RoboInter: A Holistic Intermediate Representation Suite Towards Robotic Manipulation

Hao Li[1,2,*], Ziqin Wang[3,2,*], Zi-han Ding[4], Shuai Yang[5], Yilun Chen[2,‡], Yang Tian[2], Xiaolin Hu[6], Tai Wang[2], Dahua Lin[7], Feng Zhao[1,†], Si Liu[3,†] and Jiangmiao Pang[2,†]

[1]University of Science and Technology of China, [2]Shanghai Artificial Intelligence Laboratory, [3]Beihang University, [4]Nanyang Technological University, [5]Zhejiang University, [6]Tsinghua University, [7]The Chinese University of Hong Kong

Advances in large vision-language models (VLMs) have stimulated growing interest in vision-language-action (VLA) systems for robot manipulation. However, existing manipulation datasets remain costly to curate, highly embodiment-specific, and insufficient in coverage and diversity, thereby hindering the generalization of VLA models. Recent approaches attempt to mitigate these limitations via a plan-then-execute paradigm, where high-level plans (e.g., subtasks, trace) are first generated and subsequently translated into low-level actions, but they critically rely on extra intermediate supervision, which is largely absent from existing datasets. To bridge this gap, we introduce the ***RoboInter Manipulation Suite****, a unified resource including data, benchmarks, and models of* ***intermediate representations*** *for manipulation. It comprises* ***RoboInter-Tool****, a lightweight GUI that enables semi-automatic annotation of diverse representations, and* ***RoboInter-Data****, a large-scale dataset containing over 230k episodes across 571 diverse scenes, which provides dense per-frame annotations over more than 10 categories of intermediate representations, substantially exceeding prior work in scale and annotation quality. Building upon this foundation,* ***RoboInter-VQA*** *introduces 9 spatial and 20 temporal embodied VQA categories to systematically benchmark and enhance the embodied reasoning capabilities of VLMs. Meanwhile,* ***RoboInter-VLA*** *offers an integrated plan-then-execute framework, supporting modular and end-to-end VLA variants that bridge high-level planning with low-level execution via intermediate supervision. In total, RoboInter establishes a practical foundation for advancing robust and generalizable robotic learning via fine-grained and diverse intermediate representations.*

Code:RoboInter | Data:RoboInter-Data | Homepage

1. Introduction

The remarkable generalization of large language models (LLMs) and vision-language models (VLMs) through large-scale pretraining has inspired efforts to extend this paradigm to robotics, giving rise to end-to-end vision-language-action (VLA) models (Bjorck et al., 2025; Black et al., 2024; Brohan et al., 2023; Kim et al., 2024; Lu et al., 2025). Although web-scale multimodal data enables broad semantic reasoning, existing large-scale robot datasets (Bu et al., 2025; et al., 2023; Khazatsky et al., 2024; Wu et al., 2024) remain costly and tightly coupled to specific embodiments despite massive efforts in data collection (Fu et al., 2024b), leaving a significant gap in generalization and robustness.

To address this gap, recent research has explored frameworks that separate planning from execution. **Modular approaches** (Belkhale et al., 2024; Huang et al., 2024a, 2023; Liu et al., 2024a; Nasiriany et al., 2024) infer high-level structures before translating them into low-level actions, enabling better generalization but often relying on rule-based design. Meanwhile, many **end-to-end VLAs** (Cen et al., 2025; Deng et al., 2025; Du et al., 2023; Lin et al., 2025; Niu et al., 2024; Shi et al., 2025; Wu et al., 2025; Yang et al., 2025b; Zawalski et al., 2024; Zhou et al., 2025b) introduce intermediate representations (e.g., subtasks, visual traces, 2D/3D grounding, etc.) as extra input conditions or supervision signals. Despite their differences, both directions converge on introducing planning as an intermediate representation, which we summarize as the ***plan-then-execute*** paradigm.

* *Equal contributions, ordered by coin toss.* ‡ *Project leader.* † *Corresponding authors. Email: lihaohn@mail.ustc.edu.cn, wzqin@buaa.edu.cn*

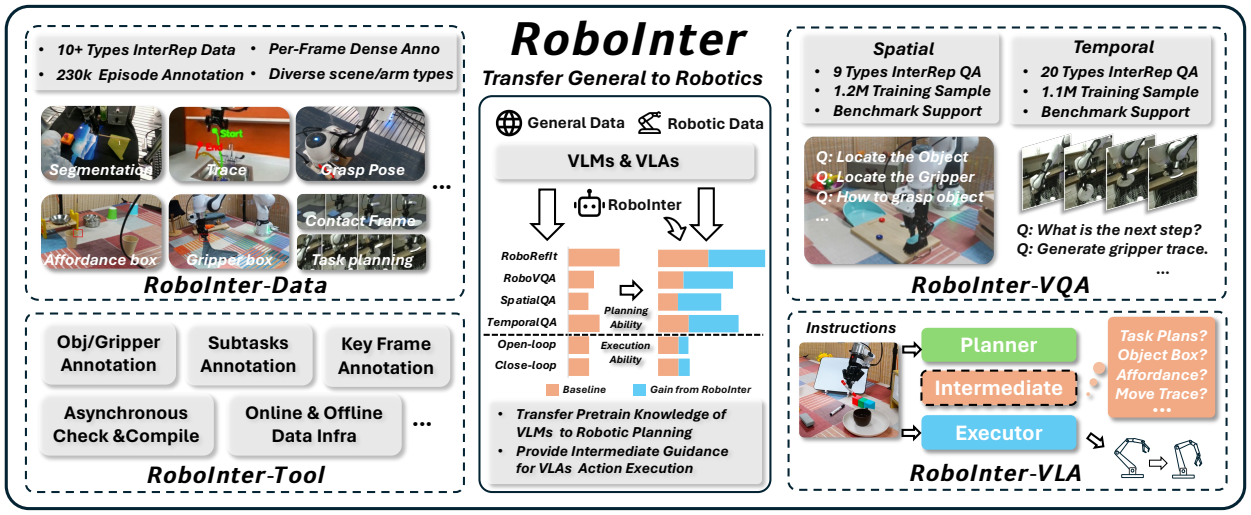

Figure 1. **RoboInter manipulation suite** includes annotation tools, annotated data, curated VQA dataset, and their applications in VLMs and VLAs. RoboInter provides a dataset with over 230k episodes (mainly from Droid (Khazatsky et al., 2024) and RH20T (Fang et al., 2023)) and 10+ types of intermediate representation annotations, named *RoboInter-Data*; a curated embodied VQA benchmark and dataset covering 29 spatial- and temporal-level categories, *RoboInter-VQA*; and an integrated *plan-then-execute* framework for training VLM and VLA models, *RoboInter-VLA*.

The effectiveness of this paradigm critically depends on the availability of high-quality intermediate representations. Existing datasets (et al., 2023; Khazatsky et al., 2024) typically pair visual inputs with overall instructions and robot actions, but they rarely provide the fine-grained intermediates required for ***plan-then-execute***. Collecting and annotating new data remains costly and infrastructure-intensive, which fails to leverage the advantages of community-driven open-source data and makes it unavailable for large-scale pretraining. Recent efforts (Li et al., 2025b; Yuan et al., 2024) have explored automated annotation for existing datasets, yet with limited success. For example, LLARVA (Niu et al., 2024) leverages a pretrained gripper detector to generate large-scale traces, but it is sensitive to distribution shift. ECoT (Zawalski et al., 2024) annotates pseudo-intermediate textual planning with object grounding via Gemini (Team, 2023). ShareRobot (Ji et al., 2025) combines automated annotation with manual verification, but only at a small scale and with labels misaligned to step-wise actions. Overall, current datasets lack large-scale, high-quality annotations, which limits their value for advancing research on intermediate representations for VLMs and ***plan-then-execute*** VLAs.

