BRIDGING HUMAN AND ROBOT EXPERIENCES VIA A SHARED HIERARCHICAL SEMANTIC LATENT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale egocentric human videos provide rich and scalable manipulation experience for robot policy learning, yet substantial embodiment gaps in visual appearance, kinematic structure, and action spaces hinder effective knowledge transfer. Existing methods often decouple cross-embodiment knowledge alignment from robot policy learning, making it difficult for human manipulation knowledge to directly and fully contribute to robot action learning. To address this issue, we propose Nexus-VLA, a hybrid policy learning framework that enables cross-embodiment knowledge transfer through end-to-end human-robot joint training, allowing human manipulation experience to directly contribute to robot policy learning. We first introduce a hierarchical cross-embodiment semantic alignment paradigm that progressively establishes shared manipulation semantics from task intent to fine-grained interactions, enabling more complete use of the manipulation knowledge contained in human videos. Building on this hierarchy, we design a shared manipulation latent representation that jointly encodes multi-level shared knowledge into an internal representation directly used for continuous action generation, improving policy learning efficiency while effectively leveraging human knowledge. Across RoboCasa-GR1, OOD settings, and real-world robot experiments, Nexus-VLA achieves an 80.5% task success rate on RoboCasa-GR1 and demonstrates strong advantages in human-data utilization, learning efficiency, and cross-scenario generalization.

1 INTRODUCTION

Because collecting robot demonstrations at scale remains costly and challenging (Khazatsky et al., 2024), massive egocentric human videos, which capture rich real-world manipulation and interaction knowledge, provide a scalable source of data for robot policy learning (Grauman et al., 2022). However, substantial differences between humans and robots in visual appearance, kinematic structure, and action spaces (Chen et al., 2025a) make it challenging to extract manipulation representations that are semantically consistent across embodiments, thereby hindering the effective transfer of human manipulation knowledge to robot policies. Despite these differences, humans and robots share similar manipulation semantics and interaction structures when performing the same task. Therefore, a central challenge in learning robot policies from human videos is to identify such shared representations and transform them into a unified representation that effectively supports robot action generation.

To better exploit human data, recent studies have begun to use embodiment-invariant manipulation semantics in human demonstrations as bridges between humans and robots (Bharadhwaj et al., 2023). However, existing approaches often establish human-robot correspondences through only one aspect of semantic consistency, such as aligning high-level task semantics (Kim et al., 2026) or matching human hand motion with robot end effectors (Chen et al., 2026), leaving shared cross-embodiment knowledge only partially modeled. Manipulation semantics are not a single flat signal, but form a hierarchy that progressively connects abstract intent to concrete physical interaction. Manipulation intent describes what to do, spatial semantics specify what to interact with and where, motion semantics characterize how to perform the operation, and interaction semantics further capture how the action affects the object. We argue that these coarse-to-fine semantic levels jointly define the shared structure of human and robot manipulation, suggesting that cross-embodiment correspondence should be progressively established across different semantic levels.

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.

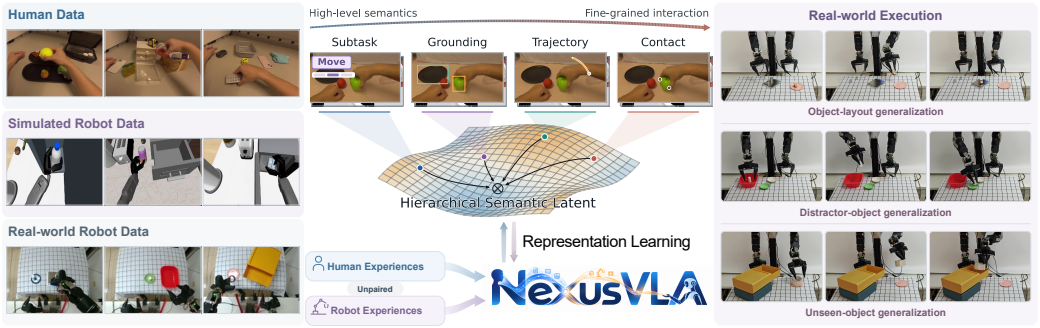

Figure 1: Nexus-VLA bridges human and robot experiences through a Hierarchical Semantic Latent, enabling cross-embodiment knowledge transfer and real-world generalization.

However, explicitly modeling manipulation semantics separately at different levels is insufficient to capture their underlying structure. These semantics are not independent; rather, they describe the same manipulation process at different levels of abstraction and form a progressively refined, hierarchically constrained structure. Modeling each level in isolation makes it difficult to preserve cross-level dependencies and consistency, motivating a unified representation of multi-level manipulation semantics. However, directly using explicit semantics as intermediate policy representations introduces additional computation and decision latency during inference (Bai et al., 2026), while their structured or discrete form is also ill-suited for directly driving continuous action generation (Wei et al., 2026). The key challenge is therefore to transform hierarchical manipulation knowledge shared across embodiments into an internal representation suitable for action generation, allowing knowledge from human demonstrations to directly participate in robot policy decision-making. In this way, multi-level knowledge from human demonstrations can be more completely transferred to robot policies, enabling more effective use of human data.

