TWO BRIDGES, ONE PATHWAY: FROM VLMS TO GENERALIZABLE VLAS WITH EMBODIED TRAJECTORY-COUPLED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) are powerful general-purpose reasoners, yet converting them into robot control policies (VLAs) is surprisingly difficult. The root cause is a two-fold gap: VLMs are trained on internet-scale images with language-understanding objectives, while VLAs must perceive robot scenes and predict motor actions. Fine-tuning a VLM directly on robot action data forces the model to cross both gaps at once—the learning curve is steep and the rich generalizations learned during pretraining tend to degrade rather than transfer. We argue that this gap can be bridged gradually with the right intermediate data. We introduce *embodied trajectory-coupled (ETC) data*—vision-language supervision derived from the same robot scenes and trajectories used for action learning. Because ETC data shares the visual context of robot operation while retaining familiar language-understanding objectives, it provides a natural stepping stone between VLM pretraining and VLA fine-tuning. Building on this, we design a three-stage training recipe. *Distribution Bridging* first adapts the VLM to embodied visual-language semantics. *Objective Bridging* then gradually shifts the model toward action prediction while preserving the acquired representations. *Retentive Adaptation* finally specializes the policy to the target deployment domain. We further show that mixing task-relevant out-of-distribution ETC data with a small amount of in-distribution action data only enables the model to generalize to novel visual-language conditions without requiring additional robot demonstrations. Simulation and real-robot experiments confirm that this gradual bridging strategy is the key to transferring VLM generalization into robust, deployable robot policies.

1 INTRODUCTION

Building a generalizable vision-language-action (VLA) policy from a pretrained vision-language model (VLM) requires crossing two distinct mismatches at the same time. On the *input* axis, the VLM is trained on broad web-scale visual-language data, while a manipulation policy must operate on embodied scenes saturated with robot-specific viewpoints, object configurations, and interaction geometry. On the *objective* axis, the VLM is optimized for next-token visual-language understanding, while the policy must generate executable actions. These two mismatches are orthogonal: closing one does not automatically close the other.

Prior work has sought to narrow the VLM-to-VLA gap through *embodied reasoning enhancement* before policy training (Team et al., 2025c; Du et al., 2026; Zhai et al., 2025; Yang et al., 2026; Team et al., 2025b) and *multimodal co-training* during VLA adaptation (Cai et al., 2026b; Zitkovich et al., 2023; Intelligence et al., 2025; Zhou et al., 2025; Liu et al., 2026; Cheang et al., 2025). These approaches aim to inherit common sense and visual-language understanding from VLMs (Driess et al., 2023; Zitkovich et al., 2023), enabling stronger embodied scene understanding (Ji et al., 2025; Team et al., 2025c; Yu et al., 2026b) and more generalizable policies (Intelligence et al., 2025; Qu et al., 2026; Wu et al., 2026; Team et al., 2026). The former aims to strengthen embodied understanding before policy learning and has shown value as a VLA initialization, yet the acquired capabilities may still degrade once subsequent training shifts to action-only supervision (Driess et al., 2026; Kachaev et al., 2025; Dalal et al., 2026). The latter maintains multimodal supervision during action learning

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

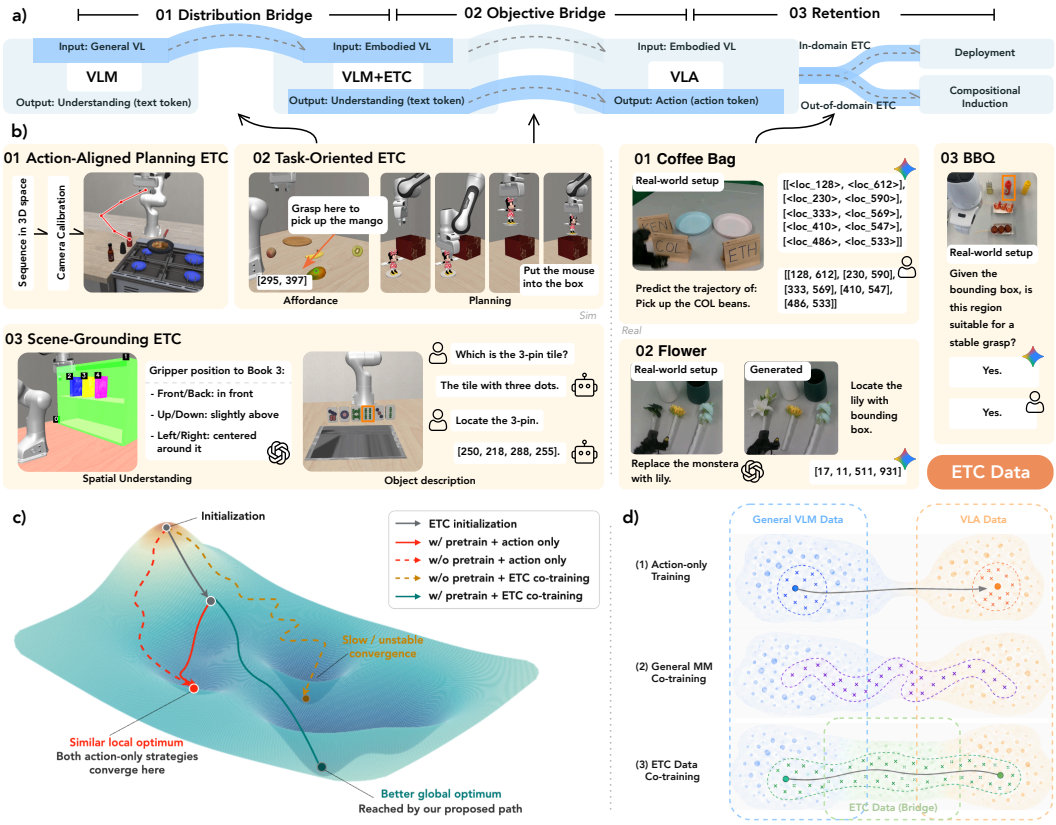


Figure 1: **Overview of the ETC-guided VLM-to-VLA paradigm.** (a) Three-stage ETC-based adaptation pathway: Distribution Bridging, Objective Bridging, and Retentive Adaptation, under a shared next-token prediction objective. (b) ETC examples and the offline construction pipeline. (c) A staged route from VLM initialization toward a generalizable VLA policy, compared with direct action-only adaptation and ETC co-training. (d) ETC data bridges general vision-language data and robot action data, sharing embodied scene semantics with the latter while retaining vision-language supervision.

to preserve and exploit pretrained VLM knowledge. Recent studies have begun to disentangle the effects of different co-training modalities and embodied supervision signals (Lin et al., 2026a;b; Sun et al., 2026; Mazzaglia et al., 2026). However, these efforts mainly study such supervision within a particular training stage or objective, leaving how the same embodied supervision should transfer, persist, and compose across the full VLM-to-VLA adaptation pathway less understood. More broadly, it remains unclear what forms of multimodal supervision can most effectively bridge VLMs and VLAs and how they should be organized across adaptation stages. A key challenge is to acquire manipulation-relevant representations that action learning can exploit, while preventing the shift to action prediction from eroding the embodied information acquired beforehand. We therefore investigate three questions: (1) what kind of multimodal supervision most effectively bridges the dual gap between VLMs and VLAs? (2) how can such supervision be organized into a unified training recipe across adaptation stages to support knowledge transfer and preservation? and (3) can such supervision further expand the policy's capability boundary to unseen visual-language conditions without corresponding action demonstrations?

Our central hypothesis is that the dual VLM-to-VLA mismatch can be bridged through a data-centric intermediate supervision, which we call *embodied trajectory-coupled (ETC) data*. ETC is pair-wise vision-language supervision derived from the same embodied scenes and trajectories as robot action data, shown in Figure 1b. It retains the VLM's next-token prediction objective while matching the embodied content of VLA training. As illustrated in Figure 1d, this dual alignment enables

distribution alignment without action learning, followed by action grounding on an already aligned embodied representation. In this way, ETC separates distribution alignment from action grounding. We further develop a scalable offline pipeline that automatically derives ETC supervision from existing robot trajectories without additional human annotation.

Building on this insight, we organize a three-stage training paradigm under a shared next-token prediction objective (Figure 1a). **(i) Distribution Bridging** fine-tunes the VLM backbone on ETC data, transporting the input distribution from common visual-language data onto the embodied manifold while leaving the objective unchanged. **(ii) Objective Bridging** introduces action-token prediction on this aligned representation while continuing ETC co-training from the same scenes and trajectories, preventing action learning from eroding the embodied alignment established in the first stage. **(iii) Retentive Adaptation** further adapts the policy to deployment domains with continued ETC supervision, maintaining reusable embodied competence while extending the policy's capability boundary. Figure 1d and Figure 1c conceptually illustrate these two roles, respectively: ETC bridges general vision-language data and robot action data, while the three-stage pathway is designed as a smoother optimization route toward generalizable VLA policies than direct adaptation.

Beyond retaining embodied competence, we find that the two stages jointly trained with ETC and action data can support a new capability, which we call *compositional induction*. By introducing task-relevant ETC supervision for unseen visual-language conditions, while pairing it with only limited in-distribution action data, the policy learns to associate new embodied conditions with already acquired action behaviors. This enables generalization to held-out visual-language compositions without collecting action demonstrations for those compositions.

Our contributions are three-fold:

- We introduce **Embodied Trajectory-Coupled (ETC) data**, a data-centric bridge between VLMs and VLAs that is objective-aligned with VLM training and distribution-aligned with robot action data, together with an offline pipeline for deriving ETC supervision from existing trajectories.
- We develop a three-stage ETC-based adaptation pathway—**Distribution Bridging**, **Objective Bridging**, and **Retentive Adaptation**—under a shared next-token prediction objective, progressively aligning embodied understanding, action generation, and downstream adaptation.
- We identify **Compositional Induction**: task-relevant out-of-distribution ETC, paired with limited in-distribution action data, enables policies to generalize to held-out visual-language compositions without action demonstrations for those conditions.

2 Related Work

Knowledge Transfer and Preservation in Vision-Language-Action Models. Adapting VLMs into generalizable VLA policies requires transferring multimodal knowledge into action learning while preserving the semantic and reasoning capabilities acquired during pretraining (Black et al., 2024; Bjorck et al., 2025; Liu et al., 2025; Zawalski et al., 2024; Ishitoya et al., 2026). Recent work explores multimodal co-training and embodied adaptation to improve generalization and retain visiolinguistic competence (Lin et al., 2026a; Zhang et al., 2026; Hancock et al., 2025; Cai et al., 2026a; Chen et al., 2025b; Grover et al., 2025; Li et al., 2026b; Yu et al., 2026a; Li et al., 2025), while others mitigate interference from action learning by insulating or pretraining the action expert before joint VLA optimization (Driess et al., 2026; Xu et al., 2026; Huang et al., 2026). However, two gaps remain. First, generic multimodal co-training relies on supervision whose distribution is distant from embodied action data, making it an indirect bridge from VLM knowledge to manipulation-relevant behavior (Lin et al., 2026a; Chen et al., 2025a; Ji et al., 2025). Second, subsequent action adaptation can degrade previously acquired vision-language representations and the generalization capabilities they support (Du et al., 2026; Zhang et al., 2026; Jiang et al., 2025; Chen et al., 2025b; Kachaev et al., 2025; Xu et al., 2026). More broadly, existing studies often examine these adaptation mechanisms as separate interventions, leaving how they interact when composed across successive VLM-to-VLA stages underexplored. These limitations motivate a multi-stage adaptation framework that explicitly studies how embodied supervision supports knowledge transfer, retention, and downstream capability expansion across VLA training stages.