To address this gap, we propose the ***RoboInter Manipulation Suite***, illustrated in Figure.1. Built on **RoboInter-Tool**, a lightweight GUI for semi-automatic per-frame annotation of embodied videos, we introduce **RoboInter-Data**, a large-scale, per-frame annotation dataset of intermediate representations for robotic manipulation. As shown in Table.1, *RoboInter-Data* provides over 230k episodes across 571 distinct scenes, surpassing LLARVA (Niu et al., 2024), ECoT (Zawalski et al., 2024), and ShareRobot (Ji et al., 2025) in both scale and diversity. Unlike prior datasets that either cover a limited number of scenes (Chen et al., 2024a) or rely solely on automatic annotation (Li et al., 2025b; Zawalski et al., 2024), *RoboInter-Data* uniquely combines large-scale coverage with human-in-the-loop verification. To our knowledge, this is the first real-world manipulation dataset to provide dense per-frame alignment across more than ten categories of intermediate representations, including subtasks, primitive skills, segmentation, gripper/object bounding boxes, placement proposals, affordance boxes, grasp poses, traces, contact points, etc. All annotations are temporally aligned with executed actions, robot states, one third-person and one wrist-view observation, enabling end-to-end action learning.

Table 1. **Comparison of embodied datasets and their annotations.** *Emb.-VQA* denotes the availability of curated embodied VQA benchmarks and datasets; *E2E-ACT* indicates whether the dataset temporally aligned with executed actions; *Curated-CoT* specifies multi-intermediate chain-of-thought support.

Dataset	#Video	#Scene	Dense	**Emb. -VQA**	**E2E -ACT**	**Curated -CoT**	Subtask &Skill	Affor--dance	Contact Point	Gripper Box	Object Box	Trace	Annotation Type
LLARVA	–	311	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	Auto
Hamster	136k	–	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	Auto
RH20T-P	38k	7	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗	Human+Auto
ECoT	60k	12	✓	✗	✓	✓	✓	✗	✗	✓	✓	✗	Auto
AgiBot-World	1M	106	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	Human
VLA-OS	10k	–	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	Auto
ShareRobot	51k	102	✗	✓	✗	✗	✓	✓	✓	✗	✗	✓	Human+Auto
Robo2VLM	176k	463	✗	✓	✗	✗	✓	✗	✓	✗	✗	✓	Auto
VeBrain	12k	–	✓	✓	✗	✗	✓	✗	✓	✗	✗	✗	Human
Ours	**230k**	**571**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**Human+Auto**

Leveraging these fine-grained annotations, we introduce **RoboInter-VQA**, which exposes and benchmarks existing VLMs from the perspectives of embodied generation and understanding, comprising 9 spatial and 20 temporal embodied-scene VQA categories to enhance the embodied abilities of VLMs. Built upon the high-level VLM planner trained on these curated VQA data, we introduce **RoboInter-VLA**, an integrated plan-then-execute framework that supports both modular and end-to-end VLA variants, enabling rapid adaptation from planning to execution, and then we systematically investigate the impact of intermediate representations on the generalization and controllability of VLA variants. Through extensive experiments, we show that RoboInter-Data substantially improves the reasoning and grounding capabilities of VLM planners, particularly in understanding and generating various embodied intermediate representations for manipulation, thereby strengthening VLMs' foundational embodied abilities. Open- and closed-loop evaluations of manipulation tasks further demonstrate that the pretrained planners and our intermediate representation data provide performance and generalization gains to the VLAs. By analyzing the trade-offs among different VLA variants, we establish a unified foundation for leveraging these data in future research. We will open-source RoboInter with its corresponding data, benchmarks, and models to the community, and hope they can pave the way for applications of intermediate representations for embodied AI.

2. Related Works

Embodied intermediate representation. Embodied intermediate representations have been widely explored. Prior work leverages 2D trace (Gu et al., 2023), optical flow (Xu et al., 2024), subtasks (Belkhale et al., 2024; Zhang et al., 2024), key points (Wen et al., 2023), future images (Cai et al., 2026; Lv et al., 2025; Zhao et al., 2025) and grounding box (Huang et al., 2025; Sundaresan et al., 2023) to guide action generation. While some representations (e.g., grounding box and flow) can be estimated by vision foundation models (Dong and Fu, 2024; Ravi et al., 2024), the results are not always reliable across different embodied scenes. Web-scale resources remain under-aligned with manipulation tasks, so embodied alternatives such as RoboBrain (Team et al., 2025) and VeBrain (Luo et al., 2025) curate task-specific embodied datasets and train VLMs to generate intermediate representations. A complementary direction (Li et al., 2025b; Yuan et al., 2024) focuses on collecting a large-scale dataset for a single intermediate representation and utilizing it for action generation.

Robotic manipulation datasets. Numerous prior works (et al., 2023; Fang et al., 2023; Gao et al., 2025b; Khazatsky et al., 2024; Tian et al., 2025) have introduced real-world or simulated robot datasets encompassing diverse sources, scenarios, and skills. While these data do not provide native

intermediate representation labels for real-world manipulation tasks, subsequent efforts have provided some annotations. RH20T-P (Chen et al., 2024a) augments RH20T with primitive-level labels of subtasks, RT-H (Belkhale et al., 2024) provides language descriptions of grasp poses and motions. LLaRVA (Niu et al., 2024) and Hamster (Li et al., 2025b) leverage gripper tracking to generate large-scale 2D trajectories for pretraining. High-level planning approaches, such as Robo2VLM (Chen et al., 2025a), sample frames and then supply diverse VQA annotations to enhance embodied scene understanding. RoboInter introduces ***pre-frame*** dense annotations data across varied intermediate representations to advance both embodied understanding and end-to-end action learning.

Embodied planning for execution. Embodied planning is important in diverse and complex real-world settings. ***Plan-then-execute*** systems can be categorized into implicit and explicit forms. Implicit methods operate as black boxes, where these VLA methods (Black et al., 2024; Li et al., 2025a, 2023a) primarily rely on implicit reasoning by fine-tuning pretrained VLMs of various pretraining paradigms. Explicit methods generate interpretable intermediate representations and directly utilize the knowledge from VLMs, such as $\pi_{0.5}$ (Intelligence et al., 2025), and Rekep (Huang et al., 2024b). ECoT (Zawalski et al., 2024) introduces explicit CoT into VLA by prompting a VLM to autoregressively generate both CoT text and discrete action tokens. VLA-OS (Gao et al., 2025a) further explores various model designs to combine planning and action generation. We offer large-scale intermediate representation data and the pretrained VLM Planners, jointly enabling both implicit and explicit reasoning in VLA variants through efficient information aggregation.

3. Dataset

3.1. RoboInter-Data

As illustrated in Figure 2, RoboInter-Data builds upon extensive manipulation datasets and provides large-scale, high-quality annotations of diverse intermediate representations.

Data collection. To enhance dataset diversity, we collected two types of raw manipulation data: (1) ***In-the-Wild setting*** (i.e., diverse indoor scenarios), emphasizing the diversity of scenes and instructions, mainly collected from Droid (Khazatsky et al., 2024) and OXE (et al., 2023). (2) ***Table-Top setting*** (i.e., tabletop interaction scenarios), highlighting the high quality and skills diversity, collected from RH20T (Fang et al., 2023). By integrating raw teleoperated video recordings of these datasets, followed by rigorous screening and pre-processing, we constructed a high-quality, large-scale database consisting of 230k manipulation episodes (third-person videos) in total.