As illustrated in Figure 1, we propose Nexus-VLA, an efficient mixed policy-learning framework for jointly learning from human and robot demonstrations. Nexus-VLA encodes hierarchical manipulation knowledge shared across embodiments into a Hierarchical Semantic Latent that directly participates in action prediction, allowing shared knowledge from human demonstrations to effectively contribute to robot policy learning while improving cross-scenario generalization. Specifically, we first introduce a hierarchical cross-embodiment semantic alignment paradigm. With structured manipulation experts, we generate consistent supervision for human and robot demonstrations across manipulation intent, spatial semantics, motion semantics, and interaction semantics, progressively establishing coarse-to-fine semantic correspondences between the two embodiments. Building on this hierarchy, we aggregate these interdependent shared semantics through joint multi-level prediction objectives into a Semantic Latent, which is combined with vision-language features for action prediction. The robot action learning objective further shapes this latent representation through end-to-end optimization, encouraging it to preserve cross-embodiment semantic consistency while remaining directly useful for action generation. Nexus-VLA jointly trains on mixed human and robot demonstrations in a single stage and, at inference time, directly predicts actions from the Semantic Latent and vision-language features without explicitly generating intermediate semantics.

We systematically evaluate Nexus-VLA on RoboCasa-GR1, multiple out-of-distribution (OOD) scenarios, and real-robot platforms. Nexus-VLA achieves an 80.5% task success rate on RoboCasa-GR1 and demonstrates strong generalization across OOD tasks and real-robot settings. These results show that connecting human experience with robot action learning through hierarchical semantic latent representations enables more effective use of large-scale human demonstrations and improves both robot policy learning and cross-scenario generalization.

In summary, the main contributions of this paper are as follows:

1. **An Efficient Single-Stage Co-Training Framework:** We overcome the limitations of traditional methods that rely on paired cross-embodiment data or multi-stage pre-training by realizing end-to-

end joint optimization over unaligned human videos and robot demonstrations, which significantly improves both human data utilization and policy learning efficiency.

2. **A Unified Hierarchical Semantic Latent Representation:** We are the first to encode multi-level manipulation semantics from abstract intents to fine-grained physical interactions into a compact latent variable. This representation retains rich operational structures while bypassing the need for explicit intermediate semantic inference, directly interfacing with continuous action generation.

3. **Extensive Evaluation Across Simulated and Real-World Environments:** Nexus-VLA achieves an 80.5% task success rate on the RoboCasa-GR1 benchmark and demonstrates outstanding cross-scene generalization and data efficiency across various out-of-distribution (OOD) scenarios and real-robot platforms.

2 Related Work

Learning Robot Policies from Human Videos. Leveraging rich manipulation experience from large-scale egocentric human videos to learn robot policies has been widely explored recently. One line of work first pre-trains visual, kinematic, or action-related representations on human videos, and subsequently adapts or fine-tunes the policy using robot demonstrations (Yang et al., 2025b; Ma et al., 2026). Further efforts introduce explicit cross-embodiment alignment stages to facilitate knowledge transfer from humans to robots (Zheng et al., 2026; Yu et al., 2026). However, such multi-stage pipelines allow human knowledge to act on final policy learning only indirectly. Consequently, recent studies have begun to explore direct co-training on mixed human and robot demonstrations (Kareer et al., 2025), yet straightforward data mixing still struggles to establish consistent cross-embodiment manipulation supervision.

Cross-Embodiment Manipulation Alignment. To bridge embodiment discrepancies, constructing cross-embodiment stable manipulation representations has emerged as a core strategy for connecting human experience with robot control. Existing work typically establishes human-robot correspondences at different levels of manipulation semantics, ranging from high-level task intents (Kim et al., 2026; Lee et al., 2026) and mid-level spatial affordances or geometric grounding (Yang et al., 2025a; Xu et al., 2025), to low-level end-effector trajectories and physical interactions (Chen et al., 2026). Other methods narrow the embodiment gap by leveraging paired videos or unified pose spaces (Xie et al., 2026; Kareer et al., 2025). However, these approaches generally focus on single-level or local semantic alignment, isolating the inter-level dependencies across semantic granularities and thus limiting the depth and completeness of cross-embodiment knowledge transfer.