Embodied Understanding and Reasoning in Vision-Language Models. A parallel line of work equips VLMs with manipulation-relevant understanding through embodied multimodal supervision.

Before policy training, robotics-oriented VQA (Sermanet et al., 2024; Chen et al., 2025a; Hancock et al., 2025), spatial grounding (Du et al., 2024; Zhou et al., 2026; Yuan et al., 2024; Qu et al., 2025), intermediate-representation learning (Li et al., 2026a), and embodied foundation models (Lee et al., 2025; Ji et al., 2025; Team et al., 2025a;c; Yang et al., 2025) introduce manipulation-relevant priors over spatial and interaction reasoning, affordances, planning, and temporal dynamics (Hao et al., 2025; Zhai et al., 2025; Chen et al., 2025b; Ma et al., 2026). Yet these studies largely treat embodied supervision as either an initialization source or a task-specific auxiliary signal. Our work instead studies trajectory-coupled supervision as a shared bridge across adaptation stages, and analyzes its role in inducing compositional generalization to unseen visual-language conditions.

3 METHOD

3.1 EMBODIED TRAJECTORY-COUPLED DATA

We define *Embodied Trajectory-Coupled (ETC) data* as vision-language supervision derived from the same embodied scenes and trajectories as robot action data. ETC retains the VLM's next-token prediction objective while matching the VLA embodied distribution. This dual alignment enables distribution bridging before action learning and action grounding over aligned representations.

We instantiate ETC as VQA-style supervision in three categories of decreasing abstraction, forming a graded spectrum from common-sense understanding to motor behaviors (examples in Figure 1b). **Scene-Grounding ETC** captures task-agnostic scene perception: spatial queries probe geometric relationships (relative direction to the end effector, cross-view object correspondence), and object-description queries identify, localize, or describe manipulation-relevant attributes of a target object. **Task-Oriented ETC** is instruction-conditioned: affordance queries predict optimal grasp points or validate whether a region affords a stable grasp, and task-planning queries decompose an instruction into ordered sub-actions, recognize the current phase, or anticipate the next action. **Action-Aligned Planning ETC** sits closest to motion, forecasting the gripper's future trajectory as five equidistant image-plane waypoints sampled along the recorded end-effector trajectory for a given sub-task—the final stepping stone before action tokens (Gu et al., 2023). **ETC Construction.** Our ETC construction pipeline is grounded in recorded robot trajectories. Ground-truth targets are obtained either from structured trajectory signals or from Gemini-2.5-Pro generation: geometry- and motion-related targets are computed directly from robot data, using camera calibration or by unifying actions into the camera frame when calibration is unavailable (Xie et al., 2026), while semantic and reasoning labels are generated by Gemini and manually verified. The resulting targets are paired with predefined question templates to form VQA-style supervision for training. Full question templates, construction details, and examples are provided in Appendix C.1.

3.2 THREE-STAGE ETC-BASED ADAPTATION

We adapt a pretrained VLM into a VLA through three stages, all sharing the same next-token prediction objective. The stages map directly onto the dual VLM-to-VLA mismatch: Stage 1 closes the input-distribution gap on the VLM objective; Stage 2 closes the objective gap on the already-aligned distribution; Stage 3 keeps both gaps closed while shifting to deployment scenes.

Distribution Bridging (Stage 1). The first stage operates purely on the VLM backbone and transports its input distribution from general visual-language data onto the embodied manifold, while leaving the next-token objective unchanged. We train the backbone on ETC data alone, spanning all three categories defined in Section 3.1, and unfreeze the ViT tower so that embodied supervision can align semantic visual features with action-relevant geometry.

Objective Bridging (Stage 2). The second stage introduces the action objective on top of the aligned distribution from Stage 1, co-training large-scale pretraining action data with ETC drawn from the same scenes and trajectories. ETC's role shifts from *establishing* the embodied distribution to *preserving* it, so that the action objective extends the model from understanding to generation without overwriting Stage 1's alignment.

Retentive Adaptation (Stage 3). The third stage adapts the policy to downstream deployment scenes by retaining the co-training scheme of Stage 2, now pairing action data with ETC constructed for the target scenes. The input distribution shifts to new objects, viewpoints, and layouts, but ETC's

dual-alignment property carries forward: continued ETC supervision preserves reusable embodied competence, while the action objective extends the policy's capability boundary onto the target tasks. Stage 3 thus realizes *capability expansion without knowledge erosion* as a training recipe.

3.3 Compositional Induction via ETC Co-training

ETC co-training further enables *compositional induction*. For held-out target compositions, we provide ETC supervision but no corresponding action demonstrations. Through the shared backbone, the policy combines conditions learned from ETC with action skills learned from in-distribution demonstrations, enabling generalization to unseen compositions. We validate this in simulation and on real robots.

4 Experiments

4.1 Experiment Setup

Training and Evaluation. We initialize Stage 1 from PaliGemma (Beyer et al., 2024) and initialize the Stage 2 VLA from the resulting Distribution-Bridged VLM, following PiFAST (Pertsch et al., 2025) for action tokenization and training. ETC and discretized action tokens are trained under a unified next-token prediction formulation. During co-training, each optimizer update accumulates gradients from separate action and ETC batches, with their batch sizes set at an ETC-to-action ratio of 1:4. For MM ablations, we replace the ETC stream with generic multimodal data at the same ratio. Full training details are provided in Appendix A. We evaluate on three simulation benchmarks—LIBERO (Liu et al., 2023), SimplerEnv (Li et al., 2024), and VLABench (Zhang et al., 2025)—and a real-world WidowX platform with two tasks: *Inserting Flower* and *Placing Coffee Bag*, both instantiated on a BridgeV2-pretrained Pifast (Walke et al., 2023). These benchmarks vary in how much they require embodied understanding beyond action execution alone, and we report success rate under each benchmark's standard protocol. Each benchmark is paired with domain-matched ETC supervision: ShareRobot (Ji et al., 2025) and VLA-OS (Gao et al., 2026) are adopted for SimplerEnv and LIBERO, respectively, while VLABenchVQA and real-world ETC are constructed in-house. Full statistics are in Appendix C.

Embodied Understanding Benchmarks. We also probe the VLM backbone's multimodal understanding on a held-out set of ETC and general multimodal data, evaluated on both the initialized VLM and the midtrain VLA backbone to track whether embodied understanding is acquired and preserved across stages. Full evaluation details are provided in Appendix D.

4.2 Analyzing the Three-Stage ETC-based Adaptation Pathway

Table 1: **Results across Distribution Bridging and Objective Bridging settings.** Rows are grouped by Stage 2 supervision; *w/ Distribution Bridging* annotations report absolute score changes over the corresponding *w/o Distribution Bridging* row.

				VLABench T1			VLABench T2		
Stage 2 signal	Backbone / Stage 1 init.	Simpler	LIBERO	SR	PS	IS	SR	PS	IS
Action only	SigLIP–Gemma	38.5	**91.5**	–	–	–	–	–	–
	w/o Distribution Bridging	79.2	89.4	37.2	48.2	50.8	23.6	31.2	30.6
	w/ Distribution Bridging	81.2 ↑2.0	90.7 ↑1.3	44.3 ↑7.1	53.1 ↑4.9	53.7 ↑2.9	28.3 ↑4.7	35.1 ↑3.9	36.5 ↑5.9
+ MM	w/o Distribution Bridging	67.7	81.6	38.1	49.0	51.8	24.1	32.8	35.3
	w/ Distribution Bridging	72.9 ↑5.2	83.3 ↑1.7	61.0 ↑22.9	72.6 ↑23.6	74.0 ↑22.2	39.3 ↑15.2	46.4 ↑13.6	47.3 ↑12.0
+ ETC	w/o Distribution Bridging	80.2	86.9	42.4	52.5	51.9	22.6	31.6	36.4
	w/ Distribution Bridging	**84.4** ↑4.2	86.8 ↓0.1	**64.8** ↑22.4	**77.3** ↑24.8	**77.8** ↑25.9	**41.4** ↑18.8	**49.8** ↑18.2	**51.1** ↑14.7
+ MM+ETC	w/o Distribution Bridging	70.8	82.5	41.2	49.8	49.3	26.0	33.4	32.7
	w/ Distribution Bridging	72.9 ↑2.1	90.1 ↑7.6	61.1 ↑19.9	72.7 ↑22.9	74.4 ↑25.1	36.9 ↑10.9	44.9 ↑11.5	46.3 ↑13.6

4.2.1 DISTRIBUTION BRIDGING: ALIGNING THE VLM TO THE EMBODIED MANIFOLD

Effective Distribution Bridging requires adapting embodied representations across both vision and language. We show that this stage provides a stronger basis for subsequent action learning, with its benefit arising primarily from joint vision-language adaptation rather than language-only tuning.

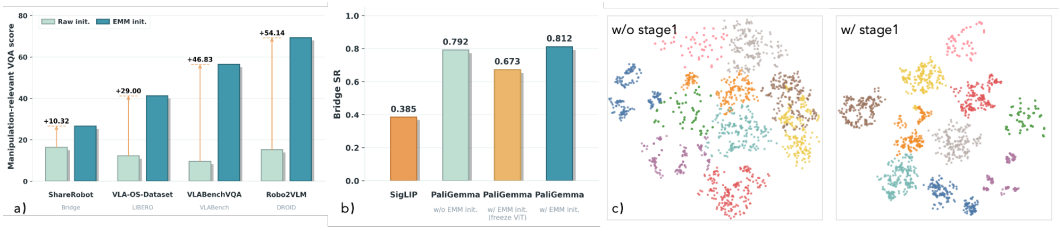


Figure 2: **Distribution Bridging aligns the VLM to the embodied manifold.** (a) ETC supervision improves performance on embodied understanding benchmarks. (b) Stage 1 strategies with identical action-only Stage 2 training. (c) t-SNE comparison on 10 VLABench tasks: w/ vs. w/o Stage 1.

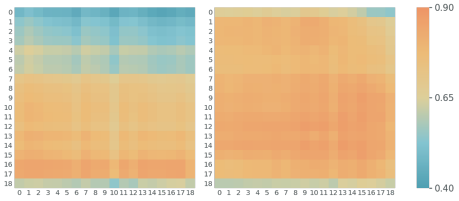
Does Distribution Bridging Provide a Better Starting Point for Action Learning? The Distribution-Bridged VLM backbone consistently outperforms the raw PaliGemma backbone across nearly all Stage 2 settings, with the largest gains on VLABench (Table 1). Under action-only training, which isolates the effect of initialization, Distribution Bridging improves SR by $+2.0$ points on SimplerEnv, $+1.3$ on LIBERO, $+7.1$ on VLABench T1, and $+4.7$ on T2. The Distribution-Bridged VLM achieves higher ETC-bench scores than the raw VLM (Figure 2a), and the t-SNE visualization (Figure 2c) further shows that the Distribution-Bridged VLM separates task-conditioned embodied inputs that the raw VLM mixes together, suggesting embodied task structure has been injected into the VLM representation as a stronger initialization.

Why Is Distribution Bridging Needed beyond General VLM Pretraining? Figure 2b supports our claim that general VLM representations remain misaligned with embodied manipulation scenes. With identical action-only Stage 2 training, raw PaliGemma clearly outperforms SigLIP–Gemma (79.2 vs. 38.5), showing the value of joint vision–language pretraining. Yet ETC-based Distribution Bridging further improves SR by $+2.0$ points to 81.2, indicating that general VLM pretraining alone is insufficient for embodied alignment. Freezing the ViT during bridging instead reduces SR to 67.3, further suggesting that the remaining gap requires visual adaptation, not just language-side tuning.

4.2.2 OBJECTIVE BRIDGING: PRESERVING EMBODIED ALIGNMENT UNDER ACTION LEARNING

Is ETC Co-training Still Needed after Distribution Bridging? We focus on SimplerEnv and VLABench, since LIBERO is less diagnostic for embodied preservation: even SigLIP–Gemma reaches 91.5 SR with action-only training (Table 1). We also exclude SigLIP–Gemma from subsequent analyses because its lack of joint vision–language pretraining makes it unsuitable for evaluating ETC's preservation effect. With the same Distribution-Bridged backbone, ETC co-training improves SR by $+3.2$, $+20.5$, and $+13.1$ points on SimplerEnv, VLABench T1, and T2, respectively. CKA analysis (Figure 3a) shows that ETC co-training keeps the Stage 2 backbone substantially closer to the Distribution-Bridged representation than action-only training, indicating reduced representational forgetting. Consistently, attention visualizations (Figure 4a) show more concentrated attention on task-relevant objects and interaction regions. Held-out embodied-understanding evaluation further confirms this preservation: ETC co-training achieves ShareRobot/VLA-OS scores of $91.92/60.00$, compared with $0.00/0.05$ under action-only training (Table 13). Together, these results show that ETC co-training preserves both the embodied representation and task-relevant visual-language grounding acquired during Distribution Bridging.

Why Is ETC More Effective than General Multimodal Supervision? Under the same Distribution-Bridged backbone and action data, ETC co-training achieves higher SR than general MM co-training, with absolute gains of 11.5, 3.5, 3.8, and 2.1 points on SimplerEnv, LIBERO, VLABench Track 1, and Track 2, respectively (Table 1). This suggests that the effectiveness of auxiliary supervision depends on its alignment with action learning. ETC is trajectory-coupled and manipulation-aligned, more closely matching the distribution of robot action data, whereas generic MM data is broader



Orange: more similar layer representations; cyan: less similar.

(a)

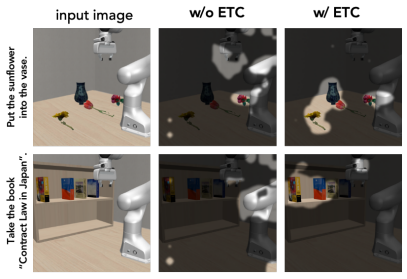
Stage 2 supervision	Track 1			Track 2		
	SR	PS	IS	SR	PS	IS
Action only	44.3	53.1	53.7	28.3	35.1	36.5
+ Scene-grounding ETC	56.2	69.6	73.6	40.7	48.1	47.3
+ Task-oriented ETC	59.3	71.9	74.8	31.4	42.3	50.7
+ Action Planning ETC	62.1	74.4	77.1	**43.8**	**49.9**	48.2
+ Full ETC	**64.8**	**77.3**	**77.8**	41.4	49.8	**51.1**

(b)

Figure 3: **Objective Bridging: ETC co-training preserves embodied alignment under action learning.** (a) CKA between the Stage 2 backbone and the Distribution-Bridged VLM. Left: action-only Objective Bridging; right: ETC co-training. (b) VLABench performance under different ETC category ablations during Objective Bridging on Track 1 (in-domain) and Track 2 (cross-category).

but distributionally farther from robot interaction and thus less directly policy-relevant. Notably, combining MM with ETC also underperforms ETC alone on most metrics, suggesting that generic MM is not simply complementary replay but may compete with manipulation-relevant supervision during action learning. Thus, the benefit of multimodal co-training depends not merely on replaying vision-language knowledge, but on whether the auxiliary supervision is aligned with the embodied distribution of policy learning.

Which Category of ETC Benefits More? We ablate three ETC categories under the same Distribution-Bridged backbone and action data (Figure 3b). All outperform action-only training, but their gains reflect different levels of proximity to action learning. Action-Aligned Planning ETC, which directly predicts future gripper trajectories and lies closest to the action objective, achieves the highest execution scores among individual categories on both Track 1 and the cross-category Track 2, suggesting that its motion-coupled geometry transfers across task compositions. Task-Oriented ETC instead emphasizes affordance and task reasoning, and yields the strongest understanding among individual categories, reaching a Track 2 IS of 50.7. Scene-Grounding ETC provides more general perceptual and spatial supervision, improving all metrics without dominating any. Combining all three categories yields the highest scores across all Track 1 metrics and further improves Track 2 IS to 51.1, while Action-Aligned Planning ETC retains slightly higher Track 2 execution scores (43.8/49.9 SR/PS). These results suggest that ETC forms a complementary supervision spectrum: supervision closer to action learning primarily benefits execution, while more abstract embodied supervision strengthens perception and understanding.



(a)

Method	LIBERO	Simpler	VLABench SR T1	T2	T3	T4	Pretrain
Ours	86.8	**84.4**	**64.8**	**41.4**	**31.0**	**31.2**	✗
$\pi_{0.5}$	85.9	35.4	53.2	33.2	15.2	5.6	✗
π_0-FAST	**89.4**	79.2	37.2	23.6	26.1	22.5	✗
$\pi_{0.5}$	**96.9**	46.9	**66.5**	38.2	30.5	31.9	✓
π_0-FAST	85.5	**76.5**	65.5	**46.9**	**43.4**	**52.8**	✓

(b)

Figure 4: **Objective Bridging improves manipulation-relevant representations and achieves competitive downstream performance.** (a) Attention visualizations on representative VLABench scenes, comparing action-only Objective Bridging with ETC co-training. (b) Performance after Distribution and Objective Bridging, compared with representative VLAs. ✗denotes models trained without initialization from a large-scale pretrained VLA checkpoint, while ✓denotes initialization from a pretrained VLA base model.

How Do Distribution and Objective Bridging Interact? ETC co-training yields substantially larger gains after Distribution Bridging. On VLABench Track 1, adding ETC over action-only training improves SR by only 5.2 points without Distribution Bridging, compared to 20.5 points with it. On

Track 2, the corresponding gain shifts from -1.0 to $+13.1$ points. This corresponds to interaction effects of $+15.3$ and $+14.1$ points on T1 and T2, respectively, with similar amplification in PS and IS. These results demonstrate the complementary benefits of the two bridging stages, particularly for cross-category generalization.

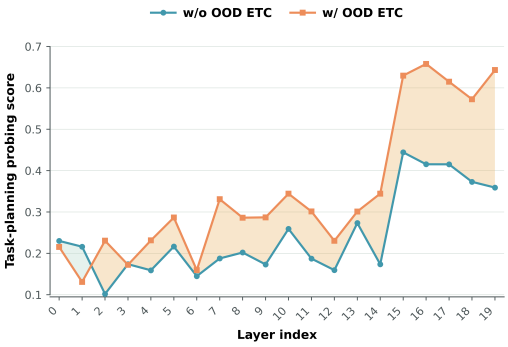
Comparison with Representative VLAs. We further compare our policy after both bridging stages with representative VLAs to provide absolute performance context (Figure 4b). Among methods without pretrained VLA initialization, ours achieves the highest scores on five of six metrics, including 84.4 on SimplerEnv and 64.8/41.4/31.0/31.2 on VLABench T1–T4. Compared with pretrained VLAs, ours achieves higher SimplerEnv SR (84.4 vs. 46.9/76.5) and comparable performance on VLABench T1 (64.8 vs. 66.5/65.5), while pretrained methods retain advantages on other benchmarks. Since these methods differ in their pretraining recipes and action data, we treat this comparison as performance context rather than a strictly controlled comparison.

4.2.3 Retentive Adaptation: Extending Capability without Knowledge Erosion

The real-robot evaluation further validates **Retentive Adaptation** on a downstream deployment task (Figure 5a). Consistent with the simulation results, continuing ETC supervision through Stage 2 and Stage 3 yields the strongest real-robot performance, whereas introducing ETC only during Stage 3 is substantially less effective. This shows that Stage 3 benefits from adapting a policy with reusable embodied competence retained from the preceding stages.

Stage 1	Stage 2	Stage 3	In-domain Score
✗	✗	✗	66.67
✗	✗	✓	61.11
✗	✓	✗	75.00
✗	✓	✓	**95.83**
✓	✗	✗	66.67
✓	✗	✓	72.22
✓	✓	✗	<u>91.67</u>
✓	✓	✓	**95.83**
From scratch			63.88

(a)



(b)

Figure 5: (a) **Retentive Adaptation: extending capability without knowledge erosion.** Three-stage ETC adaptation on the WidowX *pick coffee bag into the plate* task. ✓/✗ denotes whether ETC supervision is enabled/disabled at each stage; action training at Stages 2 and 3 is always performed. All stage-wise variants undergo BridgeV2 action pretraining, whereas *From scratch* skips BridgeV2 pretraining and trains the PaliGemma-initialized π_0-FAST directly on the downstream task. (b) **OOD ETC strengthens task-planning representations.** Layer-wise task-planning probing on VLABench Track 2 with and without OOD ETC.

4.3 Compositional Induction via Constructible OOD ETC

Beyond preserving and adapting in-distribution knowledge, ETC enables compositional induction to held-out visual–language conditions. Throughout these experiments, OOD information is introduced only through ETC supervision, while all action data remain strictly in-domain; OOD trajectories are used only to construct ETC; their actions are never used for action training. This setting allows us to examine whether ETC can associate unseen visual–language conditions with manipulation behaviors acquired solely from in-domain action data.

4.3.1 OOD ETC Induces Compositional Generalization

We first evaluate this effect on VLABench Track 2 (Table 2). We augment mid-training with OOD ETC at a 94:6 in-domain to out-of-distribution ratio while keeping all action demonstrations strictly in-domain. Compared with ID ETC only, adding OOD ETC improves IS by 2.2 points (51.1 to 53.3), while SR and PS decrease by 2.5 and 1.6 points, respectively. The layer-wise task-planning

probing in Figure 5b shows higher scores with OOD ETC, particularly in later layers, suggesting more accessible task-planning information for held-out conditions. Together with the higher IS but lower SR/PS, this indicates improved understanding without corresponding gains in action execution.

Table 2: **Compositional induction via OOD ETC supervision** on VLABench Track 2 and two real-robot tasks. All scores are reported on a 0–100 scale. Real-robot variants share the same Distribution-Bridged and Objective-Bridged initialization.