Annotations & Check with RoboInter-Tool. To obtain accurate and comprehensive labels, *RoboInter-Tool* is employed to perform the following annotations: ***(1) Task decomposition & key-frame annotation.*** Each manipulation video is decomposed into clips using 15 predefined primitive skills, and ChatGPT (OpenAI, 2023) is employed to produce preliminary references for language annotations. Human annotators utilize *RoboInter-Tool* to segment video clips, assigning each clip to a primitive skill, and simultaneously completing clip-level and video-level language annotations. The *contact frame* where the robot arm contacts the manipulated object is also recorded. ***(2) Recognizing the manipulated object.*** After recording the object that the robot arm interacts with, *RoboInter-Tool* automatically transports the annotation to *SAM2* (Ravi et al., 2024) for object segmentation and tracking, and the result is asynchronously returned for review. A re-annotation and inspection mechanism reduces the impact of segmentation or tracking errors on the quality. ***(3) Locating the end-effector.*** As many raw recordings lack reliable camera parameters, directly projecting 3D coordinates to obtain accurate 2D end-effector traces is often infeasible. We estimate a calibration matrix to improve projection accuracy and complement parameter-missing episodes via gripper detection and point tracking, enabling reliable reconstruction of the end-effector's 2D trace. Details are provided in Appendix A.6.

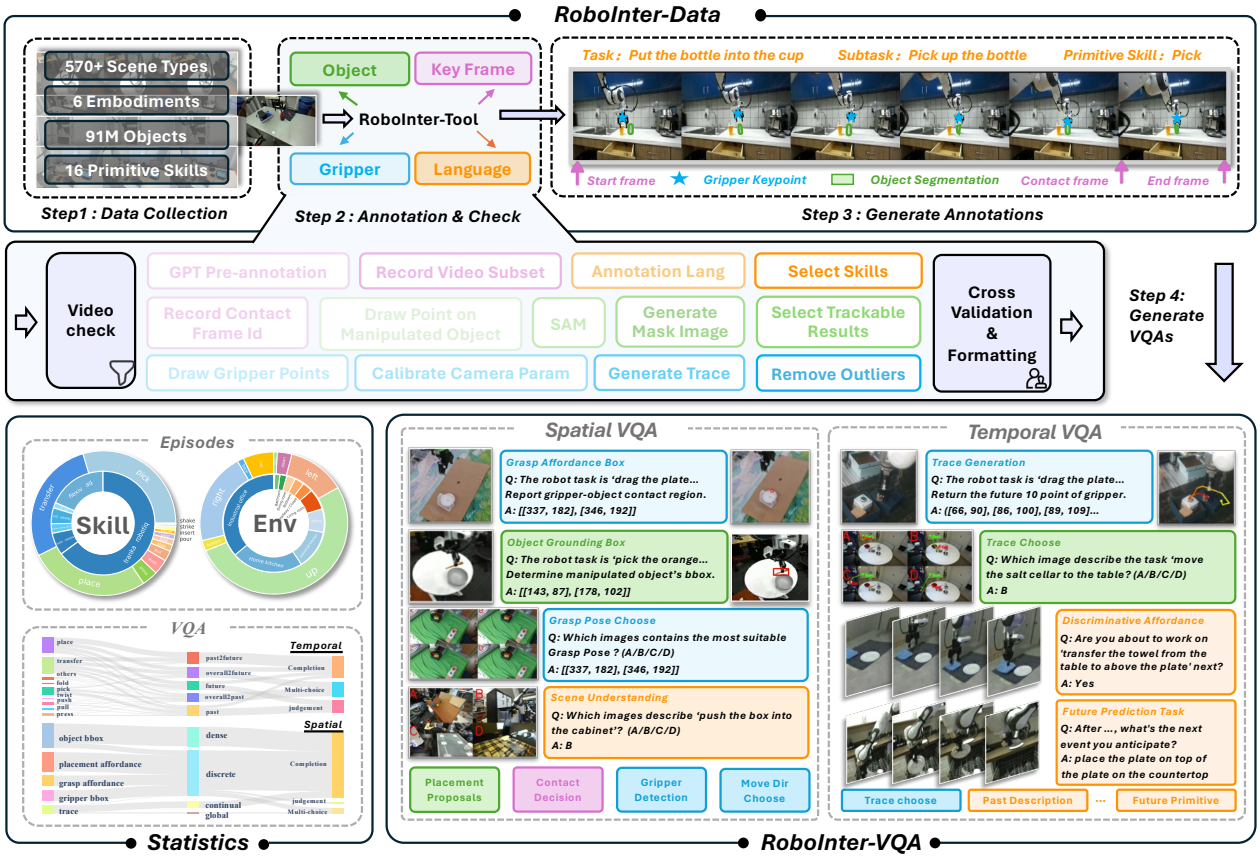

Figure 2. **Overview of RoboInter-Data and RoboInter-VQA.** We collect and annotate 230k manipulation episodes to obtain 10 types of intermediate representation annotations through Data Collection and Annotation & Check. We further construct a large-scale, diverse set of VQA spanning spatial and temporal dimensions. Statistics of raw manipulation episodes and the curated VQA are also provided.

Post-processed annotations. By reorganizing the above annotations, we derive additional intermediate representations: ***(1) Grasp annotation.*** The grasp affordance box is inferred from the 2D end-effector location at the annotated contact frame. The contact points are the pre-defined key points of the gripper at the moment of contacting, while the corresponding robot state (i.e., the 6D end-effector pose) defines the grasp pose. ***(2) Placement annotation.*** The object position at the end of the subtask is treated as the target placement location. ***(3) Gripper annotation.*** Anchor points enclosing the gripper are identified and projected from 3D to 2D using camera parameters and robot states, forming the gripper bounding box on the 2D image coordinates.

3.2. RoboInter-VQA

VQA task construction. As illustrated in the RoboInter-VQA section of Figure.2, we convert the annotations into diverse VQA tasks to enhance VLM capabilities. Tasks are organized along two axes: (i) intermediate representation type (spatial vs. temporal) and (ii) target capability (understanding vs. generation). ***(1) Spatial VQA for understanding.*** We design three selection tasks and one judgment task to train spatial comprehension, including selecting the correct object bounding box or grasp pose, matching scenes to instructions, and determining whether contact occurs. ***(2) Spatial VQA for generation.*** It includes five prediction tasks that require generating spatial intermediate representations for downstream execution and complete specific content based on spatial reasoning, which includes object bounding box, grasp pose, placement proposal, key points, and the gripper

bounding box. ***(3) Temporal VQA for understanding.*** To evaluate how the VLMs understand motion traces and the relationships between subtasks and observations, we design selection tasks and judgment tasks for temporal information. We design five selection tasks for movement directions of grippers, the matching of trace and description, subtask/primitive discrimination, and execution stage identification. Four judgment tasks assessing task success and next-step feasibility. ***(4) Temporal VQA for generation.*** We formulate tasks that require trace generation and multi-step planning under varying levels of contextual completeness. Prompts condition on different amounts of prior information (e.g., past subtasks or overall instructions) and ask the model to predict the subsequent steps or multi-step planning. Video-based visual inputs are also used to summarize past events and predict feasible next steps. Trace generation is evaluated under both easy and challenging settings (with or without initial waypoints). Details are included in the Appendix.A.7.3.

3.3. Data Statistics

Statistics of RoboInter-Data. As shown in Figure.2, we provide high-quality, scene-diverse intermediate representation annotations for 230k manipulation episodes. This dataset includes 6 types of robot arms, 571 types of scenes, and 15 types of primitive skills. With the assistance of the RoboInter-Tool, we produce nearly 61M-frame object grounding annotations, about 70M-frame gripper trace annotations, 190k affordance box and placement proposal annotations, and nearly 760k language clip annotations. Despite this scale, the annotations maintain high quality.

Statistics of RoboInter-VQA. Our VQA data is also large in scale, comprising approximately 1M spatial generation entries, 172k spatial understanding entries, 131k temporal generation entries, and 935k temporal understanding entries. To prevent information leakage between training and validation, we carefully designate 7,246 videos as the evaluation pool with the remaining data used for training, and sample validation sets for each question category from this pool.