Latent Manipulation Representations for Robot Control. Beyond extracting shared semantics, transforming these semantics into internal representations suitable for robot control represents another core challenge. One category of methods explicitly expresses the semantics as structured intermediate information (e.g., affordance maps or predicted trajectories) (Li et al., 2025; Xu et al., 2025; Yang et al., 2025a). While interpretable, such methods require generating explicit intermediate representations during inference, which introduces computational overhead and decision latency; structured signals cannot directly serve as an effective interface for continuous action generation. Another line of work encodes manipulation information into compact latent representations (Ye et al., 2025; Chen et al., 2025b; Bu et al., 2025). Although well-suited for end-to-end control, these latent spaces lack explicit multi-level semantic supervision, making it difficult to form structured semantic constraints and weakening their capability to support precise action generation.

3 Method

As illustrated in Figure 2, Nexus-VLA performs end-to-end mixed training over unpaired human and robot demonstrations within a unified policy learning framework, enabling human manipulation knowledge to directly contribute to robot action learning. A shared vision-language backbone encodes both types of demonstrations and aggregates four types of cross-embodiment manipulation semantics (Intent, Spatial, Motion, and Interaction) spanning high-level task understanding to low-level physical interaction into a Hierarchical Semantic Latent, forming a unified internal representation for manipulation. We further organize human and robot actions into a unified action

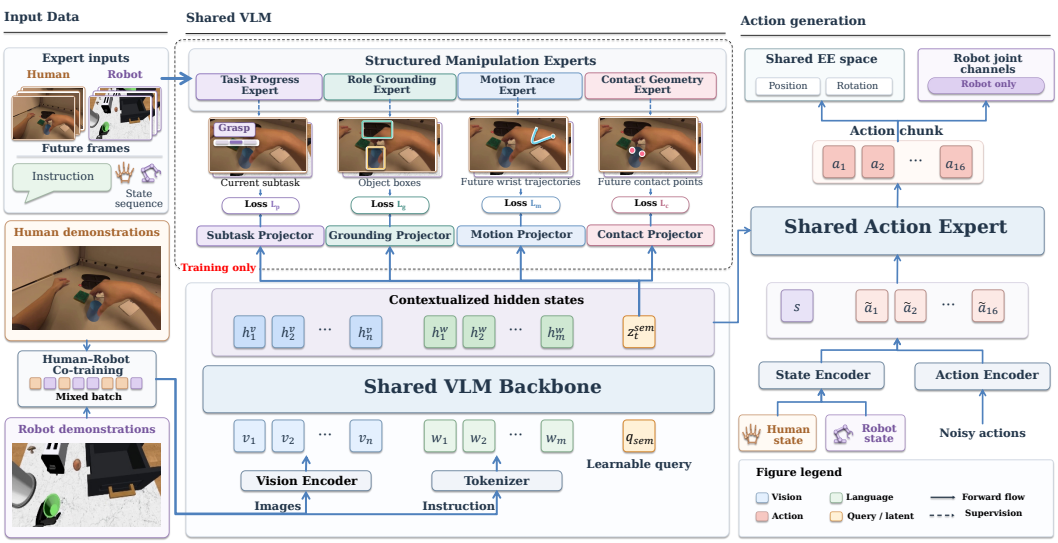

Figure 2: Overview of the Nexus-VLA framework. Hierarchical semantic latents enable joint human–robot learning and action generation.

representation space, allowing semantic and action supervision to jointly shape the Semantic Latent, which, together with vision-language features, conditions continuous robot action generation.

3.1 Hierarchically Shaped Shared Manipulation Latent Representation

To transform cross-embodiment manipulation knowledge into an internal representation that can directly support action generation, Nexus-VLA constructs a Hierarchical Semantic Latent and shapes it with multi-level manipulation semantics ranging from high-level task understanding to concrete physical interaction.

3.1.1 Hierarchical Shared Manipulation Semantics

Despite differences in visual appearance, kinematic structure, and execution form, humans and robots share similar task goals, object spatial relations, motion processes, and physical interactions when performing the same manipulation. We organize such cross-embodiment shared information into four semantic levels (Intent, Spatial, Motion, and Interaction) according to their level of abstraction, progressively grounding manipulation semantics from high-level task understanding to concrete actions and physical interactions.

Specifically, Intent represents the local task goal, i.e., "what to do"; Spatial represents the spatial locations of the manipulated object and target region, i.e., "what to manipulate and where"; Motion represents the motion trend of the acting embodiment toward subsequent interaction, i.e., "how to move"; and Interaction represents effective interaction locations and contact states, i.e., "how to interact." These four types of semantics can be defined consistently for both human and robot demonstrations, providing shared semantic supervision for cross-embodiment representation learning and subsequently shaping the Hierarchical Semantic Latent.

3.1.2 Hierarchical Semantic Latent Construction

To provide a unified internal representation for manipulation semantics at different levels, we append a learnable semantic token $\mathbf{q}_{\text{sem}} \in \mathbb{R}^d$ to the end of the vision-language token sequence and define its final hidden state as the Hierarchical Semantic Latent.