Simulation evaluation				
Benchmark	Stage 2 supervision	SR	PS	IS
VLABench Track 2	ID ETC only	**41.4**	**49.8**	51.1
	ID ETC + OOD ETC	38.9	48.2	**53.3**
Real-robot evaluation				
Task	Stage 3 supervision	ID Score	OOD Score	
Coffee bag	Action only	<u>91.67</u>	41.67	
	+ ID ETC	**95.83**	69.44	
	+ OOD ETC (Human-derived)	<u>91.67</u>	<u>80.56</u>	
	+ OOD ETC (Gemini)	<u>91.67</u>	**91.67**	
Flower	Action only	**47.22**	27.08	
	+ ID ETC	<u>45.83</u>	27.08	
	+ OOD ETC (Human-derived)	37.50	**50.00**	
	+ OOD ETC (GPT-Image-2)	38.89	<u>37.50</u>	

We then validate whether this effect extends to real-robot deployment. For real-robot evaluation, OOD conditions comprise unseen bag-label–plate-color combinations in the coffee-bag task and an unseen flower category in the flower task; these conditions are excluded from action demonstrations. To isolate the contribution of downstream ETC, all variants start from policies trained with both Distribution Bridging and Objective Bridging and differ only in their Stage3 supervision (Table 2). On the coffee-bag task, OOD ETC improves OOD performance from 41.67 under action-only adaptation to 80.56, and further outperforms in-domain ETC (69.44). On the flower task, in-domain ETC yields no improvement over action-only adaptation (27.08 for both), whereas OOD ETC raises OOD performance to 50.00. These results suggest that explicitly covering unseen deployment conditions with ETC provides additional OOD gains beyond in-domain ETC supervision.

Together, the simulation and real-robot results demonstrate that targeted OOD ETC composes unseen visual–language conditions with previously acquired manipulation behaviors, enabling generalization without corresponding OOD action demonstrations.

4.3.2 Cost-Efficient Construction of OOD ETC.

Effective ETC for novel conditions need not rely on costly manual annotation or additional robot demonstrations. In the coffee-bag task, Gemini-generated annotations on real-world images outperform human annotations, improving the score from 80.56 to **91.67**. In the flower task, photorealistic target-condition images generated by GPT-Image-2 enable ETC construction without collecting new visual observations, improving the score from 27.08 to 37.50, although still below the real-image, human-annotated setting of 50.00. Together, automated annotation and generative visual construction provide a scalable route to compositional induction without new action demonstrations.

5 Conclusion

We formulate VLM-to-VLA adaptation as a dual mismatch in input distribution and training objective, and propose an ETC-guided three-stage pathway to bridge these gaps. Our results show that ETC provides a stronger VLA initialization, remains more effective than general multimodal supervision during action learning, and supports reliable adaptation on real-robot deployment tasks. Experiments in simulation and on real robots demonstrate that this pathway improves generalizable policy learning. Beyond this staged adaptation, low-cost OOD ETC enables *compositional induction*, allowing policies to generalize to unseen visual-language compositions without collecting corresponding OOD action demonstrations.

AI USE STATEMENT

We used Gemini-2.5-Pro to assist in generating semantic and reasoning annotations for ETC data, and GPT-Image-2 to synthesize OOD images for real-robot experiments. We also used ChatGPT to assist with manuscript organization, language editing, and experimental analysis. All AI-assisted content was reviewed and validated by the authors, who take full responsibility for the methodology, experimental results, and conclusions.

REPRODUCIBILITY STATEMENT

We provide implementation details and training configurations in Appendix A, ETC data construction procedures and statistics in Appendix C, and evaluation protocols in Appendix D. Additional real-world experimental details are provided in Appendix E to facilitate reproduction of our results.

REFERENCES

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

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

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.

Lukas Blecher. LaTeX-OCR: pix2tex – using a ViT to convert images of equations into LaTeX code. `https://github.com/lukas-blecher/LaTeX-OCR`, 2022. Software repository, accessed 2026-05-28.

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*, 2026a.

Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, et al. Xiaomi-robotics-0: An open-sourced vision-language-action model with real-time execution. *arXiv preprint arXiv:2602.12684*, 2026b.

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

Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R. Sanketi, and Ken Goldberg. Robo2VLM: Visual question answering from large-scale in-the-wild robot manipulation datasets. *arXiv preprint arXiv:2505.15517*, 2025a.

William Chen, Suneel Belkhale, Suvir Mirchandani, Oier Mees, Danny Driess, Karl Pertsch, and Sergey Levine. Training strategies for efficient embodied reasoning. *arXiv preprint arXiv:2505.08243*, 2025b.

Dwip Dalal, Shivansh Patel, Chahit Jain, Jeonghwan Kim, Utkarsh Mishra, Alex Baratian, Hyeonjeong Ha, Heng Ji, Svetlana Lazebnik, and Unnat Jain. Generalizable vla finetuning via representation anchoring and language-action alignment. *arXiv preprint arXiv:2607.13429*, 2026.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Danny Driess, Jost Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *Advances in Neural Information Processing Systems*, 38:102867–102888, 2026.

Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024.

Yiyang Du, Zhanqiu Guo, Xin Ye, Liu Ren, and Chenyan Xiong. Embodiedmidtrain: Bridging the gap between vision-language models and vision-language-action models via mid-training. *arXiv preprint arXiv:2604.20012*, 2026.

Chongkai Gao, Zixuan Liu, Zhenghao Chi, Junshan Huang, Xin Fei, Yiwen Hou, Yuxuan Zhang, Yudi Lin, Zhirui Fang, and Lin Shao. Vla-os: Structuring and dissecting planning representations and paradigms in vision-language-action models. *Advances in Neural Information Processing Systems*, 38:136705–136736, 2026.

Shresth Grover, Akshay Gopalkrishnan, Bo Ai, Henrik I Christensen, Hao Su, and Xuanlin Li. Enhancing generalization in vision-language-action models by preserving pretrained representations. *arXiv preprint arXiv:2509.11417*, 2025.

Jiayuan Gu, Sean Kirmani, Paul Wohlhart, Yao Lu, Montserrat Gonzalez Arenas, Kanishka Rao, Wenhao Yu, Chuyuan Fu, Keerthana Gopalakrishnan, Zhuo Xu, et al. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. *arXiv preprint arXiv:2311.01977*, 2023.

Asher J Hancock, Xindi Wu, Lihan Zha, Olga Russakovsky, and Anirudha Majumdar. Actions as language: Fine-tuning vlms into vlas without catastrophic forgetting. *arXiv preprint arXiv:2509.22195*, 2025.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, et al. Mimo-embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025.

Chengyue Huang, Mellon M Zhang, Robert Azarcon, Glen Chou, and Zsolt Kira. Maps: Preserving vision-language representations via module-wise proximity scheduling for better vision-language-action generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32451–32462, 2026.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL `https://arxiv.org/abs/2504.16054`.

Yuri Ishitoya, Jeremy Siburian, Masashi Hamaya, Kuniaki Saito, Cristian C Beltran-Hernandez, and Mai Nishimura. Clap: Direct vlm-to-vla adaptation via language-action grounding. *arXiv preprint arXiv:2607.08974*, 2026.

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1724–1734, 2025.

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

Nikita Kachaev, Mikhail Kolosov, Daniil Zelezetsky, Alexey K Kovalev, and Aleksandr I Panov. Don't blind your vla: Aligning visual representations for ood generalization. *arXiv preprint arXiv:2510.25616*, 2025.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

Hao Li, Zi-han Ding, Shuai Yang, Yilun Chen, Yang Tian, Xiaolin Hu, Tai Wang, Dahua Lin, Feng Zhao, Si Liu, et al. Robointer: A holistic intermediate representation suite towards robotic manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 113314–113381, 2026a.

Haoyang Li, Guanlin Li, Youhe Feng, Chen Zhao, Zhuoran Wang, Yang Li, Qizhe Wei, Shifeng Bao, Haitao Shen, Yihan Zhao, et al. Training vision-language-action models with dense embodied chain-of-thought supervision. *arXiv preprint arXiv:2606.30552*, 2026b.

Jinming Li, Yichen Zhu, Zhibin Tang, Junjie Wen, Minjie Zhu, Xiaoyu Liu, Chengmeng Li, Ran Cheng, Yaxin Peng, Yan Peng, et al. Coa-vla: Improving vision-language-action models via visual-textual chain-of-affordance. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9759–9769. IEEE, 2025.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

Fanqi Lin, Kushal Arora, Jean Mercat, Haruki Nishimura, Paarth Shah, Chen Xu, Mengchao Zhang, Mark Zolotas, Maya Angeles, Owen Pfannenstiehl, et al. A systematic study of data modalities and strategies for co-training large behavior models for robot manipulation. *arXiv preprint arXiv:2602.01067*, 2026a.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Weifeng Lin, Siyuan Huang, Hao Li, Tingwei Chen, Ruichuan An, Xinyu Wei, Jianbo Liu, and Hongsheng Li. Rethinking vlm representation for vla initialization. *arXiv preprint arXiv:2605.25802*, 2026b.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Yicheng Liu, Shiduo Zhang, Zibin Dong, Baijun Ye, Tianyuan Yuan, Xiaopeng Yu, Linqi Yin, Chenhao Lu, Junhao Shi, Luca Jiang-Tao Yu, et al. Faster: Toward efficient autoregressive vision language action modeling via neural action tokenization. *arXiv preprint arXiv:2512.04952*, 2025.

Yicheng Liu, Zibin Dong, Baijun Ye, Tianyuan Yuan, Tao Jiang, Anqi Yang, Shicheng Cao, Haonan Liu, Yue Sun, Zihan Guo, et al. G0.5: One autoregressive stream for robot reasoning and action. *arXiv preprint arXiv:2608.11739*, 2026.

Xiangkai Ma, Lekai Xing, Han Zhang, Wenzhong Li, and Sanglu Lu. Unifying perception and action: A hybrid-modality pipeline with implicit visual chain-of-thought for robotic action generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22380–22390, 2026.

Pietro Mazzaglia, Cansu Sancaktar, Markus Peschl, and Daniel Dijkman. Hybrid training for vision-language-action models. In *International Conference on Learning Representations*, volume 2026, pp. 135535–135556, 2026.

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

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

Delin Qu, Haoming Song, Qizhi Chen, Zhaoqing Chen, Xianqiang Gao, Dong Wang, Xinyi Ye, Qi Lv, Modi Shi, Guanghui Ren, Cheng Ruan, Maoqing Yao, Haoran Yang, Jiacheng Bao, Bin Zhao, and Xuelong Li. Eo-1: An open unified embodied foundation model for general robot control, 2026. URL `https://arxiv.org/abs/2508.21112`.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 645–652. IEEE, 2024.

Nan Sun, Yuan Zhang, Yongkun Yang, Wentao Zhao, Peiyan Li, Jun Guo, Wenxuan Song, Pengxiang Ding, Runze Suo, Yifei Su, et al. Revisiting embodied chain-of-thought for generalizable robot manipulation. *arXiv preprint arXiv:2606.03784*, 2026.

BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025a.

Gemini Robotics Team, Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, et al. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*, 2025b.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025c.

Xiaomi Robotics Team, Jun Guo, Piaopiao Jin, Jason Li, Peiyan Li, Yingyan Li, Futeng Liu, Wanli Peng, Optimus Qin, Yifei Su, et al. Xiaomi-robotics-1: Scaling vision-language-action models with over 100k hours of real-world trajectories. *arXiv preprint arXiv:2607.15330*, 2026.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, et al. A pragmatic vla foundation model. *arXiv preprint arXiv:2601.18692*, 2026.

Sicheng Xie, Lingchen Meng, Zijie Diao, Haidong Cao, Zhiying Du, Shuyuan Tu, Jiaqi Leng, Qiuyue Wang, Mingsheng Li, Shuai Bai, Zuxuan Wu, and Yu-Gang Jiang. Unify robot actions in camera frame, 2026. URL https://arxiv.org/abs/2511.17001.

Kechun Xu, Zhenjie Zhu, Anzhe Chen, Rong Xiong, and Yue Wang. Apt: Action expert pre-training improves instruction generalization of vision-language-action policies. *arXiv preprint arXiv:2606.12366*, 2026.

Ganlin Yang, Tianyi Zhang, Haoran Hao, Weiyun Wang, Yibin Liu, Dehui Wang, Guanzhou Chen, Zijian Cai, Junting Chen, Weijie Su, et al. Vlaser: Vision-language-action model with synergistic embodied reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 65736–65765, 2026.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14203–14214. IEEE, 2025.

Qize Yu, Jiadi You, Yuran Wang, Jiaqi Liang, Bowen Ping, Yang Tian, Yue Chen, Minghong Cai, Zeying Gong, Ruihai Wu, et al. Affordancevla: A vision-language-action model empowering action generation through affordance-aware understanding. *arXiv preprint arXiv:2606.06155*, 2026a.

Ryan Yu, Pushi Zhang, Starrick Liu, Brae Liu, Miracle Kang, Shalfun Li, Lights Shi, Ellie Ma, Ping Yang, Chris Pan, et al. Wall-oss-0.5 technical report. *arXiv preprint arXiv:2605.30877*, 2026b.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

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

Jianke Zhang, Xiaoyu Chen, Qiuyue Wang, Mingsheng Li, Yanjiang Guo, Yucheng Hu, Jiajun Zhang, Shuai Bai, Junyang Lin, and Jianyu Chen. Vlm4vla: Revisiting vision-language-models in vision-language-action models. *arXiv preprint arXiv:2601.03309*, 2026.

Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11142–11152, 2025.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *Advances in Neural Information Processing Systems*, 38: 28404–28481, 2026.

Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Yaxin Peng, Chaomin Shen, Feifei Feng, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5377–5395, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

A Implementation Details

A.1 Co-training Strategy

During both Objective Bridging and Retentive Adaptation, we co-train on ETC and action data. Each optimization step combines one action batch with one ETC batch: we compute the action loss and ETC loss in separate forward/backward passes, accumulate or explicitly sum their gradients, and then perform a single optimizer update on the combined gradients. This decouples the two objective evaluations while keeping a joint update, reducing interference between action learning and ETC supervision throughout co-training.

A.2 Hyperparameters

Stage 1 (Distribution Bridging). All backbones are fine-tuned on ETC supervision using the ms-swift framework with full-parameter fine-tuning (ViT included, `freeze_vit=false`), DeepSpeed ZeRO-3, and BF16 precision. Key hyperparameters are given in Table 3.

Table 3: Stage 1 ETC pretraining hyperparameters (shared across all source datasets).

Hyperparameter	PaliGemma-3B	Qwen2.5-VL-3B
Max steps	20,000	20,000
Learning rate	1×10^{-5}	1×10^{-5}
Warmup ratio	0.05	0.05
LR schedule	cosine	cosine
Per-device batch size	2	2
Gradient accumulation	1	1
Max sequence length	8,192	8,192
Sequence packing	✓	✓
Precision	BF16	BF16

Stage 2 (Objective Bridging). Bridge and LIBERO experiments use the AR-VLA-cotraining framework with a PiFAST backbone; VLABench experiments use the PiFAST framework. Both share the cosine-decay optimizer schedule; differences are listed in Table 4.

Table 4: Stage 2 Objective Bridging hyperparameters.

Hyperparameter	Bridge	LIBERO	VLABench
Architecture	PiFAST	PiFAST	PiFAST
Stage 1 init	ShareRobot	VLA-OS	VLABench VQA
Camera views	1 (primary)	2 (primary + wrist)	3 (primary + wrist + second image)
Max steps	100,000	30,000	100,000
Learning rate	2.5×10^{-5}	2.5×10^{-5}	2.5×10^{-5}
Warmup steps	1,000	1,000	1,000
LR schedule	cosine	cosine	cosine
Global batch size	128	128	32
Optimizer	AdamW (8-bit)	AdamW (8-bit)	AdamW
Action horizon	10	10	10
Action dim	7 (EEF delta)	7 (EEF delta)	7 (EEF delta)
ETC co-train ratio	0.25	0.25	0.25
Image augmentation	rot.$\pm5°$, crop, color jitter	rot.$\pm5°$, crop, color jitter	—
Action chunk encoding	rel. to current	rel. to current	step-wise delta

A.3 Data Mixing Strategy

Stage 1 (Distribution Bridging). For each ETC source, we uniformly sample its sub-categories to maintain a balanced distribution across sub-types.

Stage 2 (Objective Bridging). We use the joint-update co-training schedule described in Section A.1. The ETC batch size is set to 25% of the action batch size (`cotrain_ratio`$= 0.25$), and ETC

sub-categories are sampled uniformly. In the +MM ablations on Bridge and LIBERO, the ETC stream is replaced by generic MM data (A-OKVQA, COCO, and LaTeX-OCR) at the same 25% ratio, with the three MM datasets mixed uniformly.

A.4 Hardware and Compute

Training. Stage 1 ETC pretraining is conducted on 4 NVIDIA H200 141G GPUs. Objective Bridging is conducted on 8 NVIDIA H200 141G GPUs. All training runs use BF16 mixed precision.

Evaluation. All simulation evaluations are conducted on NVIDIA H100 80G GPUs (CUDA 12.8). All models use FP32 precision during inference.

B Additional Backbone Results

The main experiments focus on PaliGemma backbones. Here we report additional backbone ablations using (i) **SigLIP+Gemma**, a PiFAST model initialized from a raw SigLIP-400M vision encoder and Gemma-2B LLM without joint VLM pretraining, and (ii) **Qwen2.5-VL-3B**, evaluated with and without Stage 1 ETC pretraining where runs are available.

Table 5: SigLIP+Gemma backbone ablation under action-only and co-training conditions. Scores are absolute points on a 0–100 scale. LP-train and LB-train denote LIBERO-Plus evaluation for policies trained on LIBERO-Plus and LIBERO, respectively.

Cotrain data	Bridge	LIBERO	LP-train	LB-train
Action only	**38.5**	**91.5**	32.4	40.6
+ MM	2.1	84.6	**39.7**	37.9
+ ETC	17.7	89.4	36.7	38.2
+ MM+ETC	18.8	85.9	—	**43.7**

Table 6: Qwen2.5-VL-3B backbone results on LIBERO. Scores are absolute points on a 0–100 scale.

Init	Cotrain data	LIBERO
Qwen2.5-VL-3B (w/o ETC)	Action only	74.3
	+ MM	73.2
	+ ETC	**78.2**
	+ MM+ETC	70.0
Qwen2.5-VL-3B (w/ ETC)	Action only	66.6
	+ MM	72.0
	+ ETC	**73.2**
	+ MM+ETC	72.6

C Data

C.1 ETC Data Curation Pipeline

Our VQA pipeline grounds questions in real-robot recorded trajectories, inspired by the Robo2VLM paradigm of leveraging synchronized proprioceptive and visual modalities. Ground-truth answers are produced from structured trajectory signals or Gemini-2.5-Pro generation and are manually verified. For most question types, a chain of thought is generated and merged with the answer during training. The pipeline covers the following five reasoning categories:

Trajectory Prediction. Given the initial frame and a language instruction, the model predicts five equidistant waypoints along the recorded end-effector trajectory to abstract the full manipulation motion, evaluating the ability to anticipate complete motion from a single image.

Spatial Understanding. View Correspondence requires the model to match a BBox-highlighted object in one viewpoint to its numeric ID in another, demanding cross-view geometric reasoning

Table 7: VQA construction examples grouped by capability.

Capability	Question Prototype	Ground-truth Source
Trajectory Prediction		
`subtask_trajectory_prediction`	Predict the coordinates of five key points of the robot's gripper trajectory. Output as [(x1,y1), (x2,y2), (x3,y3), (x4,y4), (x5,y5)]. The robot is asked to {instruction}.	Computed from real robot data / generated by Gemini
Spatial Understanding		
`view_correspondence`	The first image includes the bounding box of a specific object. Which bounding box in the second image corresponds to the same object, based on attributes relevant for manipulation? Answer the correct ID.	Generated by Gemini
`relative_direction`	In the image, which direction is the {target object} relative to the robot's end effector?	Computed from real robot data
Affordance		
`affordance_validation`	Given the bounding box, is this region suitable for a stable grasp? Answer yes or no.	Generated by Gemini
`affordance_localization`	Where is the optimal grasp point for {target object}? Provide the pixel coordinates as [x, y].	Computed from real robot data
Object Description		
`target_object_identification`	Which object in the image is the target object? Answer the ID number of its bounding box. The robot is asked to {instruction}.	Generated by Gemini
`target_object_localization`	In the image, locate the {target object} by providing its bounding box. Output the bounding box as [x_min, y_min, x_max, y_max] in image pixel coordinates.	Generated by Gemini
`target_object_description`	Describe the target object of the task. The robot is asked to {instruction}. Focus on attributes relevant for manipulation, such as type, shape, color, position, and orientation. Keep your answer concise.	Generated by Gemini
Task Planning		
`action_understanding`	The robot is asked to {instruction}. Which phase of the grasp action is shown? Choose from [pre grasp, contact, immobilization, detach, post grasp].	Computed from real robot data
`predict_action`	After the action shown in the image, what will be the robot's next action for the task {instruction}?	Generated by Gemini
`subtasks_sequencing`	Please describe the sequence of actions the robot will perform to complete the task: {instruction}. Keep your answer concise.	Generated by Gemini

without depth or calibration data. Relative Direction asks the model to determine the 3D spatial relation between the target object and the end-effector, choosing from left, right, forward, backward, upper, lower. Directions are defined in a gripper-centric frame: left/right along the gripper's local x-axis (perpendicular to the opening plane), forward/backward along the local y-axis (approach direction), and upper/lower along the world z-axis (vertical displacement).

Affordance Reasoning. Affordance Validation presents a highlighted grasp region and asks for a Yes/No judgment on grasp stability. Affordance Localization requires regressing the optimal grasp point for a specified target object.

Goal-Conditioned Reasoning. Target Object Identification selects the numeric ID of the task-specified object from an annotated initial frame. Target Object Localization requires regressing a bounding box for that object. Target Object Description generates a natural-language description of the object's manipulation-relevant (and in some cases, semantic) attributes, testing open-domain visual recognition.

Task Planning. Action Understanding classifies the current manipulation phase from the onset frame and task instruction. Predict Action forecasts the next action from an end-of-phase frame. Sub-task Sequencing generates a complete natural-language action plan from the initial frame and full instruction, evaluating long-horizon procedural reasoning.

Table 7 summarizes the question prototypes used to construct VQA examples for each ETC capability, together with the source of the ground-truth answer.

C.2 Simulation ETC Data Sources and Statistics

For simulation experiments, we construct ETC supervision from three source datasets: ShareRobot (Ji et al., 2025) (Bridge experiments), VLA-OS (Gao et al., 2026) (LIBERO experiments), and VLABench VQA annotations derived via the pipeline described above (VLABench experiments; Track 1 in-domain and Track 2 OOD). Tables 8–10 report the category and sub-type breakdown for each source.

Table 8: ShareRobot ETC data statistics (Bridge experiments). Total: 44,862.

Category	Sub-type	Samples
Affordance	Affordance grounding/validation	6,522
Trajectory	Trajectory prediction	6,870
Planning	Discriminative affordance (+)	3,147
	Discriminative affordance (−)	3,147
	Future prediction	3,147
	Generative affordance	3,147
	Past description	3,147
	Planning remaining steps	3,147
	Planning	3,147
	Planning with context	3,147
	Success prediction (+)	3,147
	Success prediction (−)	3,147
	Planning subtotal	*31,470*
	Total	**44,862**

Table 9: VLA-OS ETC data statistics (LIBERO experiments). Total: 388,808.

Category	Samples
Object BBox grounding	143,668
Gripper localization	83,512
Move-target prediction	78,117
Sub-task decomposition	78,117
Task planning	5,394
Total	**388,808**

C.3 ETC Data Examples

Figure 6 shows representative ETC supervision samples constructed from real-robot trajectories, covering the question types described in Section C.1.

C.4 MM Data for Ablation

For the + `MM` and + `MM+ETC` ablations in Table 1, we use A-OKVQA (Schwenk et al., 2022) (24,903 samples), COCO object detection (Lin et al., 2014) (94,208 samples), and LaTeX-OCR (Blecher, 2022) (192,446 samples) as generic multimodal supervision, totaling 311,557 samples.

D Understanding Evaluation Details

This section describes how the benchmark scores reported in the main paper are computed, and provides per-category breakdowns for all experimental conditions. The evaluation covers two groups of benchmarks: *embodied multimodal benchmarks* (ETC), including ShareRobot, VLA-OS, VLABenchVQA, and Robo2VLM—all constructed from the same ETC data sources described in Section C.2—and *general multimodal benchmarks* (MM), including LaTeX-OCR, COCO, and

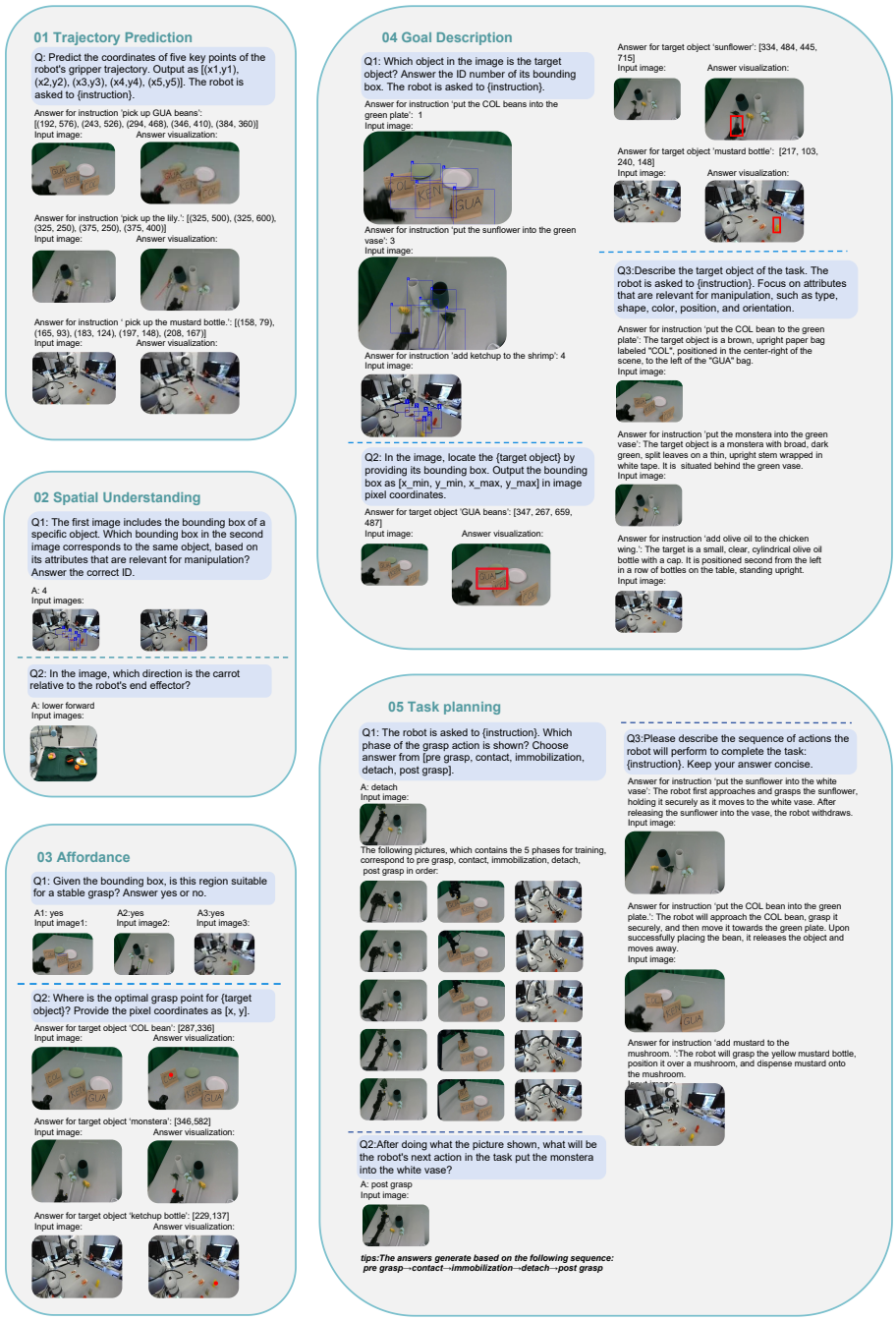


Figure 6: Representative ETC supervision samples constructed from real-robot data. Each row corresponds to a different question type.

Table 10: VLABench ETC data statistics (VLABench experiments). T1 = Track 1 in-domain; T2 = Track 2 OOD.

Category	Sub-type	T1	T2
Affordance	Affordance validation (+)	2,850	1,281
	Affordance validation (−)	2,851	1,281
	Affordance localization	28,433	1,281
	Subtotal	*34,134*	*3,843*
Goal Description	Target object description	1,912	430
	Target object identification	26,577	1,290
	Target object localization	28,749	1,286
	Subtotal	*57,238*	*3,006*
Spatial Understanding	Relative direction	1,913	430
	View correspondence	30,820	930
	Subtotal	*32,733*	*1,360*
Task Planning	Action understanding	46,209	2,094
	Sub-task sequencing	9,583	698
	Subtotal	*55,792*	*2,792*
Trajectory Prediction	Trajectory waypoints	30,674	2,081
	Total	**210,571**	**13,082**

A-OKVQA. Table 12 reports pre-action VLM scores across the four ETC benchmarks; Table 13 details Objective Bridging co-training results per benchmark.

Evaluation protocol. All benchmarks are evaluated in an offline prediction-then-scoring manner. Each test example contains one or more images, a natural-language instruction or question, and a reference answer. The model response is decoded deterministically and saved before metric computation. The scorer assigns each example a continuous score and a binary correctness label; reported accuracy is the number of correct examples per 100 examples. Each dataset is decomposed into task-specific sub-scores so that localization, trajectory prediction, planning, recognition, and open-text generation are reported separately.

Metric convention. Performance results are reported as absolute scores on a 0–100 scale; differences between results are measured in score points. Benchmark averages are unweighted means over the visible sub-scores in each table. We keep the paper tables keyed by semantic experimental conditions rather than internal run IDs: *Init* specifies the VLM initialization, *Training mix* specifies the VL co-training source, and *Action data* specifies the robot dataset used for VLA action training.

D.1 Scoring Rules

Several scoring rules are shared across datasets.

D.1.1 Text Similarity

For open-ended short answers, planning outputs, and natural-language descriptions, we use token-level F1 unless otherwise specified. Prediction and reference are lowercased and tokenized; let the multiset overlap be m. Precision $= m/|\text{pred}|$, recall $= m/|\text{ref}|$, and the text score is their harmonic mean. Empty prediction and empty reference receive score 1; if only one side is empty, the score is 0.

For LaTeX-OCR, we use character-level sequence similarity instead, because correctness is often determined by local character differences (braces, subscripts, symbols).

For spatial-relation questions in VLABenchVQA, we extract a set of relation labels (`left`, `right`, `forward`, `backward`, `upper`, `lower`) from both prediction and reference and compute F1 over the relation sets. In the reported evaluation the correctness threshold is 1.0 (exact set match).

An example is counted as correct when its best text similarity score reaches the dataset-specific threshold (default 0.8 for short answers; 0.35 for long descriptive VLABenchVQA tasks).

D.1.2 NORMALIZED EXACT MATCHING

For yes/no, ID-choice, and VQA-style tasks, answers are normalized (lowercased, punctuation and extra whitespace removed) before comparison. Yes/no answers are mapped to Boolean values; ID-choice answers are parsed by extracting the first integer.

For Robo2VLM, the evaluator first extracts a final answer from a `final answer:` field if present, otherwise uses the last non-empty line. The example is correct if the normalized prediction exactly matches the reference, or if the normalized reference appears as a substring of the prediction.

D.1.3 BOUNDING BOX IOU

For BBox localization, predictions and references are parsed as (x_1, y_1, x_2, y_2). The score is intersection-over-union (IoU); the example is correct if IoU exceeds the task threshold (ShareRobot: 0.5; VLABenchVQA: 0.6; VLA-OS: 0.7).

D.1.4 POINT DISTANCE

For point localization, the error is the Euclidean pixel distance d. The continuous score is $\exp(-d/\tau)$; the example is correct if $d \leq \tau$. VLA-OS gripper-position tasks use $\tau = 20$ px; VLABenchVQA affordance-point tasks use $\tau = 50$ px.

D.1.5 TRAJECTORY RMSE

Predicted and reference trajectories are parsed as sequences of 2D pixel coordinates and resampled to 50 points along arc length if needed. The trajectory error is the root-mean-squared pixel error over aligned points: $\text{RMSE} = \sqrt{\text{mean}_i \|p_i - g_i\|_2^2}$. The continuous score is $\exp(-\text{RMSE}/\tau)$ with $\tau = 20$ px; the example is correct if RMSE ≤ 20 px.

D.1.6 COCO JOINT DETECTION F1

For COCO, we use a dataset-specific metric that jointly requires category-label agreement and BBox overlap. Class names are normalized to the 80 COCO categories (e.g., "bike" $\rightarrow$ "bicycle"). Predictions and ground-truth objects are matched greedily by descending IoU, with each object used at most once. A *class-aware* match requires class equality and IoU ≥ 0.6. Precision, recall, and F1 are computed from global TP/FP/FN across the dataset. A ***BBox-only*** variant ignores class labels and uses IoU alone.

D.2 DATASET-SPECIFIC METRICS

D.2.1 SHAREROBOT

ShareRobot contains three test subsets: affordance, trajectory, and planning. Affordance examples are scored by BBox IoU (threshold 0.5). Trajectory examples are scored by trajectory RMSE (threshold 20 px). Planning examples are divided into yes/no (normalized exact match) and open-text questions (token-F1, threshold 0.8). The overall ShareRobot score reported in our tables is the unweighted mean of the affordance, planning, and trajectory sub-scores.

D.2.2 VLA-OS

VLA-OS contains five subsets: `bbox`, `gripper_position`, `planning`, `subtask`, and `move`. BBox examples use BBox IoU (threshold 0.7). Gripper-position examples use point distance ($\tau = 20$ px). Planning, subtask, and move examples use token-F1 (threshold 0.7).

D.2.3 VLABENCHVQA

VLABenchVQA spans affordance, goal description, spatial understanding, task planning, and trajectory prediction. Point affordance questions use point distance ($\tau = 50$ px); yes/no questions use normalized exact match; text affordance uses token-F1 (0.8). BBox tasks use IoU (0.6); ID-choice tasks use parsed-ID exact match; long goal descriptions use token-F1 (0.35). Spatial-relation examples use relation-set F1 (threshold 1.0). Task-planning examples use token-F1 (0.8); trajectory examples use RMSE (20 px).

D.2.4 ROBO2VLM

Robo2VLM contains VQA and reasoning subsets, both evaluated with final-answer extraction followed by normalized exact matching (or token-F1 as a continuous score).

D.2.5 LATEX-OCR

LaTeX-OCR evaluates formula transcription. The evaluator extracts the final answer (preferring an explicit `<answer>` block, then `final answer:`, then the last non-empty line) and computes character-level sequence similarity against the reference. The example is correct if similarity ≥ 0.8.

D.2.6 COCO

COCO is evaluated as open-set object detection with the joint Detection F1 metric (IoU 0.6) described in Section D.1. The primary score is the global class-aware F1; the stricter per-image exact-match accuracy is reported as a diagnostic.

D.2.7 A-OKVQA

A-OKVQA is evaluated as VQA with multiple acceptable answer forms (direct answers, multiple-choice text, and the correct-option field). The score is the maximum token-F1 over all candidates; the example is correct if the best score ≥ 0.8.

D.3 AGGREGATION AND REPORTING

For all datasets except COCO, each example contributes one binary correctness label to the overall accuracy and one continuous score to the sub-score average. For COCO, the primary score is the global class-aware F1 computed from accumulated TP/FP/FN rather than per-image accuracy.

Table 11 summarizes all thresholds used in the reported evaluation.

Table 11: Evaluation thresholds used in the reported results. Bounded similarity and overlap thresholds are shown on a 0–100 scale; error and distance thresholds retain pixel units.

Metric	Threshold (points or px)
ShareRobot BBox IoU	50
VLABenchVQA BBox IoU	60
VLA-OS BBox IoU	70
COCO detection IoU	60
Trajectory RMSE	20 px
VLA-OS gripper point dist.	20 px
VLABenchVQA affordance point	50 px
Short text token-F1	80
VLA-OS text token-F1	70
VLABenchVQA long text token-F1	35
LaTeX-OCR char. similarity	80
A-OKVQA token-F1	80
Spatial relation F1	100

D.4 Benchmark Score Tables

Pre-action VLM scores. Table 12 compares backbone VLMs of different initializations on each benchmark before any action training.

Table 12: Pre-action VLM absolute scores on a 0–100 scale across the four ETC benchmarks before any action training.

ShareRobot (Bridge-side)

Init	Avg.	Plan
PaliGemma-raw	16.33	16.33
PaliGemma-Bridge	26.65	26.65

VLA-OS (LIBERO-side)

Init	Avg.	Move	Plan	Subtask	BBox	Grip
PaliGemma-raw	12.21	17.74	22.82	20.46	0.00	0.00
PaliGemma-LIBERO	41.21	25.02	91.44	47.18	32.42	9.96

VLABenchVQA

Init	Avg.	Aff.	Traj.	Plan	Goal	Spatial
PaliGemma-raw	9.55	24.67	0.00	4.92	7.13	11.04
PaliGemma-VLABench	56.38	47.63	12.94	84.09	47.54	89.69

Robo2VLM-VQA

Init	Avg.	VQA
PaliGemma-raw	15.15	15.15
PaliGemma-DROID	69.29	69.29

Objective Bridging co-training results. ShareRobot (Bridge-side) covers affordance reasoning, high-level planning, and trajectory reasoning. VLA-OS (LIBERO-side) decomposes scene understanding into move-target prediction, task planning, subtask decomposition, object BBox grounding, and gripper localization. The general MM benchmark measures open-domain VL skill retention; A-OKVQA is evaluated as multiple-choice VQA, COCO as class-aware object detection, and LaTeX-OCR as expression transcription. For COCO, the main score is the dataset-level class-aware F1 at IoU 0.6; the diagnostics table additionally reports box-only F1@0.6 (localization only), mIoU over matched boxes, and image-level exact-match accuracy. Tables 13–13 report all sub-scores.

E Real-World Experiments

E.1 Task Suite and Deployment Splits

We evaluate real-world adaptation on two WidowX tasks following the Bridge dataset. Each task has an in-domain split used for task-specific Retentive Adaptation and an OOD split that changes task-relevant visual-language factors without adding OOD action demonstrations. All real-world experiments use the same training hyperparameters as the simulation experiments, except that we set the maximum number of training steps to 30,000.

WidowX Tasks. The coffee-bag task requires selecting a handwritten-labeled coffee-bean bag and placing it onto a specified plate, e.g., *"Put the ETH coffee bag on the blue plate."* During task-specific Retentive Adaptation, the model only sees a fixed set of bag labels and plate colors; the OOD split changes both factors, requiring an unseen label–plate color combination. The flower task requires inserting an instructed flower into a green or white vase, e.g., *"Put the monstera in the green vase."* Its Retentive Adaptation split contains seen flower–vase pairs, while the OOD split replaces a seen flower category with an unseen one. Thus, the OOD factors in both tasks are held out from action demonstrations during Retentive Adaptation.

E.2 Real-World Data, Supervision, and Evaluation Protocol

Figure 7 summarizes the real-world ETC data sources used in our Retentive Adaptation ablations. For coffee-bag placement, we separately train with human ETC, computed from real-robot trajectories, and Gemini ETC, generated from the first trajectory frame, to test whether automatically generated annotations can match human trajectory-derived supervision. For flower insertion, we additionally compare real OOD data with GPT-Image-2 synthesized OOD images, evaluating whether generated visual variations can support scalable OOD ETC construction.

Table 14 summarizes the ETC data used in our real-world experiments. We distinguish different ETC sources and train them separately in the corresponding ablations, with the resulting policy

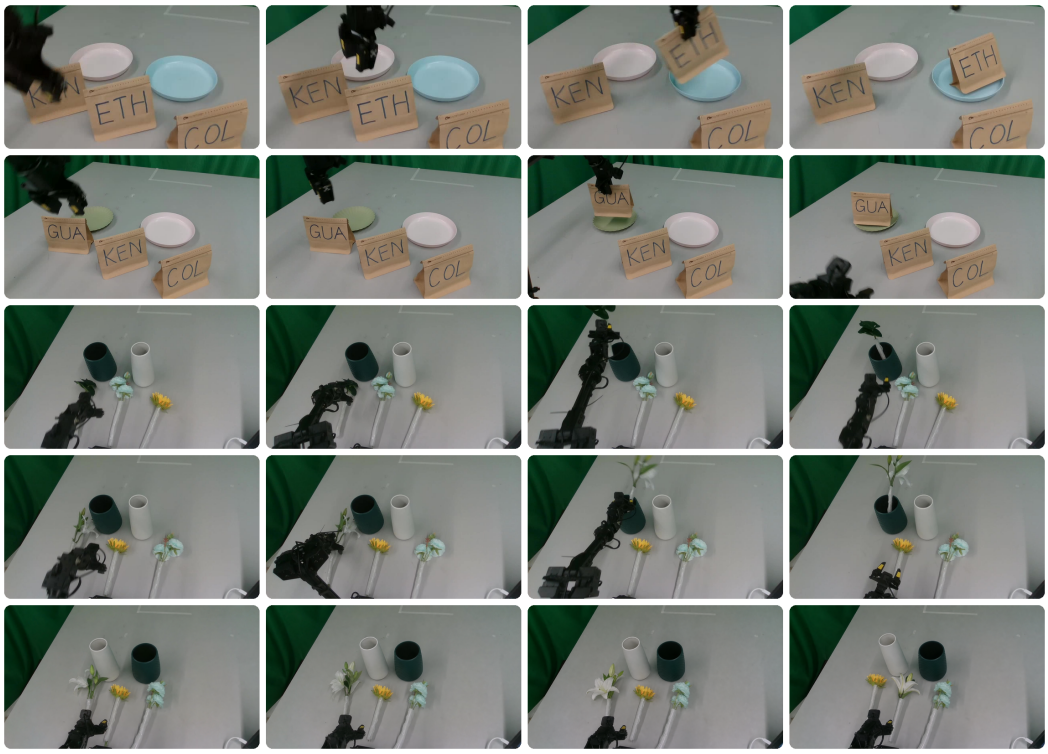


Figure 7: Real-world task and data examples used in the WidowX experiments. Rows show, from top to bottom: coffee-bag in-domain rollouts, coffee-bag OOD rollouts, flower in-domain rollouts, flower OOD rollouts, and GPT-Image-2 synthesized flower OOD examples.

Table 13: Objective Bridging co-training benchmark details. All results are absolute scores on a 0–100 scale.

ShareRobot (Bridge; all variants use Bridge as action data)

Init	Training mix	Avg.	Aff.	Plan	Traj.
PaliGemma-ETC	Action only	0.00	0.00	0.00	0.00
PaliGemma-ETC	+ MM+ETC	38.04	5.86	15.38	92.89
PaliGemma-ETC	+ ETC	91.92	93.72	88.73	93.31
PaliGemma	Action only	0.00	0.00	0.00	0.00
PaliGemma	+ ETC	37.57	2.93	13.54	96.23
PaliGemma	+ MM+ETC	69.29	94.56	20.00	93.31

VLA-OS (LIBERO; all variants use LIBERO as action data)

Init	Training mix	Avg.	Move	Plan	Subtask	BBox	Grip
PaliGemma-ETC	Action only	0.05	0.18	0.00	0.08	0.00	0.00
PaliGemma-ETC	+ MM+ETC	72.46	43.76	100.00	54.38	75.28	88.88
PaliGemma-ETC	+ ETC	60.00	41.66	100.00	52.78	32.61	72.97
PaliGemma	Action only	0.05	0.19	0.00	0.08	0.00	0.00
PaliGemma	+ ETC	58.16	40.48	100.00	51.02	27.69	71.59
PaliGemma	+ MM+ETC	72.30	41.87	100.00	53.65	76.72	89.28

General MM benchmark

Init	Training mix	Action data	Avg.	A-OKVQA	COCO	LaTeX-OCR
PaliGemma-ETC	Action only	Bridge	0.89	0.00	0.00	2.66
PaliGemma-ETC	+ MM	Bridge	43.76	51.56	13.70	66.03
PaliGemma-ETC	+ MM+ETC	Bridge	1.09	0.00	0.00	3.27
PaliGemma	Action only	Bridge	23.65	0.03	9.02	61.91
PaliGemma	+ MM	Bridge	31.11	32.49	12.25	48.58
PaliGemma	+ MM+ETC	Bridge	19.20	2.79	8.56	46.25
PaliGemma-ETC	Action only	LIBERO	0.00	0.00	0.00	0.00
PaliGemma-ETC	+ MM	LIBERO	22.97	14.76	9.39	44.77
PaliGemma-ETC	+ MM+ETC	LIBERO	17.62	0.00	8.37	44.48
PaliGemma	Action only	LIBERO	0.00	0.00	0.00	0.00
PaliGemma	+ MM	LIBERO	23.28	14.67	10.25	44.93
PaliGemma	+ MM+ETC	LIBERO	17.54	0.00	9.02	43.60

COCO detection diagnostics (Cls-F1@0.6 = class-aware F1; Box-F1@0.6 = localization only; mIoU = mean IoU over matched boxes; Exact = image-level exact-match accuracy)

Init	Training mix	Action data	Cls-F1@0.6	Box-F1@0.6	mIoU	Exact
PaliGemma-ETC	Action only	Bridge	0.00	0.05	0.00	0.00
PaliGemma-ETC	+ MM	Bridge	13.70	14.23	77.46	8.24
PaliGemma-ETC	+ MM+ETC	Bridge	0.00	0.04	0.00	0.00
PaliGemma	Action only	Bridge	9.02	10.06	76.24	5.09
PaliGemma	+ MM	Bridge	12.25	12.92	75.89	6.36
PaliGemma	+ MM+ETC	Bridge	8.56	9.47	75.49	4.81
PaliGemma-ETC	Action only	LIBERO	0.00	0.06	0.00	0.00
PaliGemma-ETC	+ MM	LIBERO	9.39	10.28	75.43	5.13
PaliGemma-ETC	+ MM+ETC	LIBERO	8.37	9.30	75.23	4.60
PaliGemma	Action only	LIBERO	0.00	0.04	0.00	0.00
PaliGemma	+ MM	LIBERO	10.25	11.29	75.73	5.47
PaliGemma	+ MM+ETC	LIBERO	9.02	10.02	75.28	4.34

performance reported in Table 2. Here, *human* denotes ETC supervision directly computed from real-robot trajectories, while *Gemini* denotes annotations generated by prompting Gemini with the first frame of each real-robot trajectory. For the flower task, we additionally include GPT-Image-2-generated images of OOD scenes to evaluate whether synthetic visual variations can provide useful OOD ETC supervision.

Table 14: Real-world ETC data statistics.

Task	ETC source	Records	Images
Coffee bag	ID ETC (Human)	2880	2160
	ID ETC (Gemini)	3960	2520
	OOD ETC (Human)	1120	840
	OOD ETC (Gemini)	1540	980
Flower	ID ETC (Gemini)	1635	872
	OOD ETC (Human)	540	288
	OOD ETC (GPT-Image-2)	470	188

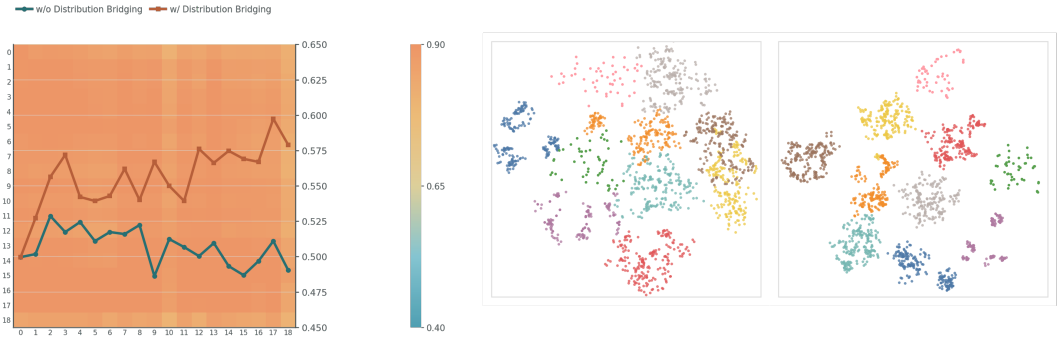


Figure 8: **Distribution Bridging provides an embodied initialization while preserving the pretrained VLM structure.** Left: CKA similarity between the original VLM and the ETC-pretrained VLM, together with affordance prediction probing on VLABench. Right: t-SNE visualizations of VLABench task-conditioned representations before and after Distribution Bridging, where colors denote different task categories.

We use the following task-specific scoring method for WidowX evaluation. Scores are averaged over 18 in-domain and 12 OOD rollouts for each task and reported as absolute points on a 0–100 scale. Coffee-bag placement receives 100 points if the correct bag is placed on the correct plate, 25 points if the correct bag and plate are selected but the bag is dropped or not reliably grasped, and 0 otherwise. Flower insertion receives 100 points for successful insertion, 75 if the correct flower is grasped but dropped before completion, 50 if the correct flower and vase are selected without insertion, 25 if the correct flower is selected but the wrong vase is targeted, and 0 otherwise.

F How ETC Reshapes Embodied Representations

F.1 Distribution Bridging Shapes VLM Representations

Figure 8 examines how Distribution Bridging changes the VLM before any action training. The CKA analysis shows that ETC pretraining preserves substantial similarity to the original VLM representation, indicating that embodied adaptation does not overwrite the pretrained visual-language structure. At the same time, layer-wise affordance prediction probing improves after Distribution Bridging, showing that the backbone encodes more manipulation-relevant affordance information.

The t-SNE visualization provides a complementary view of this change. Without Distribution Bridging, representations from different VLABench tasks are more entangled. After Distribution Bridging, task-conditioned embodied inputs form more separable clusters, suggesting that ETC pretraining injects task-discriminative embodied structure into the VLM representation and provides a stronger initialization for downstream policy learning.

F.2 Preserving Embodied Alignment during Objective Bridging

Figure 9 studies whether the embodied alignment acquired from Distribution Bridging is preserved during Objective Bridging. Under action-only Objective Bridging, the policy backbone drifts away

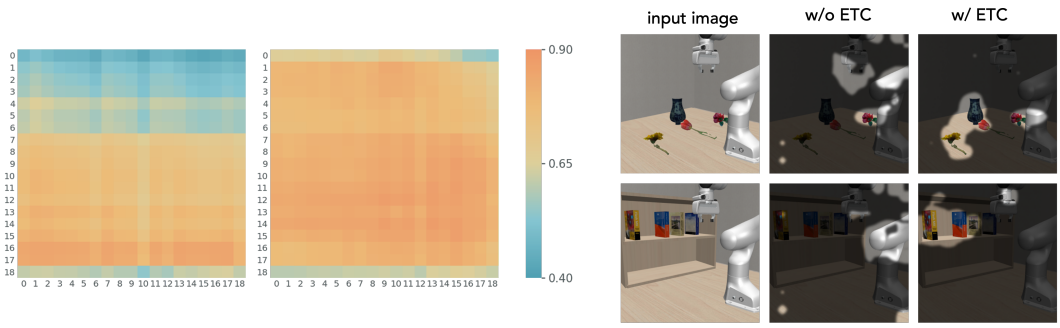


Figure 9: **Objective Bridging preserves the embodied alignment acquired during Distribution Bridging.** Left: CKA similarity between the ETC-pretrained representation and VLA backbones after Objective Bridging with action-only supervision or action–ETC co-training. Right: attention visualizations on representative VLABench scenes, comparing the input image, action-only Objective Bridging, and ETC co-training.

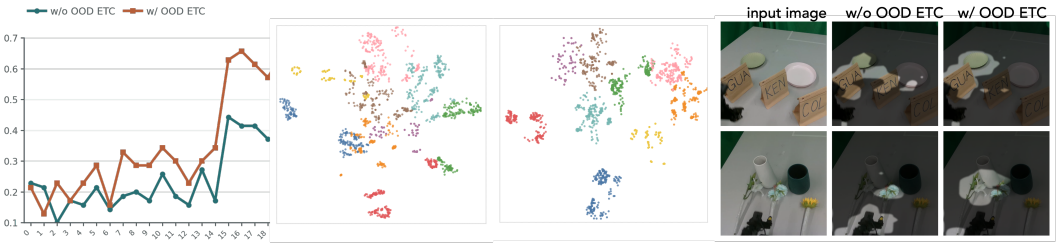


Figure 10: **OOD ETC induces representations that better support compositional generalization.** Left: layer-wise task-planning probing on VLABench Track 2 with and without OOD ETC induction. Middle: t-SNE visualizations of OOD VLABench Track 2 representations, comparing models trained without and with OOD ETC. Right: attention visualizations on real-world OOD scenes, comparing the input image, the model without OOD ETC, and the model with OOD ETC.

from the ETC-pretrained representation, as reflected by lower CKA similarity. In contrast, Objective Bridging with ETC co-training keeps the representation substantially closer to the ETC-pretrained backbone. This suggests that ETC supervision regularizes policy learning and mitigates representational forgetting during action adaptation.

The attention visualizations provide a qualitative view of this preservation effect. Without ETC co-training, the model often attends diffusely or focuses on regions weakly related to the instruction. With ETC co-training, attention is better concentrated on task-relevant objects and interaction regions. Thus, Objective Bridging preserves not only feature-space similarity to the embodied initialization, but also the visual-language grounding needed for downstream policy execution.

F.3 OOD ETC Induces Compositional Generalization

Figure 10 examines whether OOD ETC can reshape the policy representation toward unseen deployment conditions. We focus on OOD inputs, where the task-relevant object compositions are absent from action demonstrations. Compared with the model trained without OOD ETC, OOD ETC induction improves layer-wise task-planning probing on OOD embodied inputs, indicating that the backbone better recognizes how unseen objects or compositions should participate in the task. The t-SNE visualization further shows more structured clusters after OOD ETC induction, suggesting that these unseen compositions become more separable in the learned representation.

The attention visualizations make this effect concrete in real-world OOD scenes. Without OOD ETC, attention is often scattered or biased toward irrelevant regions under unseen labels or object categories. With OOD ETC, attention shifts toward the task-relevant objects and interaction regions. These

results suggest that OOD ETC provides a lightweight mechanism for binding unseen visual-language factors to existing manipulation skills for compositional generalization.