4. RoboInter-VLA

In this section, as illustrated in Figure.3, we present **RoboInter-VLA**, a family of models following a **plan-then-execute** paradigm, consisting of a *Planner* and an *Executor*. Rather than a monolithic design, RoboInter-VLA supports multiple variants and enables flexible adaptation from planning to execution. The *Planner* performs high-level decision-making by combining general and embodied reasoning to produce intermediate representations, which guide the *Executor* in translating multimodal observations and language instructions into low-level actions.

4.1. Model Architecture

VLMs as Planner. The Planner model acquires embodied capabilities through a visual question answering formulation with a co-training strategy. To capture both spatial and temporal information, we adopt VLM architectures that support single- and multi-image inputs, including the Qwen-VL series (Wang et al., 2024b) and LLaVA-One-Vision (Li et al., 2024a). Each model consists of a base LLM, a vision encoder, and an MLP-based vision–language projector. The Planner generates outputs autoregressively and is optimized using a cross-entropy loss.

VLAs as Executor. To systematically and lightweightly derive from the Planner, we build Executor on a Qwen2.5-VL backbone with a Diffusion Transformer (DiT) action head (Peebles and Xie, 2023). We further utilize an *information aggregator* that gathers the hidden states of all input and output tokens, as well as intermediate representations, and compresses them into conditioning features with a controllable length. The *Executor* consumes multi-view visual observations (e.g., primary and

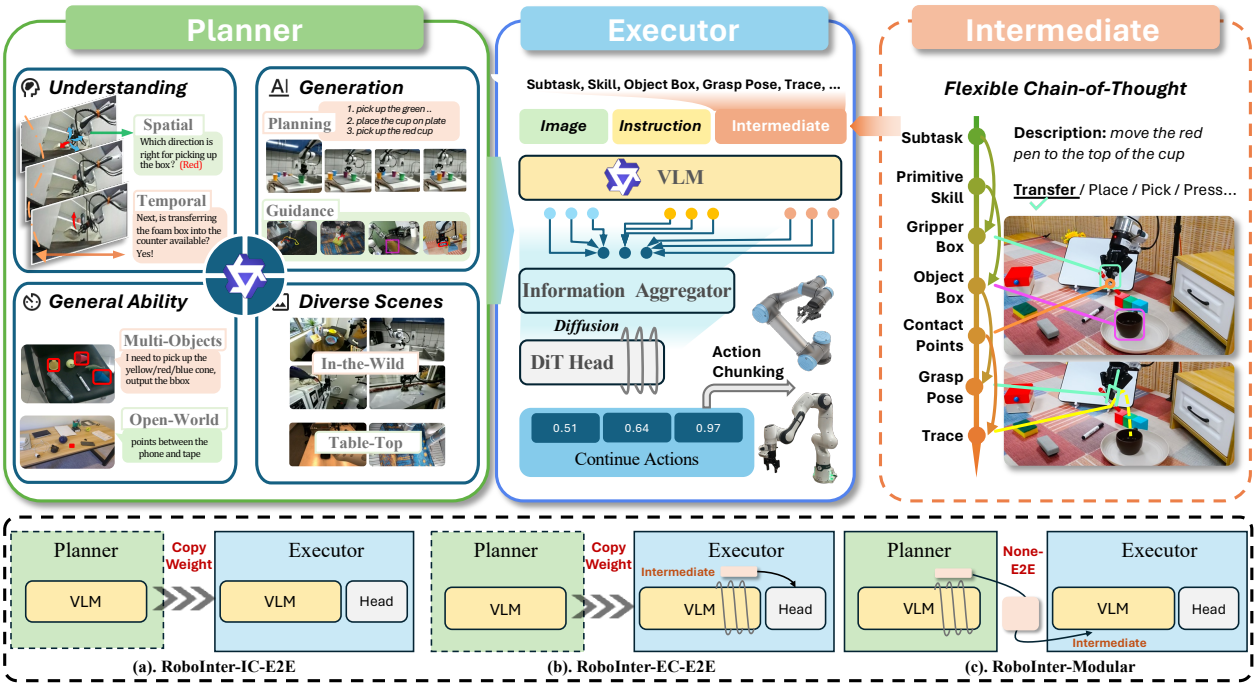

Figure 3. **Framework of RoboInter-VLA.** Our model follows a ***plan-then-execute*** paradigm with a VLM-based *Planner* and an *Executor*. The Planner exhibits enhanced understanding and generation for manipulation, strong general grounding abilities, and robust perception across diverse scenes. The Executor shares the VLM backbone with the Planner. Three variants are supported, and intermediate representations in Flexible Chain-of-Thought (F-CoT) bridge planning and execution.

wrist), language instructions, and intermediate representations (based on primary observation), and produces multi-step action chunks via a diffusion loss.

4.2. Plan-Then-Execute Paradigms

As shown in Figure.3, by leveraging the pretrained Planner, we provide three paradigms to enhance downstream action execution: (1) *RoboInter-IC-E2E (Implicitly-Conditioned End-to-End)*, which directly injects the VLM from a pretrained Planner into the end-to-end Executor, using it as a stronger vision-language feature extractor. This approach can yield robust embodied perception and more accurate task-relevant visual cues. (2) *RoboInter-EC-E2E (Explicitly-Conditioned End-to-End)*, where the Executor is initialized with the VLM of the Planner and jointly optimizes both intermediate-representation reasoning and action generation. (3) *RoboInter-Modular (Modular Planner-to-Executor)*, a non-E2E hierarchical design that treats the Planner and Executor as independent modules. During training, the Executor is conditioned on ground-truth intermediate representations to generate actions. During inference, it relies on the intermediate representations predicted by the Planner.

Flexible chain-of-thought for intermediate representations. To support the explicitly conditioned and modular architectures, we introduce **F-CoT**, a chain-of-thought composed of multiple intermediate representations. F-CoT plays two roles: (i) as VQA supervision for training the Planner, and (ii) as action-aligned guidance for the Executor. In *RoboInter-IC-E2E*, the VLM generates F-CoT content, which is directly consumed by the DiT head. In *RoboInter-Modular*, the Planner produces the F-CoT content and the Executor conditions on it. F-CoT flexibly combines representations such as subtasks, skills, object bounding boxes, affordance boxes, motion traces, etc., in textual or visual form, allowing users to select subsets tailored to specific embodied tasks. We denote textual F-CoT as

Table 2. **Performance comparison on third-party benchmarks.** Including **Embodied**, **Grounding**, and **General** benchmarks for general VLMs (upper) and embodied VLMs (lower).

Model Name	Embodied			Grounding			General					
	Where-2place↑	RoboRefIt-test↑	Robo-VQA↑	Refcoco-g-val↑	Refcoco+ val↑	Refcoco-val↑	Text-VQA↑	CO-CO↑	OCR-bench↑	MME↑	MM-VET↑	POPE↑
InternVL3-1B	2.65%	7.0%	30.5	79.8%	73.2%	83.0%	75.1	23.7	798	1907	58.9%	90.7%
InternVL3-2B	1.86%	27.5%	27.7	87.6%	84.0%	85.8%	77.0	**27.9**	835	2186	62.2%	89.6%
InternVL3-8B	1.95%	27.7%	27.9	**89.6%**	**88.2%**	**92.5%**	80.2	26.5	880	**2410**	**81.3%**	91.1%
QwenVL2.5-3B	11.8%	68.9%	37.6	85.2%	82.4%	89.1%	79.3	15.7	797	2175	61.8%	85.9%
QwenVL2.5-7B	18.9%	75.8%	38.4	87.2%	84.2%	90.2%	**84.9**	15.0	864	2306	67.1%	85.9%
LLaVA-OV-7B	7.9%	10.4%	31.4	71.9%	69.7%	73.8%	71.1	8.4	**882**	2307	67.3%	86.4%
Robobrain-2.0-3B	59.8%	30.9%	30.6	55.0%	51.5%	50.9%	81.0	27.2	811	2126	59.4%	88.1%
Robobrain-2.0-7B	63.6%	8.8%	31.6	62.9%	70.1%	76.1%	75.9	25.2	857	2076	61.4%	86.2%
RoboInter-Qwen-3B	58.3%	80.0%	43.3	87.9%	85.8%	89.5%	78.9	15.6	787	2180	61.0%	90.5%
RoboInter-Qwen-7B	65.8%	85.6%	74.4	88.4%	86.6%	91.5%	83.0	15.9	832	2281	62.3%	**91.4%**
RoboInter-LLaVAOV-7B	**66.3%**	**89.3%**	**74.5**	87.3%	84.2%	91.3%	72.2	15.8	725	2217	61.4%	90.4%

Table 3. **Performance on RoboInter-VQA *spatial* and *temporal* benchmark. G.D.** means grounding, **A.F.** denotes the Affordance. Spatial generation uses ACC@IOU>0.1 (%↑), multiple choice and T/F use ACC (%↑). For temporal, Trace employs DTW (↓) and other metrics use ACC or average BLEU (↑).

Model Name	RoboInter-VQA Spatial							RoboInter-VQA Temporal				
	Generation				Multiple Choice		T/F	Generation		Multiple Choice		T/F
	Object G.D.↑	Grasp A.F.↑	Place A.F.↑	Gripper G.D.↑	Grasp Pose↑	Grouding Choice↑	Contact↑	Trace↓	Task Planning↑	Visual Trace↑	Planning Choice↑	Task Planning↑
QwenVL2.5-3B	46.6%	12.2%	34.1%	6.1%	21.1%	21.9%	50.9%	2712	20.3	37.5%	60.0%	59.7%
QwenVL2.5-7B	51.2%	14.7%	38.2%	10.2%	27.3%	25.7%	52.5%	1702	22.4	39.0%	64.5%	60.5%
InternVL3-1B	7.8%	2.3%	8.3%	1.2%	24.8%	25.9%	50.4%	–	10.5	28.9%	54.9%	55.8%
InternVL3-2B	20.6%	3.1%	17.9%	1.9%	25.5%	27.3%	50.1%	–	7.7	35.2%	59.3%	59.4%
InternVL3-8B	32.7%	5.9%	28.2%	3.5%	25.1%	31.1%	52.9%	1035	8.1	34.0%	71.5%	60.0%
Llava-OV-7B	25.8%	5.5%	23.7%	1.6%	24.5%	31.9%	54.4%	–	11.0	37.7%	44.9%	63.5%
GPT4o-mini	6.8%	–	7.2%	1.1%	10.9%	16.8%	53.6%	1736	14.7	28.4%	66.6%	63.9%
Gemini-2.5-flash	1.7%	–	1.2%	–	32.7%	69.4%	65.5%	–	–	49.4%	–	–
Robobrain-2.0-3B	15.2%	–	–	2.8%	25.5%	26.5%	50.4%	595	16.0	29.7%	48.2%	46.8%
Robobrain-2.0-7B	–	–	–	2.5%	23.3%	21.5%	49.2%	541	15.3	29.5%	57.8%	46.4%
RoboInter-Qwen-3B	76.1%	34.9%	52.7%	61.6%	74.0%	73.4%	75.2%	332	61.2	78.8%	82.2%	88.7%
RoboInter-Qwen-7B	75.1%	37.8%	**56.9%**	62.0%	**76.1%**	75.7%	75.6%	323	**63.4**	**81.9%**	**86.5%**	**93.0%**
RoboInter-LlavaOV-7B	**82.9%**	**46.3%**	55.1%	**70.1%**	74.1%	**79.7%**	**76.3%**	**299**	62.7	**81.9%**	81.8%	83.9%

RoboInter-Te-Modular and visual-prompted F-CoT as *RoboInter-Im-Modular*.

5. Benchmarking and Experiments

5.1. Benchmarking the Planner

Enhanced grounding and embodied capability. As shown in Table 2, we evaluated on third-party spatial reasoning benchmarks, including Where2Place (Yuan et al., 2024) and RoboRefIt (Lu et al., 2023) (spatial point and grounding reasoning) and RoboVQA (Sermanet et al., 2024) (temporal task planning). Across all three benchmarks, our models substantially outperformed the base models (Qwen2.5-VL-3B/7B (Wang et al., 2024b) and LLaVA-OneVision-7B (Li et al., 2024a)). Notably, RoboBrain2.0 (Team et al., 2025) is also an embodied VLM (i.e., Planner). At the 3B scale, *RoboInter-Qwen-3B* achieved a 49.1% improvement over RoboBrain2.0 on RoboRefIt and a 12.7% improvement on RoboVQA. At the 7B scale, the corresponding gains reached 76.8% and 42.8%, respectively. For grounding, all three RoboInterVLM variants exceeded their respective base models on Refcoco (Lin

Table 4. **Open-loop evaluation in In-the-Wild setting**. We report OLS with different error thresholds (@0.1 to @0.01) and the mean value.

Method	Open-Loop Score (OLS)				mOLS
	@0.1	@0.05	@0.03	@0.01	-
VLA-OS	0.6180	0.3905	0.1928	0.0129	0.3035
Vanilla	0.6793	0.3608	0.1753	0.0189	0.3086
RoboInter-IC-E2E	0.6984	0.3810	0.1873	0.0204	0.3218
RoboInter-EC-E2E	0.7049	0.3930	0.2066	0.0314	0.3340
QwenVL+Executor	0.6749	0.3582	0.1777	0.0298	0.3102
RoboInter-Te-Modular	0.7124	0.4133	0.2332	0.0584	0.3543
RoboInter-Im-Modular	0.7056	0.4029	0.2240	0.0430	0.3439
Oracle+VLA-OS	0.7260	0.4928	0.2734	0.0200	0.3780
Oracle+Executor	0.7511	0.4640	0.2705	0.0587	0.3861

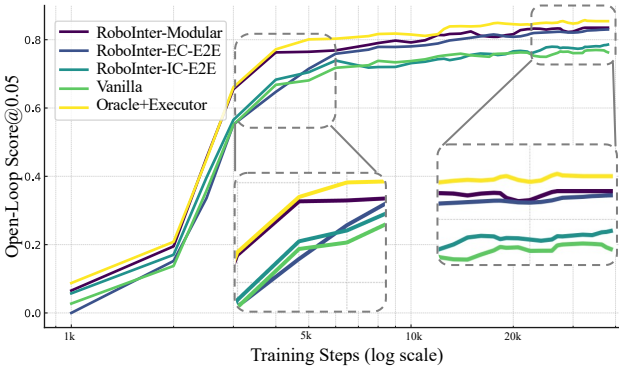

Figure 4. **Open-loop evaluation in TableTop setting**. We show the curve of OLS@0.05 from 1k to 40k training steps. We mainly report the five variances of RoboInter-VLA.

et al., 2014). Particularly, *RoboInter-Qwen-7B* eventually ranked second overall, with a 27.4% relative improvement over Robobrain2.0-7B. On general benchmarks, our models remained relatively stable on most benchmarks, indicating that ***our curated VQA data enhances the abilities of embodied reasoning and grounding***, meanwhile general capabilities of our VLMs are slightly affected.

RoboInter-VQA benchmark at the spatial and temporal level. As shown in Table 3, for spatial-based generation tasks, closed-source and general VLMs without embodied experience rarely produce accurate intermediates (typically below 40%), underscoring the importance of additional intermediate representation annotations. On simpler questions, Gemini-2.5-Flash and RoboBrain-2.0-7B lead on *Grasp Pose (choice)* with 32.7% and 23.3% ACC. For *Grounding Choice*, LLaVA-OV-7B achieves 31.9%, while Gemini-2.5-Flash is strongest at 69.4%; most other models remain near random choosing (25%) given limited understanding of manipulation scenes. For temporal, closed-source API and general VLMs largely fail to generate future traces or task planning; RoboBrain-2.0 attains a much better DTW in *Trace Generation* than Qwen-VL-2.5 (541 v.s. 1702). On *Visual Trace Choice*, Gemini-2.5-Flash remains competitive (49.4%). For *Planning Choice* and *T/F of Task Planning*, as planning aligns closely with general LLM abilities, most models transfer common-sense knowledge and show better performance. Overall, ***current closed-source and general VLMs typically lack enough embodied abilities***. Curated from diverse annotations, ***our RoboInter-VQA markedly improves the VLM abilities of understanding and generating intermediate representations***.

5.2. Open-Loop Evaluation of the Executor

Experimental settings. In this section, we examine how a pretrained Planner improves the Executor and compare different VLA paradigms. As our annotated corpus spans more than 500 distinct scenarios, comprehensive real-world validation across all scenarios is infeasible, as emphasized by HPT (Wang et al., 2024a). Following the discrete token-accuracy evaluation in OpenVLA (Kim et al., 2024), we utilize an ***Open-Loop Score (OLS)*** to evaluate the generation of *continuous action chunks*, in which per-step actions are assessed independently and compared with the ground-truth actions. OLS is computed as the average value over 100K transitions from evaluation videos, ensuring statistical stability. More details in Appendix.A.4.3. Nine Executor variants are evaluated: (a).*Vanilla*: omits any pretrained VLM from Planner, performing action learning only; (b-e).*RoboInter-IC-E2E*, *EC-E2E*, *Te-Modular* and *Im-Modular* are stated in Section.4.2; (f).*Oracle+Executor*: not end-to-end, both training and inference are guided by GT intermediate representations; (g).*QwenVL+Executor*: training is GT-guided, and inference employs intermediates from original Qwen2.5VL; (h).*VLA-OS*: we train

VLA-OS (Gao et al., 2025a) in our setting and use it as an additional baseline; (i).*Oracle+VLA-OS*: not end-to-end, VLA-OS are guided by GT intermediate representations. Two open-loop evaluation settings: (1) *In-the-Wild*: focus on scene and object generalization, we compare the convergence performance under identical training steps. (2) *Table-Top*: focus on the tabletop environment and cross-embodiment ability, we mainly examine the evaluation curve during training. The annotated corpus is divided into *In-the-Wild* and *Table-Top* subsets. We sample approximately 10% of episodes from each subset for Executor training (25k in total), of which 8% are reserved for evaluation.

Planner consistently improves the Executor's action generation. The *In-the-Wild* setting is shown in Table.4. The *Vanilla* achieves a lower mOLS score than IC-E2E (0.3086 v.s. 0.3218), which incorporates intermediate representations, indicating that ***pretrained VLM Planner can enhance the learning capability of the VLA Executor***. The mOLS score of *EC-E2E* is higher than *IC-E2E* (0.3340 v.s. 0.3218), showing that the ***explicit intermediate representations are more helpful for action guidance than the implicit***. For *Oracle + Executor*, the non-E2E architectures, utilizing ground-truth annotation in the Executor, achieve substantially the highest scores. This indicates that ***our annotations are informative and stable***. *Te-Modular* surpasses *EC-E2E* (0.3543 v.s. 0.3340), implying that ***decoupling planning and execution facilitates dedicated optimization*** of each capability and mitigates mode conflict. *Im-Modular* performs slightly worse than *Te-Modular* (0.3439 v.s. 0.3543), as visual prompting embeds within information-dense images, diluting their relative contribution. *QwenVL + Executor* employs Qwen2.5-VL as a zero-shot Planner, and its overall performance is not comparable with other Non-E2E models, ***showing that the embodied reasoning ability of our Planner is better than the general VLMs***. The evaluation curves in the *Table-Top* setting are shown in Figure.4. As Table-Top scenes are easier to interpret, most methods eventually achieve high scores. *RoboInter-Te-Modular* and *Oracle + Executor* converge faster and reach higher performance. *EC-E2E* model converges more slowly but ultimately approaches the performance of *Te-Modular*. *IC-E2E* shows stronger early-stage results and maintains a consistent advantage over *Vanilla* after 20k steps.

Ablations on intermediate representations. We ablate combinations of different intermediate representations within the open-loop Oracle+Executor setting. As shown in Table 5, coarse-grained representations such as *Subtask* and *Primitive Skill* provide only marginal improvements, as they offer stage-level guidance with limited actionable constraints during execution. In contrast, spatially grounded representations (*Object Box*, *Gripper Box*, and *Affordance*) yield substantially larger gains by providing finer-grained cues. The most significant improvement comes from *Trace*, which introduces dense, temporally grounded information and achieves the strongest overall performance. Additional results are provided in Appendix A.2.3.

Table 5. **Ablation of intermediate representation**. We report OLS under multiple thresholds. Six types of representations are evaluated, where finer-grained categories yield larger gains.

Variant	OLS				mOLS
	@0.1	@0.05	@0.03	@0.01	-
Vanilla	0.6793	0.3608	0.1753	0.0189	0.3086
+ Subtask	0.6965	0.3676	0.1770	0.0171	0.3146
+ S. + Primitive Skill	0.6983	0.3681	0.1779	0.0194	0.3159
+ S. + P. + Object Box	0.7025	0.3849	0.1988	0.0294	0.3289
+ S. + P. + O.B. + Gripper Box	0.7212	0.4032	0.2048	0.0272	0.3391
+ S. + P. + O.B. + G.B. + Affordance	0.7245	0.4083	0.2114	0.0297	0.3435
+ S. + P. + O.B. + G.B. + Aff. + Trace	**0.7511**	**0.4640**	**0.2705**	**0.0587**	**0.3861**

5.3. Closed-Loop Real-World Evaluation of the Executor

Experimental setting. We study how our dataset and the pretrained Planner affect the closed-loop success rate. Experiments are conducted in a few-shot TableTop evaluation with a real-world Franka

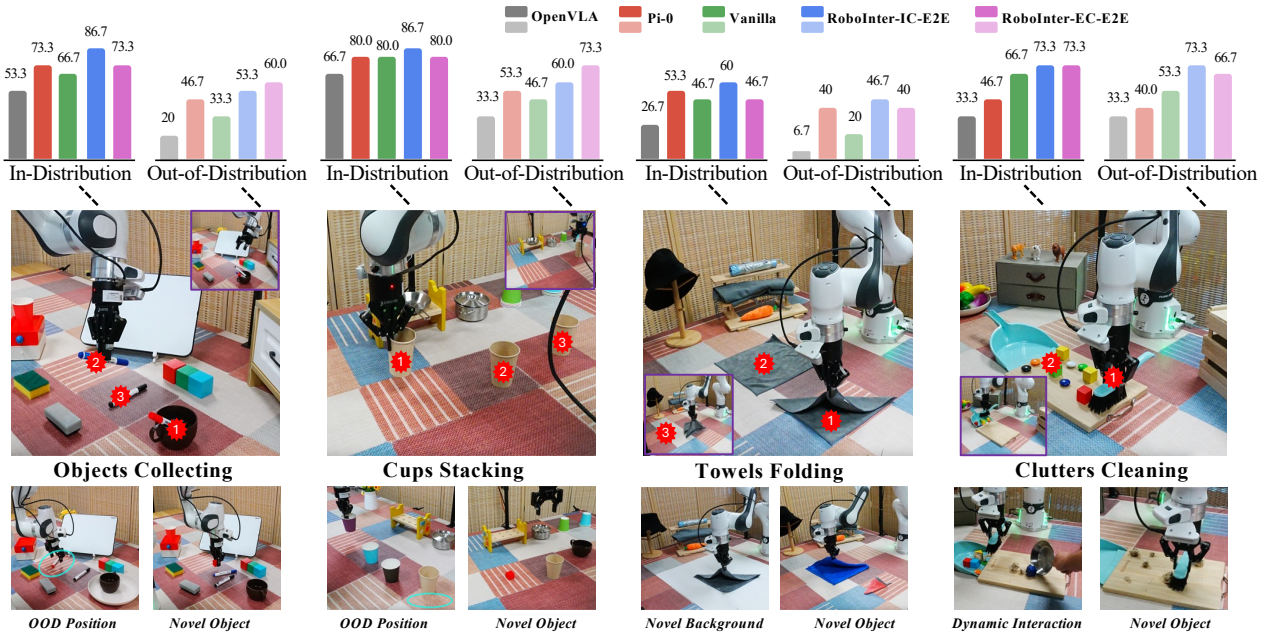

Figure 5. **Real-World Experiments.** The top charts present results from 15 in-distribution (ID) and 15 out-of-distribution (OOD) trials. The bottom panel illustrates the OOD test setup. Notably, the performance drop from ID to OOD reflects each model's generalization under distribution shift, where EC-E2E outperforms IC-E2E and exhibits a smaller ID→OOD degradation (8.3% vs. 19.0%), showing the consistent conclusion with the *Open-Loop Evaluation*. Key steps are marked with number, along with an end-execution thumbnail. Experimental results of RoboInter-Modular is in Table.6.

Research 3 arm. Observation input comprises a static third-person camera and a wrist-view camera, and no proprioceptive state. We focus on more practical E2E variants and evaluate five E2E models: (1) *OpenVLA* (Kim et al., 2024): Initialized from official pretrained weights and extended with an additional wrist-view input. (2) *Pi-0* (Black et al., 2024): Fine-tuned from the official checkpoints of Droid. (3) *Vanilla*: our baseline. (4) *RoboInter-IC-E2E*: Initialized from the Planner and further finetuned on in-distribution (ID) data. (5) *RoboInter-EC-E2E*: Initialized from the Planner, and jointly optimize action and CoT generation, with a 1:1 ratio of our annotated data and collected data (more details in A.1.1). We design four tasks: (a) Object Collecting: sequentially place three pens from a cluttered tabletop into a cup, out-of-distribution (OOD) tests are with novel objects, and alter spatial layouts. (b) Cup Stacking: Stack cups from left to right. OOD tests include novel objects, OOD positions, and continuous stacking. (c) Towel Folding: Fold two towels in sequence and stack them. OOD tests vary in the towel category and background. (d) Clutter Cleaning: Clean all items from the board with a brush. OOD tests introduce novel objects and disturbances.

Experimental results on ID and OOD testing. Across all tasks, *RoboInter-IC-E2E* consistently outperforms the *Vanilla*. In ID evaluations, IC-E2E attains an average success rate of 77.3%, compared with 65.0% of Vanilla. Under OOD conditions, the gap widens, and IC-E2E achieves a 58.3% success rate, while Vanilla reaches only 38.3%, indicating the superior generalization of *IC-E2E*. ***The pretrained VLM from the Planner is pre-exposed to rich embodied data and therefore provides stronger perceptual priors.*** Although Pi-0, which is pretrained on Droid, also demonstrates solid ID and OOD performance, the *IC-E2E* benefits from a broader representation training, thereby producing better overall results. *EC-E2E* records a lower ID success rate than *IC-E2E* (68.3% vs. 77.3%), which seems to be misaligned with the open-loop results in which *EC-E2E* was superior. Actually, the open-loop protocol enforces strict decouple between training and validation, and therefore functions more like

an OOD test. Correspondingly, under real-world OOD conditions, *EC-E2E* exceeds *IC-E2E* in Object Collecting (60.0% vs. 53.3%) and Cup Stacking (73.3% vs. 60.0%), and achieves a higher average success rate (60.0% vs. 58.3%). The ID-to-OOD drop is only 8.3% for *EC-E2E*, whereas *IC-E2E* declines by 19%. We attribute *EC-E2E*'s weaker ID accuracy to the potential modality interference from the joint training of text generation and action prediction. ***Diverse OOD knowledge from our dataset contributes to superior OOD robustness and generalization***. More results are in Appendix A.1.1.

6. Conclusion and Further Applications

Conclusion. We presented the RoboInter Manipulation Suite, a unified platform designed to advance research on intermediate representations for the plan-then-execute VLA paradigm. At its core, RoboInter-Data offers over 230k episodes with dense, per-frame annotations spanning diverse intermediate representations, establishing a new scale and quality standard for real-world manipulation datasets. Built upon this foundation, RoboInter-VQA systematically benchmarks and improves the embodied generation and understanding capabilities of RoboInter-VLMs across rich spatial and temporal reasoning tasks. RoboInter-VLA further integrates intermediate representations into both modular and end-to-end frameworks, enabling a principled study of how intermediate signals influence execution performance. We open-source RoboInter to facilitate future research on intermediate representations as a key bridge between vision-language reasoning and robotic action.

Further applications. Beyond training VLMs and VLAs, *RoboInter-Data* also supports broader research directions: **(1) Expert generative models for individual intermediate representations.** In contrast to datasets centered on a single annotation type, RoboInter-Data spans diverse embodied scenes while providing large-scale, high-quality annotation for each intermediate representation, enabling the pre-training of specialized generative models tailored to specific representations. **(2) Human–robot interaction.** Intermediate representations such as traces, bounding boxes, and subtask sequences naturally bridge high-level instructions and low-level control. Our dataset supports the development of intuitive and precise shared-autonomy systems grounded in these representations. **(3) Embodied world model learning.** All annotations in RoboInter-Data are temporally aligned with 640×360 raw manipulation videos. These intermediate representations can serve as structured control signals for controllable video generation, while the aligned information about the robot, objects, and environment provides strong supervision for learning structured embodied world models. **(4) Video action model learning.** The combination of action annotations, videos, and diverse intermediate representations makes RoboInter-Data well suited for training video action models with rich, multi-level supervision.

References

S. Belkhale, T. Ding, T. Xiao, P. Sermanet, Q. Vuong, J. Tompson, Y. Chebotar, D. Dwibedi, and D. Sadigh. Rt-h: Action hierarchies using language. arXiv preprint arXiv:2403.01823, 2024.

J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, et al. 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, et al. π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024.

A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818, 2023.

Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint arXiv:2503.06669, 2025.

J. Cai, Z. Cai, J. Cao, Y. Chen, Z. He, L. Jiang, H. Li, H. Li, Y. Li, Y. Liu, et al. Internvla-a1: Unifying understanding, generation and action for robotic manipulation. arXiv preprint arXiv:2601.02456, 2026.

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. URL https://arxiv.org/abs/2506.21539.

K. Chen, S. Xie, Z. Ma, P. R. Sanketi, and K. Goldberg. Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets. arXiv preprint arXiv:2505.15517, 2025a.

X. Chen, Y. Chen, Y. Fu, N. Gao, J. Jia, W. Jin, H. Li, Y. Mu, J. Pang, Y. Qiao, et al. Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy. arXiv preprint arXiv:2510.13778, 2025b.

Z. Chen, Z. Shi, X. Lu, L. He, S. Qian, Z. Yin, W. Ouyang, J. Shao, Y. Qiao, C. Lu, et al. Rh20t-p: A primitive-level robotic dataset towards composable generalization agents. arXiv preprint arXiv:2403.19622, 2024a.

Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24185–24198, 2024b.

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

Q. Dong and Y. Fu. Memflow: Optical flow estimation and prediction with memory. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19068–19078, 2024.

Y. Du, M. Yang, B. Dai, H. Dai, O. Nachum, J. B. Tenenbaum, D. Schuurmans, and P. Abbeel. Learning universal policies via text-guided video generation. arXiv preprint arXiv:2302.00111, 2023. URL https://arxiv.org/abs/2302.00111.

A. O. et al. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. arXiv preprint arXiv:2307.00595, 2023.

C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024a. URL https://arxiv.org/abs/2306.13394.

Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. arXiv preprint arXiv:2401.02117, 2024b.

C. Gao, Z. Liu, Z. Chi, J. Huang, X. Fei, Y. Hou, Y. Zhang, Y. Lin, Z. Fang, Z. Jiang, et al. Vla-os: Structuring and dissecting planning representations and paradigms in vision-language-action models. arXiv preprint arXiv:2506.17561, 2025a.

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, 2025b.

J. Gu, S. Kirmani, P. Wohlhart, Y. Lu, M. G. Arenas, K. Rao, W. Yu, C. Fu, K. Gopalakrishnan, Z. Xu, et al. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. arXiv preprint arXiv:2311.01977, 2023.

H. Huang, F. Lin, Y. Hu, S. Wang, and Y. Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. arXiv preprint arXiv:2403.08248, 2024a.

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, 2025.

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.

W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. arXiv preprint arXiv:2409.01652, 2024b.

P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al. pi-0.5: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054, 2025.

Y. Ji, H. Tan, J. Shi, X. Hao, Y. Zhang, H. Zhang, P. Wang, M. Zhao, Y. Mu, P. An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 1724–1734, 2025.

N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht. Cotracker: It is better to track together. In European conference on computer vision, pages 18–35. Springer, 2024.

A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. arXiv preprint arXiv:2403.12945, 2024.

M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246, 2024.

B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, et al. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024a.

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

Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, M. Liao, F. Wei, Y. Deng, S. Xu, Y. Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. arXiv preprint arXiv:2411.19650, 2024b.

X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu, H. Li, and T. Kong. Vision-language foundation models as effective robot imitators. arXiv preprint arXiv:2311.01378, 2023a.

X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. arXiv preprint arXiv:2405.05941, 2024c.

X. Li, P. Li, M. Liu, D. Wang, J. Liu, B. Kang, X. Ma, T. Kong, H. Zhang, and H. Liu. Towards generalist robot policies: What matters in building vision-language-action models. arXiv preprint arXiv:2412.14058, 2024d.

Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355, 2023b.

Y. Li, Y. Deng, J. Zhang, J. Jang, M. Memmel, R. Yu, C. R. Garrett, F. Ramos, D. Fox, A. Li, et al. Hamster: Hierarchical action models for open-world robot manipulation. arXiv preprint arXiv:2502.05485, 2025b.

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. URL https://arxiv.org/abs/2505.11917.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

F. Liu, K. Fang, P. Abbeel, and S. Levine. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024, 2024a.

Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X.-C. Yin, C.-L. Liu, L. Jin, and X. Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences, 67(12), Dec. 2024b. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

G. Lu, W. Guo, C. Zhang, Y. Zhou, H. Jiang, Z. Gao, Y. Tang, and Z. Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. arXiv preprint arXiv:2505.18719, 2025.

Y. Lu, Y. Fan, B. Deng, F. Liu, Y. Li, and S. Wang. Vl-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes. In 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 976–983. IEEE, 2023.

G. Luo, G. Yang, Z. Gong, G. Chen, H. Duan, E. Cui, R. Tong, Z. Hou, T. Zhang, Z. Chen, et al. Visual embodied brain: Let multimodal large language models see, think, and control in spaces. arXiv preprint arXiv:2506.00123, 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.

S. Nasiriany, F. Xia, W. Yu, T. Xiao, J. Liang, I. Dasgupta, A. Xie, D. Driess, A. Wahid, Z. Xu, et al. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. arXiv preprint arXiv:2402.07872, 2024.

D. Niu, Y. Sharma, G. Biamby, J. Quenum, Y. Bai, B. Shi, T. Darrell, and R. Herzig. Llarva: Vision-action instruction tuning enhances robot learning. arXiv preprint arXiv:2406.11815, 2024.

Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine. Octo: An open-source generalist robot policy. In Proceedings of Robotics: Science and Systems, Delft, Netherlands, 2024.

OpenAI. Gpt-4 technical report. arXiv:2303.08774, 2023.

W. Peebles and S. Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4195–4205, 2023.

K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine. Fast: Efficient action tokenization for vision-language-action models. arXiv preprint arXiv:2501.09747, 2025.

D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. arXiv preprint arXiv:2501.15830, 2025.

N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024. URL https://arxiv.org/abs/2408.00714.

P. Sermanet, T. Ding, J. Zhao, F. Xia, D. Dwibedi, K. Gopalakrishnan, C. Chan, G. Dulac-Arnold, S. Maddineni, N. J. Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 645–652. IEEE, 2024.

L. X. Shi, B. Ichter, M. Equi, L. Ke, K. Pertsch, Q. Vuong, J. Tanner, A. Walling, H. Wang, N. Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. arXiv preprint arXiv:2502.19417, 2025.

A. Singh, V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8317–8326, 2019.

P. Sundaresan, S. Belkhale, D. Sadigh, and J. Bohg. Kite: Keypoint-conditioned policies for semantic manipulation. arXiv preprint arXiv:2306.16605, 2023.

B. R. Team, M. Cao, H. Tan, Y. Ji, M. Lin, Z. Li, Z. Cao, P. Wang, E. Zhou, Y. Han, et al. Robobrain 2.0 technical report. arXiv preprint arXiv:2507.02029, 2025.

G. D. Team. Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023. URL https://arxiv.org/abs/2312.11805.

Y. Tian, Y. Yang, Y. Xie, Z. Cai, X. Shi, N. Gao, H. Liu, X. Jiang, Z. Qiu, F. Yuan, et al. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. arXiv preprint arXiv:2511.16651, 2025.

L. Wang, X. Chen, J. Zhao, and K. He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. Advances in Neural Information Processing Systems, 37:124420–124450, 2024a.

P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024b.

C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel. Any-point trajectory modeling for policy learning. arXiv preprint arXiv:2401.00025, 2023.

K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. arXiv preprint arXiv:2412.13877, 2024.

Z. Wu, Y. Zhou, X. Xu, Z. Wang, and H. Yan. Momanipvla: Transferring vision-language-action models for general mobile manipulation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 1714–1723, 2025.

M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song. Flow as the cross-domain manipulation interface. In 8th Annual Conference on Robot Learning, 2024. URL https://openreview.net/forum?id=cNI0ZkK1yC.

J. Yang, R. Tan, Q. Wu, R. Zheng, B. Peng, Y. Liang, Y. Gu, M. Cai, S. Ye, J. Jang, et al. Magma: A foundation model for multimodal ai agents. arXiv preprint arXiv:2502.13130, 2025a.

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, 2025b.

W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490, 2023.

W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. arXiv preprint arXiv:2406.10721, 2024.

M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine. Robotic control via embodied chain-of-thought reasoning. arXiv preprint arXiv:2407.08693, 2024.

J. Zhang, K. Pertsch, J. Zhang, and J. J. Lim. Sprint: Scalable policy pre-training via language instruction relabeling. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 9168–9175. IEEE, 2024.

Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, et al. 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.

R. Zheng, Y. Liang, S. Huang, J. Gao, H. Daumé III, A. Kolobov, F. Huang, and J. Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. arXiv preprint arXiv:2412.10345, 2024.

E. Zhou, J. An, C. Chi, Y. Han, S. Rong, C. Zhang, P. Wang, Z. Wang, T. Huang, L. Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. arXiv preprint arXiv:2506.04308, 2025a.

Z. Zhou, Y. Zhu, M. Zhu, J. Wen, N. Liu, Z. Xu, W. Meng, R. Cheng, Y. Peng, C. Shen, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. arXiv preprint arXiv:2502.14420, 2025b.