Given the visual observation $\mathbf{o}_t$ at time t and the task instruction ℓ, let $\mathbf{V}_t$ and $\mathbf{T}_t$ denote the corresponding visual and language tokens, respectively. We feed the semantic token together with both

inputs into the shared vision-language backbone F_θ, yielding

$$\mathbf{H}_t = \left[\widetilde{\mathbf{V}}_t; \widetilde{\mathbf{T}}_t; \mathbf{z}_t^{\text{sem}}\right] = F_\theta\left([\mathbf{V}_t; \mathbf{T}_t; \mathbf{q}_{\text{sem}}]\right), \tag{1}$$

where $\widetilde{\mathbf{V}}_t$ and $\widetilde{\mathbf{T}}_t$ denote the contextually encoded visual and language features, respectively, and $\mathbf{H}_t$ denotes the complete multimodal hidden sequence. The final hidden state corresponding to the semantic token, $\mathbf{z}_t^{\text{sem}} \in \mathbb{R}^d$, is defined as the Semantic Latent.

Because $\mathbf{q}_{\text{sem}}$ is placed at the end of the multimodal sequence, its hidden state attends to the preceding visual and language context through causal self-attention, forming a compact internal representation of the current manipulation scene and task instruction. Subsequent multi-level semantic prediction objectives jointly shape $\mathbf{z}_t^{\text{sem}}$, enabling it to further encode cross-embodiment manipulation semantics ranging from high-level task understanding to concrete physical interaction.

3.1.3 Multi-level Semantic Shaping

To enable the Semantic Latent to capture manipulation knowledge at different levels of abstraction, we jointly shape it using four semantic targets: Intent, Spatial, Motion, and Interaction. Let $\mathcal{K} = \{\text{intent}, \text{spatial}, \text{motion}, \text{interaction}\}$ denote the set of semantic levels. For each semantic level $k \in \mathcal{K}$, we predict the corresponding semantic target from the same Semantic Latent $\mathbf{z}_t^{\text{sem}}$:

$$\hat{\mathbf{y}}_t^k = g_k(\mathbf{z}_t^{\text{sem}}), \qquad k \in \mathcal{K}, \tag{2}$$

where g_k denotes the projector for the k-th semantic level. To ensure that the multi-level supervision primarily shapes the shared latent representation rather than relying on additional modeling capacity from task-specific predictors, we keep each projector lightweight, using it only to map the Semantic Latent to the corresponding supervision space.

Corresponding to the four shared semantics defined in Section 3.1.1, the projectors respectively output a local task description, the 2D locations of the manipulated object and target region, future H-step bimanual wrist trajectories, and future contact locations and states. All four supervision targets are automatically constructed from demonstrations by offline Structured Manipulation Experts. These experts are used only to generate training labels and are not involved in policy inference; implementation details are provided in the appendix.

For each semantic level k, let $\bar{\ell}_t^k$ denote the prediction loss computed and normalized over valid target elements when necessary. The overall multi-level semantic shaping objective is

$$\mathcal{L}_{\text{sem}} = \sum_{k \in \mathcal{K}} \lambda_k \, \bar{\ell}_t^k, \tag{3}$$

where λ_k controls the relative weight of each semantic level. Since all four supervision signals jointly shape the same $\mathbf{z}_t^{\text{sem}}$, the resulting latent captures manipulation semantics across different levels of abstraction. At inference time, these semantic targets need not be explicitly generated; the Semantic Latent is directly used for action generation.

3.1.4 Semantics-Enhanced Action Generation

After multi-level semantic shaping, the action generation module uses action tokens to cross-attend to the multimodal hidden sequence $\mathbf{H}_t$, which contains the Semantic Latent, thereby jointly leveraging visual, language, and multi-level manipulation semantics. Given the noisy action sequence $\mathbf{A}_t^\tau$, flow timestep τ, and robot proprioceptive state $\mathbf{s}_t$, the action generation module G_ϕ predicts the conditional velocity field:

$$\hat{\mathbf{v}}_t^\tau = G_\phi\left(\mathbf{A}_t^\tau, \tau, \mathbf{s}_t\,; \mathbf{H}_t\right). \tag{4}$$

Thus, Intent, Spatial, Motion, and Interaction need not be explicitly generated as intermediate semantics during inference. The complete continuous action formulation and Flow Matching training objective are presented in Section 3.2.

3.2 End-to-End Human-Robot Mixed Training

To enable human demonstrations to directly contribute to robot action learning, Nexus-VLA organizes human and robot demonstrations into a unified training action representation and jointly trains them end-to-end within the same policy optimization process.

3.2.1 DUAL-SPACE ACTION REPRESENTATION

Although human demonstrations do not provide robot joint control, human wrists and robot end effectors share a common description of motion relative to the environment. We therefore decompose the training action representation into a cross-embodiment shared wrist/end-effector motion space and a robot-specific joint control space. At time t, the unified action representation is defined as

$$\mathbf{a}_t = [\mathbf{a}_t^{\text{joint}}; \mathbf{a}_t^{\text{wrist}}] \in \mathbb{R}^{d_j + d_w}, \tag{5}$$

where $\mathbf{a}_t^{\text{joint}} \in \mathbb{R}^{d_j}$ denotes the robot joint action, and $\mathbf{a}_t^{\text{wrist}} \in \mathbb{R}^{d_w}$ denotes the bimanual wrist or dual-end-effector motion represented under a unified coordinate system and action parameterization.

For embodiment type $e \in \{\text{human}, \text{robot}\}$, we define the corresponding action validity masks as

$$\mathbf{M}^{(\text{human})} = [\mathbf{0}_{d_j}; \mathbf{1}_{d_w}], \qquad \mathbf{M}^{(\text{robot})} = [\mathbf{1}_{d_j}; \mathbf{1}_{d_w}]. \tag{6}$$

Accordingly, human demonstrations provide action supervision through the shared wrist/end-effector space, while robot demonstrations additionally provide joint-control supervision, allowing both data sources to jointly train the same action generation model.

During mixed training, we explicitly control the overall sampling ratio between human and robot demonstrations, while sampling datasets within each embodiment source proportionally to their sizes. This decouples the human-robot training ratio from the relative sizes of the underlying datasets.

3.2.2 JOINT OPTIMIZATION OBJECTIVE

Nexus-VLA jointly optimizes multi-level semantic shaping and continuous action generation, allowing the Semantic Latent to encode cross-embodiment manipulation semantics while directly supporting robot control. For continuous action generation, we adopt conditional flow matching. Following the standard formulation, we interpolate a ground-truth action chunk $\mathbf{A}_t$ with Gaussian noise to obtain $\mathbf{A}_t^\tau$ at flow timestep τ, with target velocity $\mathbf{v}_t^\star$. Conditioned on the embodiment state $\mathbf{s}_t^{(e)}$ and multimodal context $\mathbf{H}_t$, the action generator predicts the corresponding velocity field $\hat{\mathbf{v}}_t^\tau$.

To accommodate the different action supervision available for human and robot demonstrations, we extend the embodiment-specific action mask in Equation equation 6 over the action chunk to obtain $\mathbf{\Pi}^{(e)}$ and compute the flow matching loss only over valid action dimensions:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\frac{\left\|\mathbf{\Pi}^{(e)} \odot (\hat{\mathbf{v}}_t^\tau - \mathbf{v}_t^\star)\right\|_2^2}{\left\|\mathbf{\Pi}^{(e)}\right\|_1 + \epsilon}\right]. \tag{7}$$

Finally, the action learning objective is jointly optimized with the semantic shaping objective introduced in Section 3.1.3:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}}. \tag{8}$$

Semantic and action supervision are thus jointly optimized under a single end-to-end objective. At inference time, the model generates action chunks through the learned velocity field and executes the robot joint-action component.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We use RoboCasa-GR1 (Bjorck et al., 2025) as our primary simulation benchmark. Built on RoboCasa (Nasiriany et al., 2024), the benchmark contains 24 humanoid manipulation tasks using the Fourier GR-1 robot, covering object transport and interactions with articulated objects such as drawers, cabinet doors, and microwaves. In addition to the standard test scenarios, we construct six controlled OOD settings to evaluate generalization under illumination changes, visual noise, spatial configuration changes, novel object-receptacle compositions, objects observed in human demonstrations, and fully unseen objects.

Table 1: Comparison on the RoboCasa-GR1 benchmark. We report success rates (%) averaged over five task groups and all 24 tasks.

Type	Method	Close	Cuttingboard	Placemat	Tray	Plate	Avg.
Foundation VLA	GR00T-N1.6	24.2	56.9	51.9	55.1	57.6	47.6
	ABot-M0	61.7	61.6	51.0	56.8	58.0	58.3
Semantic Enhanced	VP-VLA	54.3	60.8	54.5	46.0	53.5	53.8
	IntentVLA	59.7	54.8	55.5	49.6	66.5	57.0
	ZR-0	41.7	<u>83.6</u>	<u>74.8</u>	72.0	**84.2**	69.3
Human Experience	DyPES-VLA	55.3	63.2	51.5	62.4	64.0	59.3
	JoyAI-RA 0.1	**71.7**	68.4	45.0	68.8	55.0	63.2
	ACE-Ego-0	59.0	83.2	70.5	<u>78.0</u>	76.5	<u>72.8</u>
Ours	**Nexus-VLA**	<u>70.3</u>	**89.6**	**80.0**	**83.6**	<u>81.0</u>	**80.5**

Table 2: OOD generalization on RoboCasa-GR1 under illumination changes, visual noise, novel spatial configurations and object-receptacle compositions, objects seen only in EgoDex, and fully unseen objects, with 500 rollouts per setting.

Method	Illumination	Visual Noise	Cross-Config.	Obj.-Receptacle	Ego-Object	Fully Unseen	Avg.
Baseline	62.0	58.4	51.2	45.0	40.8	31.6	48.2
Nexus-VLA	**76.8**	**73.6**	**71.0**	**66.8**	**69.4**	**55.6**	**68.9**

We train the model using 24,000 robot trajectories from RoboCasa-GR1, covering all 24 tasks, together with 21,446 filtered egocentric pick-and-place videos from EgoDex (Hoque et al., 2026). The two data sources are unpaired and jointly used to train a single policy following the human-robot mixed training strategy described in Section 3.2. We train the model for 200K steps with a global batch size of 512 and a learning rate of 3×10^{-5}. Training is conducted on two compute nodes, each equipped with eight NVIDIA H20E 141GB GPUs. The number of evaluation rollouts for each experiment is reported alongside the corresponding results.

On RoboCasa-GR1, we compare Nexus-VLA against foundation VLAs (GR00T-N1.6 and ABot-M0 (Yang et al., 2026)), semantic-enhanced methods (VP-VLA (Wang et al., 2026), IntentVLA (Lian et al., 2026), and ZR-0 (Li et al., 2026b)), and approaches leveraging human experience (DyPES-VLA (Li et al., 2026c), JoyAI-RA 0.1 (Zhang et al., 2026), and ACE-Ego-0 (Li et al., 2026a)). In real-world experiments, we additionally compare with $\pi_{0.5}$ (Intelligence et al., 2025) and GR00T-N1.6, together with our Baseline and intermediate model variants.

4.2 Overall Performance and Generalization

As shown in Table 1, on the standard RoboCasa-GR1 test distribution, Nexus-VLA achieves an average success rate of 80.5% across all 24 tasks, outperforming the strongest baseline, ACE-Ego-0, at 72.8% by 7.7 percentage points. Across the five task groups, Nexus-VLA achieves the highest success rates on Cuttingboard, Placemat, and Tray, and ranks second on both Close and Plate. These results demonstrate stable in-distribution performance across diverse manipulation tasks.

We further construct controlled OOD tests in RoboCasa-GR1 to evaluate generalization under visual changes, scene composition shifts, and novel objects. As shown in Table 2, Nexus-VLA achieves an average success rate of 68.9% across the six OOD settings, compared with 48.2% for the Baseline, corresponding to a 20.7 percentage-point gap. Under illumination changes and visual noise, Nexus-VLA achieves 76.8% and 73.6%, exceeding the Baseline by 14.8 and 15.2 percentage points, respectively. Under novel spatial configurations and object-receptacle compositions, the performance gaps further increase to 19.8 and 21.8 percentage points. These results indicate that the manipulation knowledge learned by Nexus-VLA transfers more robustly to new visual conditions and scene compositions, consistent with its hierarchical semantic modeling of task intent, object roles, and spatial relations.

Table 3: Component ablation on RoboCasa-GR1. Robot-only variants are matched by training steps, while co-training variants are matched by robot-data amount under a fixed human-robot ratio.

Method	Human-Robot Co-training	Hierarchical Alignment	Semantic Representation	Success Rate (%) ↑
Baseline	–	–	–	72.42%
+ Hierarchical Alignment	–	✓	–	75.67%
+ Human–Robot Co-training	✓	–	–	71.75%
+ Independent Latents	–	✓	Independent	76.92%
+ Unified Latent	–	✓	Unified	78.25%
Nexus-VLA	✓	✓	Unified	80.16%

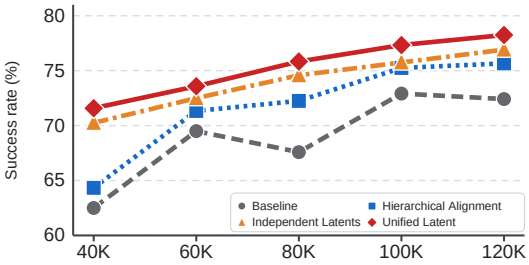

Figure 3: Training efficiency of different semantic modeling strategies.

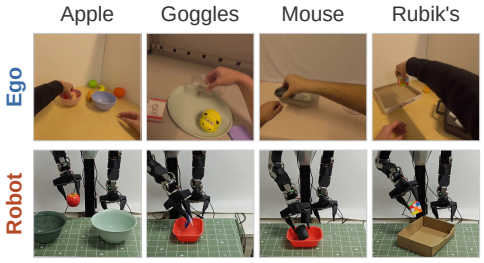

Figure 4: Cross-embodiment transfer from human experience to robot execution.

Object-level generalization further highlights the value of transferring human manipulation experience. For Ego-Objects that are absent from robot training data but observed in human demonstrations, Nexus-VLA achieves a success rate of 69.4%, exceeding the Baseline by 28.6 percentage points, the largest performance gap among the six OOD settings. Meanwhile, on Fully Unseen objects that appear in neither the human nor robot training data, Nexus-VLA still achieves 55.6%, exceeding the Baseline by 24.0 percentage points. The former demonstrates that human video experience can effectively complement robot-side unseen objects, while the latter suggests that the hierarchical semantic representation can further support generalization across object instances.

4.3 Ablation Study and Learning Efficiency

As shown in Table 3, hierarchical semantic modeling consistently improves policy learning. Compared with the Baseline, Hierarchical Alignment increases the success rate from 72.42% to 75.67%, a gain of 3.25 percentage points, indicating that multi-level supervision from task intent to fine-grained interaction provides useful structure for manipulation learning. Under the same hierarchical semantic supervision, replacing independent semantic representations with a Unified Semantic Latent further improves the success rate from 76.92% to 78.25%. This suggests that the gains arise not only from additional semantic supervision, but also from jointly representing manipulation knowledge across different semantic levels in a form that better supports action generation.

The co-training results further reveal how human data contributes to policy learning. With matched robot-data exposure, direct Human-Robot Co-training converges more slowly (71.75%), but reaches 76.32% at 200K steps, surpassing the Robot-only Baseline peak of 72.92%. In contrast, when Hierarchical Alignment and the Unified Semantic Latent are kept fixed, adding Human-Robot Co-training increases the success rate from 78.25% to 80.16%, providing an additional gain of 1.91 percentage points. This controlled comparison suggests that hierarchical semantic alignment and the unified latent representation better organize cross-embodiment shared knowledge, allowing human manipulation experience to contribute more effectively to robot policy learning.

The training curves further reveal how different semantic modeling strategies affect learning efficiency. As shown in Figure 3, Hierarchical Alignment consistently outperforms the Baseline throughout training, indicating that hierarchical semantic supervision from task intent to fine-grained

Table 4: Real-world manipulation and OOD generalization. We evaluate six real-world manipulation tasks under the standard setting and four controlled distribution shifts. Each setting is evaluated over 50 robot rollouts. All numbers are success rates (%).

Method	Standard	Object Distractor	Spatial Layout	Motion Constraint	Unseen Object	Avg.
$\pi_0.5$	64.0	48.0	50.0	44.0	40.0	49.2
GR00T-N1.6	58.0	42.0	44.0	38.0	34.0	43.2
Baseline	70.0	50.0	52.0	46.0	42.0	52.0
+ Human-Robot Co-training	74.0	56.0	58.0	52.0	48.0	57.6
+ Hierarchical Alignment + Unified Latent	78.0	68.0	66.0	62.0	58.0	66.4
Nexus-VLA	82.0	76.0	74.0	70.0	68.0	74.0

interaction provides more effective learning signals for action learning. Building on this, introducing semantic latents further accelerates performance improvement, while Unified Latent consistently outperforms Independent Latents. This indicates that the benefit does not simply come from adding multiple semantic prediction objectives, but from jointly encoding correlated manipulation knowledge across semantic levels into an internal representation for action generation. In particular, Unified Latent reaches 73.58% at 60K training steps, already surpassing the 72.92% achieved by the Baseline at 100K steps, attaining a higher success rate with approximately 40% fewer training steps. These results show that the unified hierarchical semantic representation improves not only final policy performance but also the efficiency with which semantic supervision is translated into action learning.

4.4 Real-World Experiments

We further evaluate Nexus-VLA on real-world robot manipulation tasks and construct four controlled OOD settings. Object Distractor introduces similar or irrelevant objects to interfere with target selection; Spatial Layout changes the spatial configuration between objects and targets; Motion Constraint introduces obstacles that alter feasible motion paths; and Unseen Object introduces manipulation objects absent from robot training. These settings probe semantic generalization along dimensions related to Intent, Spatial, Motion, and Interaction.

As shown in Table 4, Nexus-VLA achieves the highest success rate under both the standard setting and all four OOD settings, with an average success rate of 74.0%. Human-Robot Co-training alone provides only a modest improvement, whereas adding Hierarchical Alignment and Unified Latent leads to more substantial gains under OOD conditions, with the full Nexus-VLA achieving the best overall performance. Compared with the Baseline, the performance gap is 12.0 percentage points under the standard setting and 22.0–26.0 percentage points across the four OOD settings. This larger gap under distribution shifts suggests that Nexus-VLA not only improves standard-task performance but also learns representations that transfer more robustly across changes in targets, spatial layouts, motion constraints, and object instances.

We further investigate whether manipulation experience from human videos can transfer to real-world objects absent from robot training. To this end, we select six object categories from EgoDex that appear only in human demonstrations and construct corresponding real-world robot manipulation tasks, as shown in Figure 4. Each task is evaluated over 50 rollouts. The Robot-only Baseline achieves an average success rate of 47.0%, whereas Nexus-VLA reaches 72.7%, a difference of 25.7 percentage points. This result provides direct evidence that human-side manipulation experience can transfer to real-world robot execution even without corresponding robot demonstrations.

5 Conclusion

We present Nexus-VLA, an end-to-end framework for joint learning from unpaired human and robot demonstrations. By aligning cross-embodiment manipulation knowledge across semantic levels and encoding it into a shared Hierarchical Semantic Latent for action generation, Nexus-VLA enables human experience to directly improve robot policy learning. Experiments on RoboCasa-GR1 and real robots show improved manipulation performance, learning efficiency, and generalization. These results show that hierarchical semantic latents effectively transfer scalable human manipulation experience to robot control.

AI USE STATEMENT

In this work, we used generative AI tools only to aid and polish the writing of the manuscript, including improving clarity, grammar, and presentation. We did not use generative AI tools for other research tasks requiring disclosure, and the remaining tasks are not applicable to this work. All AI-assisted edits were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main paper provides the formulation, training objectives, experimental setup, and evaluation protocols of Nexus-VLA. The appendix further details the construction and quality control of the structured semantic supervision, the human-robot joint training procedure, model architecture, optimization settings, and additional experimental results. These details are provided to facilitate reproduction of the proposed method and experimental evaluation.

REFERENCES

Shuanghao Bai, Jing Lyu, Wanqi Zhou, Zhe Li, Dakai Wang, Lei Xing, Xiaoguang Zhao, Pengwei Wang, Zhongyuan Wang, Cheng Chi, et al. Latent reasoning vla: Latent thinking and prediction for vision-language-action models. *arXiv preprint arXiv:2602.01166*, 2026.

Homanga Bharadhwaj, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Zero-shot robot manipulation from passive human videos. *arXiv preprint arXiv:2302.02011*, 2023.

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.

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

Hanzhi Chen, Boyang Sun, Anran Zhang, Marc Pollefeys, and Stefan Leutenegger. Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27661–27672. IEEE, 2025a.

Sijin Chen, Kaixuan Jiang, Haixin Shi, Yanhui Wang, Weiheng Zhong, Haosheng Li, Bo Jiang, Yuxiao Liu, and Xihui Liu. Translation as a bridging action: Transferring manipulation skills from humans to robots. *arXiv preprint arXiv:2606.28133*, 2026.

Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19752–19763. IEEE, 2025b.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Ryan Hoque, Peide Huang, David Yoon, Jian Zhang, et al. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, volume 2026, pp. 4218–4237, 2026.

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

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Ji Woong Kim, Ke Wang, Zipeng Fu, Sirui Chen, Jeff Lai, Chelsea Finn, et al. Ego-pi: Vla finetuning for ego-centric human and robot data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1515–1524, 2026.

Sangoh Lee, Sangwoo Mo, and Wook-Shin Han. Act with intent: Distilling behavior intent for vision-language-action models. *arXiv preprint arXiv:2608.23478*, 2026.

Hao Li, Ganlong Zhao, Yufei Liu, Haotian Hou, Guoquan Ye, Tongyan Fang, Chunxiao Liu, Siyuan Huang, Jianbo Liu, Xiaogang Wang, et al. Ace-ego-0: Unifying egocentric human and robotic data for vla pretraining. *arXiv preprint arXiv:2606.17200*, 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.

Junfeng Li, Junjie He, Zhide Zhong, Yangyang Zheng, Pingyue Sheng, Jiayu Dong, Ruixin Li, Haodong Yan, Jiaguan Zhu, Tianran Zhang, et al. Dypes-vla: Learning shared dynamics priors and embodiment-specific control for cross-embodiment manipulation. *arXiv preprint arXiv:2608.06374*, 2026c.

Shijie Lian, Bin Yu, Xiaopeng Lin, Zhaolong Shen, Laurence Tianruo Yang, Yurun Jin, Haishan Liu, Changti Wu, Hang Yuan, Cong Huang, et al. Intentvla: Short-horizon intent modeling for aliased robot manipulation. *arXiv preprint arXiv:2605.14712*, 2026.

Juncheng Ma, Jianxin Bi, Yufan Deng, Xuanran Zhai, Kewei Zhang, Ye Huang, Bo Liang, Shukai Gong, Jiankai Tu, Xiaotian Tang, et al. Humanscale: Egocentric human video can outperform real-robot data for embodied pretraining. *arXiv preprint arXiv:2606.20521*, 2026.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

Zixuan Wang, Yuxin Chen, