilization of human data. Moreover, under a mixed-quality fine-tuning setting, LDA-1B improves data efficiency by **10%** through leveraging low-quality trajectories that are detrimental to baseline methods. These results highlight universal embodied data ingestion and unified latent dynamics learning as a scalable alternative to behavior-cloning-centric robot pretraining. In summary, our contributions are threefold:

- We propose LDA-1B, a scalable robot foundation model that learns generalizable interaction dynamics through unified latent dynamics pretraining.
- We construct EI-30k, a large-scale embodied interaction dataset covering diverse embodiments, environments, data qualities, with aligned end effector coordinate system.
- We demonstrate that LDA-1B achieves superior generalization and robustness across a wide range of settings, including simulation and real-world environments, contact-rich manipulation, dexterous manipulation, and long-horizon manipulation.

II. Related Work

Robot Foundation Models. Recent robot foundation models predominantly adopt the Behavior Cloning paradigm. As summarized in Table I, representative approaches including π_0 [3], RDT [22], and InternVLA [9] rely heavily on high-quality teleoperation or simulation data, which fundamentally constrains their scalability. Hybrid methods such as Being-H0 [23] and UniVLA [6] attempt to incorporate heterogeneous data with mixed quality; however, they largely depend on action alignment or auxiliary pretrained latent action models, limiting the effective data scale to around 6k hours of embodied data. In contrast, LDA-1B breaks this ceiling by adopting a unified world model formulation, enabling efficient ingestion of up to 30k hours of mixed-quality embodied data.

Model	Data Src.	#Data	Action Quality	Train.	Param.
$\pi_{0.5}$ [15]	Tele.	10k+	High	BC	3B
RDT [22]	Tele.	<10k	High	BC	1B
GraspVLA [11]	Sim.	20k+	High	BC	2B
InternVLA-M1 [9]	Sim.	<10k	High	BC	3B
Being-H0 [23]	Hum.	<10k	Mixed	Aln. + BC	14B
InternVLA-A1 [7]	Het.	10k+	High	VF + BC	3B
GR00T-N1.6 [29]	Het.	<10k	Mixed	LA + BC	1B
UniVLA [6]	Het.	<10k	Mixed	LA + BC	7B
LDA-1B	Het.	**30k+**	Mixed	UWM [41]	1B

TABLE I: **Comparison of Representative Robot Foundation Models.** This table compares the proposed LDA with recent robot foundation models in terms of data source, data quantity, action quality, training paradigm, and the number of trainable model parameters (excluding frozen components). Data source abbreviations are as follows: Tele.=teleoperation, Sim.=simulation, Hum.=human demonstration, and Het.=heterogeneous data. Training paradigm abbreviations include: BC=behavior cloning, VF=visual foresight, Aln.=alignment, LA=latent action modeling, and UWM=unified world model. Only embodied interaction data are considered, excluding internet-scale VQA data.

Unified Video Action Models. Recent works have explored joint modeling of dynamics and policy for embodied decision making. Methods such as DyWA [24], FLARE [39], DreamVLA [37] and the WorldVLA series [8, 16] demonstrate that co-training next-state prediction and policy learning can improve generalization in interactive environments. To enrich dynamics modeling, UWM [41] and UVA [20] further propose optimizing multiple objectives jointly, including video generation, forward and inverse dynamics, and action prediction. Concurrent with our work, Motus [2] adopts the UWM paradigm and integrates priors from pretrained VLM and video generation models. Despite their promising results, these approaches typically operate directly in pixel space and do not explicitly consider the roles of data quality, scale, or heterogeneity during training, which limits their ability to fully exploit large-scale, mixed-quality interaction data for robust dynamics learning.

Large-Scale Embodied Interaction Datasets. The progress in embodied AI relies on large-scale embodied datasets. Many widely used datasets are collected via teleoperation on real robots [3, 15, 19, 21] or generated in simulation [11, 9], providing high-quality action-labeled trajectories. Beyond robot-

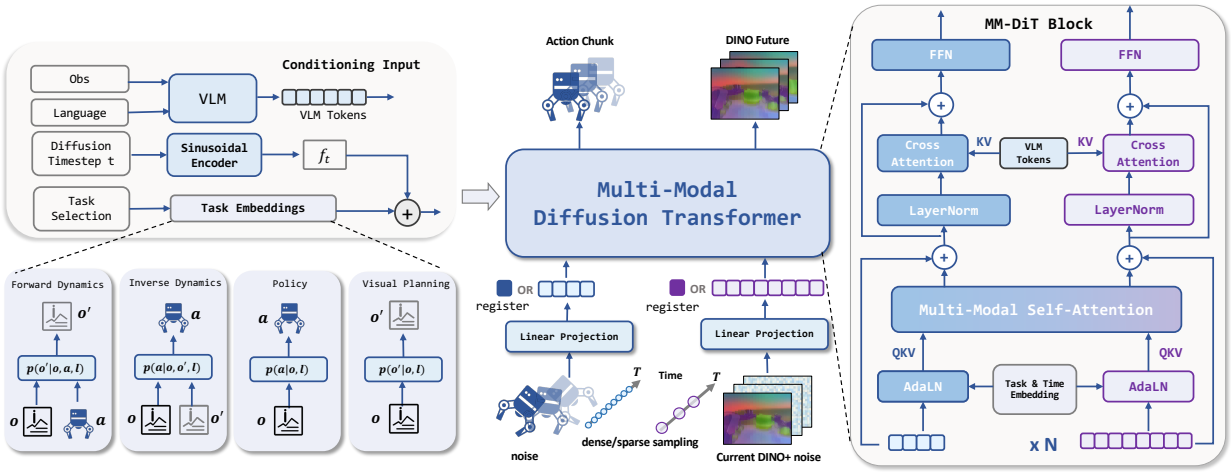

Fig. 2: **Architecture of LDA.** LDA jointly denoises action chunks and future visual latents under multiple co-training objectives, including policy learning, forward dynamics, inverse dynamics, and visual forecasting. Conditioned on VLM tokens, diffusion timesteps, and task embeddings, the model adopts a multimodal diffusion transformer architecture, where action and visual experts are decoupled and interact through a shared self-attention layer.

collected data, recent works explore human-centric embodied datasets, such as egocentric recordings with hand actions [35, 23, 12]. While these datasets significantly expand data diversity, many are either not publicly released or provide limited action supervision, making them difficult to directly integrate with robot learning pipelines. More broadly, existing embodied datasets are highly fragmented: some are closed-source, others are open but vary substantially in data formats, sensor configurations, action representations, and annotation quality. This lack of standardization poses a major obstacle to large-scale data aggregation and unified training. In contrast, our work introduces EI-30k, a large-scale embodied interaction dataset that unifies diverse data sources including robot and human trajectories from both real-world and simulated environments under consistent data formats and aligned action representations.

III. Latent Dynamics Action Model

A. Preliminary: Unified World Models

Given the current observation o_t (typically an RGB image), UWM [41] jointly models multiple conditional distributions over future observations $\boldsymbol{o}_{t+1:t+k}$ and action chunk $\boldsymbol{a}_{t+1:t+k}$, enabling unified learning of:

1) Policy: $p(\boldsymbol{a}_{t+1:t+k} \mid \boldsymbol{o}_t)$
2) Forward Dynamics: $p(\boldsymbol{o}_{t+1:t+k} \mid \boldsymbol{o}_t, \boldsymbol{a}_{t+1:t+k})$
3) Inverse Dynamics: $p(\boldsymbol{a}_{t+1:t+k} \mid \boldsymbol{o}_{t:t+k})$
4) Visual Planning: $p(\boldsymbol{o}_{t+1:t+k} \mid \boldsymbol{o}_t)$

Concretely, UWM [41] instantiates this framework using a joint diffusion model that predicts noise for both actions and future observations:

$$(\epsilon_a^\theta, \epsilon_o^\theta) = s_\theta\big(o,\, a_{t_a},\, o'_{t_o},\, t_a,\, t_{o'}\big)\,,$$

where t_a and t_o are independently sampled diffusion timesteps for actions and observations, and $\tilde{a}_{t_a}$, $\tilde{o}_{t_o}$ denote their corresponding noisy inputs. The model is trained with a standard DDPM [13] objective, jointly denoising future actions and observations conditioned on o_t. We further extend this formulation by introducing language ℓ conditioning through a VLM, enabling instruction-guided action and observation prediction.

B. Universal Data Ingestion via Multi-task Co-training

We adopt a *universal data ingestion* regime to jointly train the unified objectives described above, allowing heterogeneous embodied data to contribute according to their supervision quality. Specifically, high-quality robot and human demonstrations are co-trained with all objectives, supporting both action policy learning and dynamics modeling. Lower-quality trajectories, which may contain suboptimal or noisy actions, are used exclusively for dynamics and visual forecasting, where accurate action optimality is not required. In addition, we leverage large-scale human manipulation videos without action annotations to train the visual forecasting objective, providing supervision for instruction-conditioned future state prediction. This role-aware data usage prevents overfitting to expert-only behaviors and enables scalable learning of transferable dynamics and action representations.

To implement differentiated objectives within a single diffusion model, we introduce four learnable *task embeddings* and two learnable *register tokens*. Each task embedding corresponds to a specific training objective (policy, forward dynamics, inverse dynamics, or visual forecasting) and is added to the diffusion timestep embedding f_t to condition the denoising process. The learnable register tokens, one for action and one for visual state, serve as placeholders for modalities that are absent in a given task. For example, during policy training, the model receives noisy action tokens along with a visual register token representing the unobserved future state; in contrast, visual forecasting uses noisy future visual tokens with an action register token. This design enables a unified architecture to flexibly support different input-output structures without

modifying the network topology. Overall, the model predicts a denoising vector field v_a^θ under different task conditions and is trained using a flow-matching objective:

$$
\begin{aligned}
l^\theta_{\text{action}} &= \mathbb{E}_{\substack{(\boldsymbol{o}_{t:t+k},\boldsymbol{a}_{t+1:t+k},\ell)\sim\mathcal{D} \\ \tau_a\sim\mathcal{U}(0,T_\tau) \\ \epsilon_a\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}} \left\| v_a^\theta - (\epsilon_a - \boldsymbol{a}_{t+1:t+k}) \right\|_2^2, \\
l^\theta_{\text{obs}} &= \mathbb{E}_{\substack{(\boldsymbol{o}_{t:t+k},\boldsymbol{a}_{t+1:t+k},\ell)\sim\mathcal{D} \\ \tau_o\sim\mathcal{U}(0,T_\tau) \\ \epsilon_o\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}} \left\| v_o^\theta - (\epsilon_o - \boldsymbol{o}_{t+1:t+k}) \right\|_2^2, \\
l^\theta &= l^\theta_{\text{action}} + l^\theta_{\text{obs}}.
\end{aligned}
\tag{1}
$$

During training, action and visual losses are selectively activated according to the task specification, allowing heterogeneous data to contribute under appropriate supervision. At inference time, the same model can be flexibly invoked for different objectives by specifying the task embedding and corresponding inputs.

C. Representation of Predictive Targets

We represent predictive targets, future visual states and actions, in a unified format to maximize knowledge sharing across heterogeneous datasets. For visual prediction, we adopt latent features extracted from a pretrained DINO [34] encoder, rather than VAE-based pixel-space representations. DINO latents encode high-level semantic and spatial structure while suppressing background noise and low-level visual variations, which facilitates learning scene dynamics that generalize across diverse environments and object configurations.

For actions, we define a unified hand-centric action space based on end-effector motion, consisting of delta wrist poses and finger configurations. For parallel-jaw grippers, the finger state is represented by a single degree-of-freedom gripper width, while for multi-finger dexterous hands, finger configurations are described using keypoints expressed in the wrist coordinate frame. This design enables consistent action modeling across different embodiments and manipulation platforms.

To model temporal dynamics, visual states and actions are organized as two synchronized temporal streams with different sampling rates. Visual observations are sampled at 3 Hz, a lower frequency than actions, 10 Hz. This reduces redundant computation from highly correlated consecutive frames while preserving fine-grained action dynamics, allowing the model to maintain coherent temporal alignment between fast-varying control signals and slower-evolving visual states.

D. Architecture: MM-DiT

We adopt a Multimodal Diffusion Transformer (MM-DiT) to jointly denoise action chunks and predict future visual features within a unified diffusion framework (Fig. 2). The model operates on heterogeneous tokens while sharing a common Transformer backbone. Conditioning inputs include the current observation, language instruction, diffusion timestep, and task specification. Observations and language are encoded by a pretrained VLM into conditioning tokens. The diffusion timestep is encoded using a sinusoidal embedding, and task information is represented by a learned task embedding. All conditioning signals are injected into each Transformer block via adaptive layer normalization (AdaLN [32]).

Actions are organized as fixed-length chunks and corrupted with Gaussian noise. Future visual features (DINO [34] features) are noised in parallel. Both modalities are projected into token embeddings through modality-specific linear layers and processed jointly by MM-DiT. Each MM-DiT block applies multimodal self-attention over concatenated action and visual tokens, enabling cross-modal interaction. Modality-specific QKV projections and FFNs are retained to preserve inductive biases, while attention is shared across modalities. Language tokens are incorporated via cross-attention to provide high-level semantic guidance. Finally, modality-specific output heads predict denoised action sequences and future visual features.

E. Pre-training and Post-training

Pre-training Configurations. Our model is trained on a server cluster equipped with 48 NVIDIA H800 GPUs. The training process contains 400k iterations, resulting in a total computational cost of 4,608 GPU hours. To preserve the generalization capability and visual representation quality of the pre-trained foundation models, we keep the parameters of the VLM [**?**] and the DINO [34] encoder frozen throughout the pre-training process, updating the MM-DiT and action encoder/decoder. This design ensures that the model can learn from new data without degrading the core abilities of the base models in cross-modal understanding and fine-grained visual feature extraction.

Data-Efficient Fine-tuning. To adapt the model to target embodiments and tasks for real-world deployment, we introduce a lightweight post-training stage. This stage follows the same data regime as pretraining and effectively leverages naturally collected teleoperation data of mixed quality, without requiring expert-level demonstrations. Compared to prior fine-tuning pipelines that rely on carefully curated expert datasets, our method directly utilizes unfiltered teleoperation data, substantially improving data efficiency and reducing the cost of data collection and annotation, thereby facilitating practical deployment.

IV. Embodied Interaction Dataset (EI-30k)

We introduce the Embodied Interaction Dataset (EI-30k), a large-scale collection of embodied interaction trajectories totaling over 30k hours. It consists of 8.03k hours of real-world robot data, 8.6k hours of simulated robot data, 7.2k hours of human demonstrations with actions, and 10k hours of actionless human videos. All subdatasets are annotated with explicit quality labels, enabling systematic analysis across different fidelity levels and supporting quality-aware learning.

Data Unification. EI-30k consolidates datasets from heterogeneous platforms and tasks, which vary in storage formats, sensor modalities, and annotations. All data are converted into the LeRobot format, providing a unified representation of observations, actions, and language. This standardization facilitates plug-and-play training, flexible data composition, and

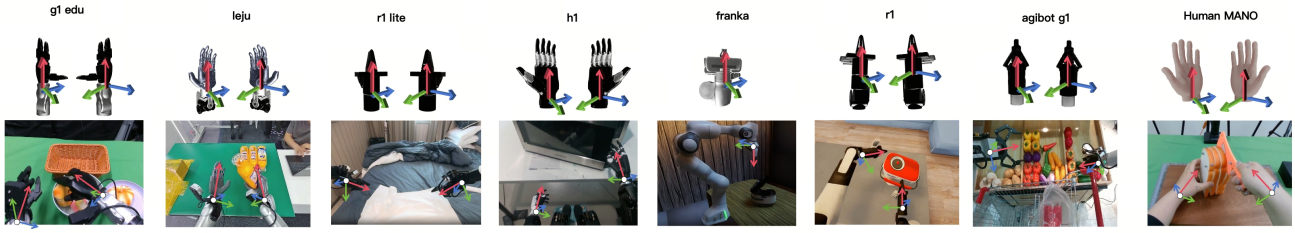

Fig. 3: **Aligned End Effector Coordinate Systems.** We manually align coordinate frames across diverse robot and human embodiments to ensure consistency. This shared representation enables joint learning from heterogeneous interaction data.

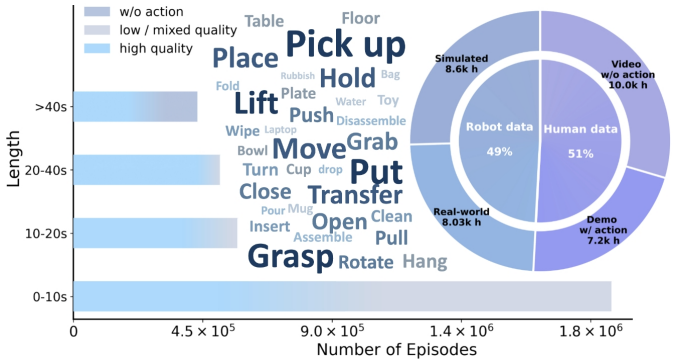

Fig. 4: **Statistics of EI-30k.** The dataset contains more than 30k hours of diverse human and robot interaction data (right). It spans varying episode lengths (left) and a rich set of manipulation tasks (center).

Model	Vis. Rep.	MMDiT	VLM	Success Rate ↑
GR00T-N1.6 [29]	-	-	Cosmos [1, 28]	47.6
StarVLA [10, 36]	-	-	Qwen3-VL [**?**]	47.8
GR00T-EI10k	-	-	Qwen3-VL	51.3
UWM-0.1B [41]	VAE	✗	-	14.2
UWM-1B	VAE	✗	Qwen3-VL	19.3
UWM(MM-DiT)	VAE	✓	Qwen3-VL	20.0
LDA(DiT)	DINO	✗	Qwen3-VL	48.9
LDA-0.5B	DINO	✓	Qwen3-VL	50.7
LDA-1B	DINO	✓	Qwen3-VL	**55.4**

TABLE II: Results on RoboCasa-GR1 [27] and impact of state representation (VAE vs. DINO [34]), model size, and the MM-DiT architecture on task success rates.

seamless integration of additional annotations, while greatly reducing engineering overhead for handling diverse sources.

Aligned Action Representation. To support consistent modeling of physical interactions across embodiments, all available action annotations are expressed as hand-centric motion in a shared coordinate frame (Fig. 3). For robots, this includes the 6-DoF end-effector pose plus gripper width or dexterous hand joints. For humans, the 6-DoF wrist pose and full MANO [33] hand parameters are recorded. Camera extrinsics are retained to decouple hand motion from egocentric head motion. All coordinate frames are manually aligned to ensure geometric consistency across datasets, enabling joint learning from both human and robot trajectories.

Quality Annotation and Cleaning. EI-30k applies systematic cleaning and quality-aware annotation. Language annotations are normalized using a vision-language model to ensure semantic consistency. Motion segments without meaningful hand-object interaction are removed, e.g., head-only or idle segments in egocentric videos. Each trajectory is assigned a quality label based on action accuracy, and annotation completeness. Unlike aggressive filtering, low-quality trajectories are preserved, allowing downstream models to exploit the full spectrum of data through quality-aware training.

V. Experiments

A. Simulation Experiments

Benchmark and Baselines. We evaluate our method on RoboCasa-GR1 [27], a simulated kitchen benchmark featuring 24 tabletop rearrangement and articulated-object manipulation tasks with the GR-1 humanoid robot and Fourier dexterous hands. The benchmark provides challenging and realistic settings that require high-DoF dexterous manipulation from egocentric RGB observations captured by a head-mounted camera. Following the GR00T [29] evaluation protocol, we finetune all models using 1,000 trajectories per task and evaluate each task with 51 trials, reporting average success rates. We compare LDA against GR00T and its strong variants, as well as UWM [41], under matched training paradigms and data. To ensure a fair comparison in terms of model capacity and pretraining, we reproduce a strong GR00T baseline (denoted as GR00T-EI10k) with 1B parameters, pretrained on our curated EI-30k high-quality subset and using Qwen3-VL as the VLM encoder.

Comparison with Baselines. As shown in Table II, the original GR00T-N1.6 [29] with 3B parameters achieves a success rate of 47.6%. When pretrained on our curated EI-30k dataset, the reproduced GR00T-EI10k with 1B parameters shows a clear improvement, reaching 51.3%, highlighting the impact of high-quality embodied data. Under the same parameter budget, LDA further improves the success rate to 55.4%. These results indicate that, beyond data quality and parameter scaling, jointly learning actions and dynamics within a unified model provides additional gains when pretrained on mixed-quality data.

Ablation Study. We further analyze key design choices under identical training data and optimization settings. UWM [41], despite jointly predicting actions and dynamics, achieves only 14.2% success due to limited model capacity and the use

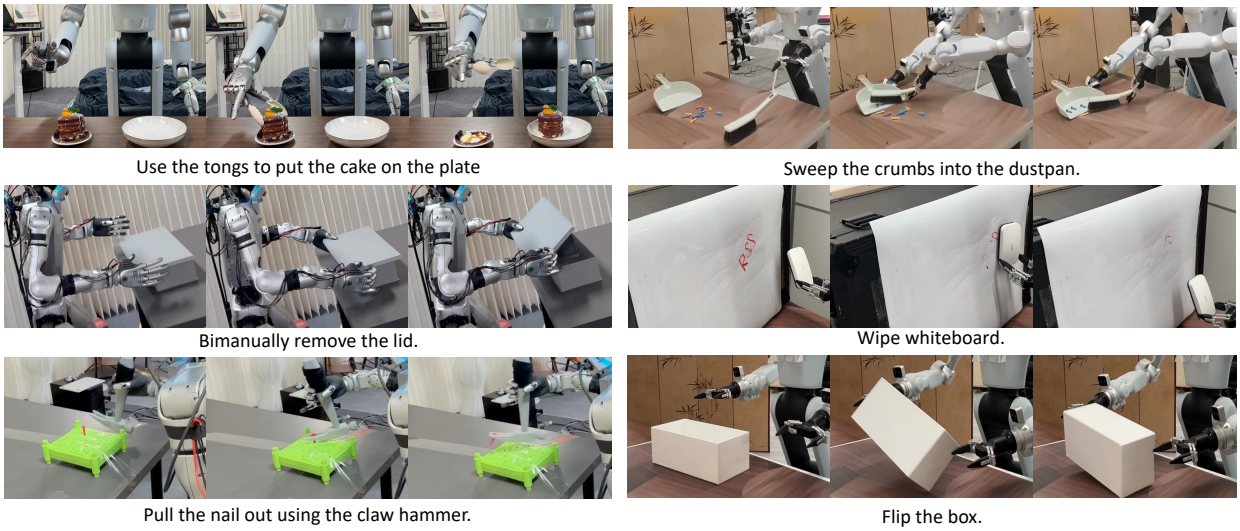

Fig. 5: **Real-World Manipulation Demonstrations Across Multiple Robotic Platforms and End-Effectors.** Galbot G1 equipped with a Sharpa dexterous hand (top-left), Unitree G1 with a BrainCo dexterous hand (middle and bottom-left), and Galbot G1 with a two-finger gripper (right).

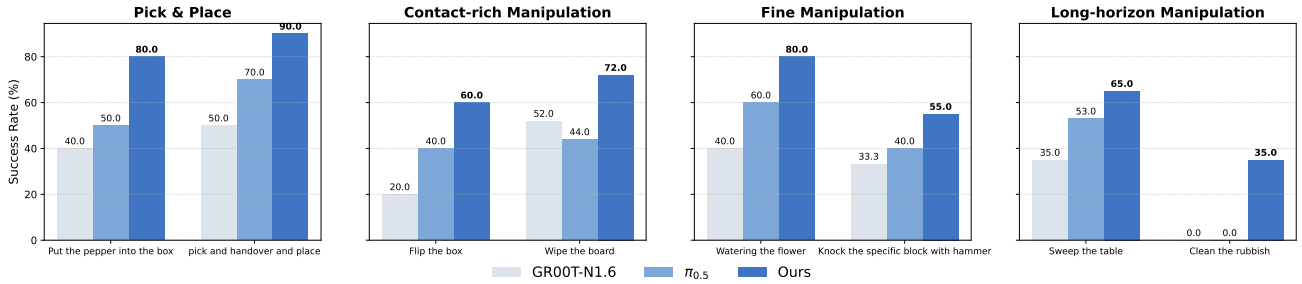

Fig. 6: **Success Rate Comparison on Real-World Gripper Manipulation Tasks.** All models are few-shot fine-tuned on Galbot and evaluated on eight tasks spanning Pick & Place, Contact-rich, Fine, and Long-horizon manipulation. LDA consistently outperforms GR00T-N1.6 [29] and $\pi_{0.5}$ [15].

of entangled VAE latent representations. Scaling UWM to 1B parameters or replacing its DiT backbone with our MM-DiT yields only marginal improvements (19.3% and 20.0%, respectively), suggesting that architectural constraints fundamentally limit its performance. In contrast, replacing pixel-space VAE latents with DINO [34] representations leads to a substantial performance gain (20.0% $\rightarrow$ 55.4%), highlighting the importance of semantically structured latent spaces for effective scaling. Finally, removing the proposed MM-DiT architecture or reducing the model size to 0.5B parameters results in performance drops of 6.5% and 4.7%, respectively, confirming the effectiveness of the multi-expert design and its favorable scaling behavior.

B. Real-world Experiments

To validate the scalability and robustness of LDA-1B, we conduct extensive real-world experiments focusing on few-shot adaptation to new embodiments, dexterous manipulation, and data efficiency under mixed-quality supervision.

Real-World Robot and Task Setup. We evaluate our method on two humanoid platforms: Galbot G1 and Unitree G1. Galbot G1 is equipped with either a two-finger gripper or 22-DoF Sharpa dexterous hands, while Unitree G1 uses 10-DoF BrainCo hands. Across all configurations, the policy receives only egocentric RGB observations from a head-mounted camera. We evaluate four categories of manipulation tasks under the gripper setting, *Pick and Place*, *Contact-rich Manipulation*, *Fine Manipulation*, and *Long-horizon Manipulation*, covering diverse contact dynamics and temporal horizons. Representative tasks include *Beat Block, Flip Box, Handover, Pick-and-Place (Pepper), Sweep Table, Clean Rubbish, Water Flower,* and *Wipe Board*. Dexterous manipulation further includes tool-use tasks such as pulling a nail with a hammer and flipping bread with a spatula, which require precise force control and coordinated finger motion. Qualitative demonstrations are shown in Fig. 5. For each task, we collect 100 teleoperated trajectories without enforcing expert-level execution. As a result, the dataset naturally exhibits mixed quality: approximately 50–80% of trajectories correspond to expert behavior, while the remainder contain suboptimal actions such as pauses, retries, or inefficient motion patterns.

Baselines and Fine-tuning Protocol. We compare LDA-1B against two strong baselines, $\pi_{0.5}$ [15] and GR00T [29]. To ensure stable and competitive performance, baseline models are

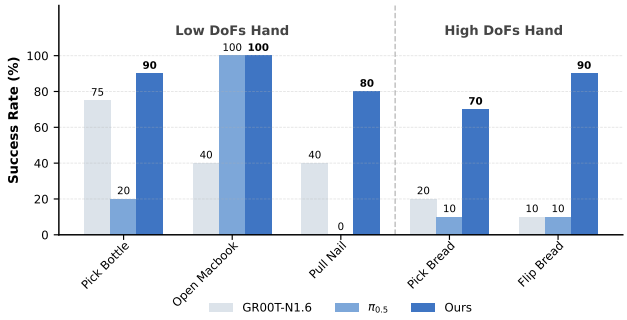

Fig. 7: **Success Rate Comparison on Real-World Dexterous Manipulation Tasks.** We evaluate the real-world performance of our model against baselines (GR00T-N1.6 and $\pi_{0.5}$) on 3 low-DoF hand (BrainCo) tasks and 2 high-DoF hand (Sharpa) tasks. Ours (dark blue) consistently outperforms baselines, especially on fine dexterous tasks (pulling nails) and high-DoF tasks.

Method	Pick & Place		
	Object	Background	OOD Pos.
$\pi_{0.5}$	26.7	20.0	6.7
GR00T	40.0	40.0	20.0
Ours	**60.0**	**60.0**	**40.0**

TABLE III: **Robust Generalization under visual and spatial perturbations.** LDA-1B achieves 60.0% success on unseen objects and backgrounds, and 40.0% under OOD positions, demonstrating effective focus on task-critical affordances over visual noise through latent dynamics pretraining.

finetuned exclusively on the filtered expert subset. In contrast, LDA-1B leverages all collected trajectories and learns directly from the full mixed-quality distribution via our Universal Embodied Data Ingestion mechanism.

Results on Gripper Manipulation. We first evaluate few-shot adaptation by deploying LDA-1B on the Galbot G1, which is excluded from our EI-30k pretraining dataset. As shown in Fig. 6, LDA-1B consistently outperforms all baselines across task categories. On simple pick-and-place tasks, LDA-1B achieves success rates of 80%–90%, indicating effective few-shot adaptation to a new robot embodiment. The performance gap widens substantially in contact-rich and long-horizon scenarios. For instance, the *Clean the Rubbish* task requires coordinated dual-arm manipulation, tool usage (dustpan), and sequential object transfer into a trash bin, where errors can easily accumulate over time. In this setting, LDA-1B achieves a 35% success rate, while both GR00T and $\pi_{0.5}$ fail entirely (0%). This result suggests that latent dynamics modeling enables LDA to better anticipate action-induced state transitions, maintain temporal consistency, and recover from intermediate failures in extended manipulation sequences.

Results on Dexterous Manipulation. We further evaluate LDA-1B on both low-DoF and high-DoF dexterous manipulation tasks, as reported in Fig. 7. On low-DoF tasks such as *Pull Nail*, which requires precise motion direction and stable contact maintenance between the hammer and the nail, LDA-1B achieves 80% success, reliably localizing targets and adjusting sensitive actions, whereas $\pi_{0.5}$ largely fails. On high-DoF tasks such as *Flip Bread*, which involve high-dimensional control, continuous contact, and coordinated wrist motion, LDA-1B attains 90% success, while $\pi_{0.5}$ reaches only 10%. These results demonstrate that pretraining on large-scale human data provides strong latent priors for dexterous control, enabling precise finger coordination and object reorientation with limited robot data. In contrast, baseline policies struggle to generalize as action dimensionality and contact complexity increase.

Method	Place the pen into the box		Bimanually remove the lid	
	63 High	63 High + 37 Low	66 High	66 High + 34 Low
$\pi_{0.5}$	60	40 (20↓)	50	40 (10↓)
Ours	**70**	**80 (10↑)**	**50**	**60 (10 ↑)**

TABLE IV: **Data-efficient mixed-quality fine-tuning.** LDA-1B improves success rates by +10% on both tasks when incorporating low-quality trajectories, while $\pi_{0.5}$ degrades significantly, demonstrating effective utilization of noisy data for enhanced generalization.

Generalization Ability. To evaluate the generalization of our policy, we test pick and place task under three conditions: novel objects, unseen backgrounds, and out-of-distribution (OOD) starting position, shown in Fig. 8. As summarized in Table III, our model maintains high success rates despite visual and spatial perturbations. The large-scale latent dynamics pretraining allows the model to ignore visual distractors (background changes) while focusing on relevant object affordances, demonstrating strong generalization relative to baselines.

Data-Efficient Fine-tuning. We analyze the value of mixed-quality data ingestion during the fine-tuning stage by post-training on two splits: (1) High-Quality Only (expert data), and (2) High + Low Quality (all 100 trajectories). As shown in Table IV, while baseline models degrade when low-quality data is added, LDA-1B effectively leverages these noisy trajectories, boosting performance by 10 percentage points, substantially improving data efficiency and reducing the cost of data collection and annotation for practical deployment.

C. Analysis of Design Choices for Model Scaling

To analyze the scaling behavior of LDA, we systematically vary model capacity, data composition, and training objectives. All models are evaluated on an unseen test set sampled from a held-out subset of *Agibot World* [5]. We report the action prediction L1 error as the primary metric, which serves as a stable and reproducible proxy for real-world performance. Fig. 10 summarizes the results under four training configurations: (i) *Policy Only*, (ii) *Policy + Visual Forecasting*, (iii) *Policy with Forward and Inverse Dynamics*, and (iv) the full co-training framework (*Ours*). These experiments jointly reveal how LDA scales under heterogeneous supervision and increasing model capacity.

Fig. 8: Generalization evaluation setup on Pick and Place task

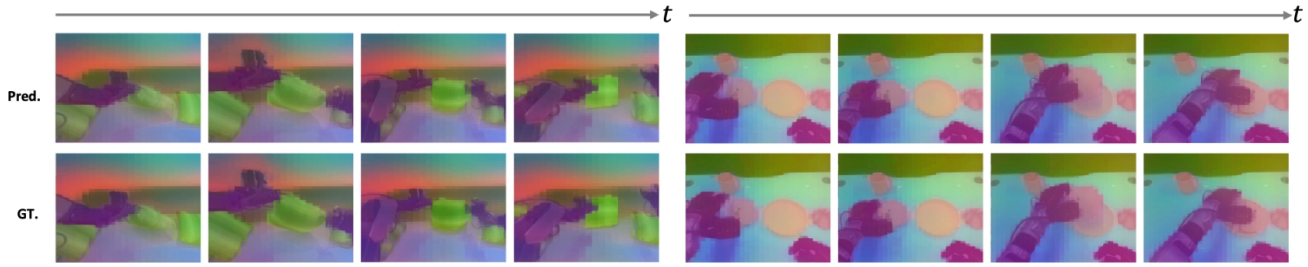

Fig. 9: **Visualization of latent forward dynamics.** Our model generates accurate future visual representations (top) aligned with ground truth (bottom) across time steps, capturing semantic object structure and motion dynamics.

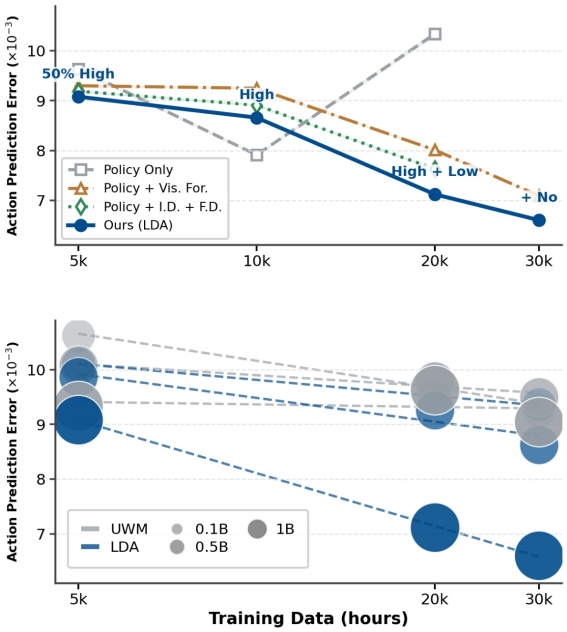

Fig. 10: Scaling Analysis of LDA, evaluated by action prediction error on an unseen test set. Top: Action prediction error decreases to 6.6 with 30k hours of training data, demonstrating effective utilization of diverse data sources. Bottom: LDA consistently outperforms UWM across model sizes (0.1B→1B) with increasing training data.

Effectiveness of Universal Data Ingestion. Effectively leveraging heterogeneous embodied data requires jointly scaling both data sources and training objectives. As shown in Fig. 10, LDA achieves its best performance only when all supervision signalspolicy learning, dynamics modeling, and visual forecastingare optimized together. When either the data scale or the training objectives are reduced, performance degrades noticeably. Using only action-labeled trajectories with a *Policy Only* objective (grey line), increasing the dataset size yields unstable behavior: while moderate scaling initially reduces error, incorporating lower-quality data leads to performance degradation. Similarly, partial co-training variants that exclude either dynamics or visual forecasting objectives (green and brown lines) improve robustness but fail to fully exploit the available data. In contrast, the full co-training framework (blue line) exhibits consistent improvement as additional heterogeneous data is introduced. Notably, even after all action-labeled trajectories are exhausted, adding 10k actionless videos continues to reduce prediction error. These results indicate that LDA can extract useful supervisory signals from low-quality data and non-action data through latent dynamics and visual forecasting, rather than treating such data as noise. Overall, these results demonstrate that Universal Data Ingestion is most effective when heterogeneous data and co-training objectives are scaled together, enabling LDA to fully utilize mixed-quality supervision.

Effectiveness of Latent Representation. Although both LDA and UWM incorporate dynamics-related supervision, their scaling behaviors diverge substantially due to differences in the structure of their latent spaces. As shown in Fig. 10, UWM quickly saturates as data scale and model capacity increase, with additional supervision yielding diminishing or even negative returns. This indicates that simply increasing data or parameters is insufficient when the latent space cannot support compositional and causal reasoning. This limitation stems from UWM's VAE-derived latent representation, which entangles appearance, geometry, and dynamics at a low-level feature granularity. Such entanglement restricts the models ability to factorize action-induced state transitions and prevents effective reuse of heterogeneous supervision during scaling. In contrast, LDA operates in a semantically structured latent space obtained from large-scale visual pretraining. This representation preserves object-level semantics and spatial coherence, enabling dynamics learning to scale smoothly with increased model capacity, richer training objectives, and more diverse datasets.

Effectiveness of Model Size Scaling. Beyond data scale, LDA

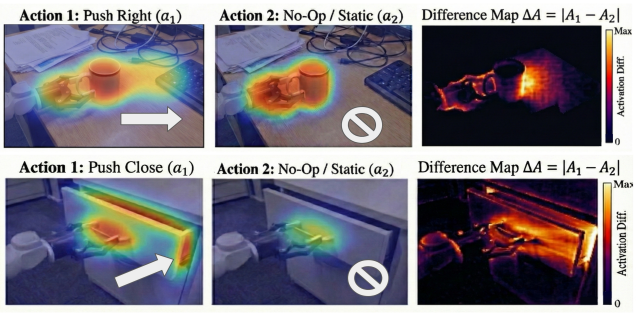

Fig. 11: Attention HeatMap: "Push Right" (top) highlights the mug's leading edge and trajectory; "Push Close" (bottom) concentrates on the contact surface. The model attends exclusively to movable regions while ignoring irrelevant background clutter.

exhibits consistent and predictable improvements as model capacity increases. As shown in Fig. 10, scaling the model from 0.1B to 0.5B and further to 1B parameters leads to monotonic reductions in action prediction error under the full co-training framework. This indicates that LDA can effectively absorb additional capacity to model increasingly complex action-dynamics relationships when sufficient heterogeneous supervision is available. The results highlight a promising scaling paradigm in which model capacity, training objectives, and heterogeneous embodied data are jointly aligned, enabling reliable performance gains.

D. Analysis of Multi-task Learning

Qualitative Analysis of Latent Forward Dynamics. Beyond quantitative prediction errors, we qualitatively examine the forward dynamics learned by LDA, visualized via PCA projections of DINO feature embeddings. As shown in Fig. 9, the model produces coherent future-state predictions that respect physical constraints such as object permanence, contact continuity, and motion consistency under the applied action. Notably, the predicted dynamics focus on task-relevant objects while remaining invariant to visual distractors that do not influence the control loop. This suggests that LDA learns a dynamics-aware latent world model, capturing how actions causally propagate through the scene rather than merely extrapolating visual appearance.

Action-Conditioned Attention. To interpret how LDA reasons about action-induced state transitions, we visualize attention maps conditioned on different action primitives. As shown in Fig. 11, we compare the attention patterns induced by an active motion command (a_1) with those under a static *No-Op* command (a_2), and compute their difference to reveal action-specific visual grounding. Across tasks, LDA consistently attends to regions that are causally relevant to the commanded interaction. In the *Push Right* scenario, the attention difference highlights the leading edge of the mug and the anticipated motion direction, reflecting awareness of object displacement. In the *Push Close* task, attention concentrates on the drawer surface where contact and force application are expected.

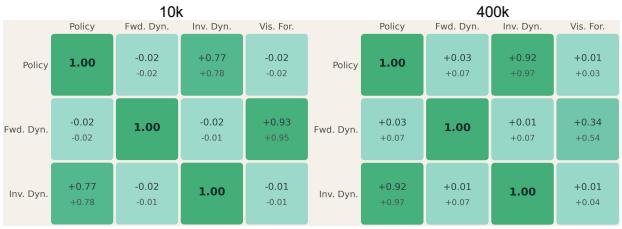

Fig. 12: Gradient cosine similarities between objectives at 10k and 400k iterations.

Importantly, background clutter and visually salient but non-interactive regions are largely suppressed. These results indicate that LDA conditions visual attention on the physical consequences of actions, selectively focusing on regions that drive state transitions rather than static appearance.

Gradient Similarity Analysis. We measure gradient cosine similarities among training objectives of the action expert at 10k and 400k iterations. Fig. 12 shows that objectives with matched targets are strongly aligned from early training and become more consistent over time: policy and inverse dynamics both predict actions, while video generation and forward dynamics both predict DINO features. Objectives with different prediction targets show weak negative correlations at 10k iterations, but this early competition disappears by 400k iterations and turns into positive alignment. This trend suggests that LDA gradually organizes heterogeneous objectives into compatible feature subspaces.

VI. Conclusion, Limitations, and Future Directions

We present **LDA-1B**, a robot foundation model that scales latent dynamics learning via universal embodied data ingestion. By assigning heterogeneous data distinct roles and leveraging over **30k hours** of human and robot trajectories in the EI-30k dataset, LDA-1B learns dynamics in a structured DINO latent space and employs a mixed-frequency multimodal diffusion transformer, enabling stable training at the **1B-parameter** scale. Experiments show strong performance across diverse manipulation and long-horizon tasks, as well as data-efficient fine-tuning on imperfect trajectories. Limitations include the reliance on fixed DINO visual features and predominantly egocentric camera viewpoints, which may constrain generalization to new visual perspectives and multimodal signals. Future work includes jointly learning visual representations and latent dynamics, extending to richer sensory modalities, automatically optimizing data roles, and fostering broader community adoption of scalable, heterogeneous data-driven robot foundation models.

Acknowledgments

This work was partially supported by New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122905). We thank Caowei Meng for teleoperation data, Haoran Liu and Jiayi Su for early exploration assistance, Yu-Wei Chao and Shengliang Deng for discussions, and Junkai Zhao for experimental equipment.

References

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

[2] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

[3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[5] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Xindong He, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025.

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

[7] Junhao Cai, Zetao Cai, Jiafei Cao, Yilun Chen, Zeyu He, Lei Jiang, Hang Li, Hengjie Li, Yang Li, Yufei Liu, et al. Internvla-a1: Unifying understanding, generation and action for robotic manipulation. *arXiv preprint arXiv:2601.02456*, 2026.

[8] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

[9] Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, et al. Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy. *arXiv preprint arXiv:2510.13778*, 2025.

[10] StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026.

[11] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Wenhao Zhang, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025.

[12] Yankai Fu, Ning Chen, Junkai Zhao, Shaozhe Shan, Guocai Yao, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Metis: Multi-source egocentric training for integrated dexterous vision-language-action model. *arXiv preprint arXiv:2511.17366*, 2025.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

[14] Yuhang Huang, Jiazhao Zhang, Shilong Zou, Xinwang Liu, Ruizhen Hu, and Kai Xu. Ladi-wm: A latent diffusion-based world model for predictive manipulation. *arXiv preprint arXiv:2505.11528*, 2025.

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

[16] Yuming Jiang, Siteng Huang, Shengke Xue, Yaxi Zhao, Jun Cen, Sicong Leng, Kehan Li, Jiayan Guo, Kexiang Wang, Mingxiu Chen, et al. Rynnvla-001: Using human demonstrations to improve robot manipulation. *arXiv preprint arXiv:2509.15212*, 2025.

[17] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In *Robotics: Science and Systems (RSS)*, 2023.

[18] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *Robotics: Science and Systems (RSS)*, 2024.

[19] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, pages 2679–2713. PMLR, 2025.

[20] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

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

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

[23] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.

[24] Jiangran Lyu, Ziming Li, Xuesong Shi, Chaoyi Xu,

Yizhou Wang, and He Wang. Dywa: Dynamics-adaptive world action model for generalizable non-prehensile manipulation. *arXiv preprint arXiv:2503.16806*, 2025.

[25] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

[26] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning (CoRL)*. PMLR, 2022.

[27] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems*, 2024.

[28] NVIDIA. Cosmos-Reason2: Physical AI common sense and embodied reasoning models. https://github.com/nvidia-cosmos/cosmos-reason2, 2025. GitHub repository. Accessed: 2026-05-09.

[29] NVIDIA, Johan Bjorck, Nikita Cherniadev Fernando Castaeda, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. In *ArXiv Preprint*, March 2025.

[30] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[31] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alexander Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and RT-X models. *arXiv preprint arXiv:2310.08864*, 2023.

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

[33] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

[34] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

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

[36] Jinhui Ye, Ning Gao, Senqiao Yang, Jinliang Zheng, Zixuan Wang, Yuxin Chen, Pengguang Chen, Yilun Chen, Shu Liu, and Jiaya Jia. StarVLA-α: Reducing complexity in vision-language-action systems. *arXiv preprint arXiv:2604.11757*, 2026.

[37] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. *CoRR*, abs/2507.04447, 2025. doi: 10.48550/ARXIV.2507.04447. URL https://doi.org/10.48550/arXiv.2507.04447.

[38] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

[39] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025.

[40] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

[41] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.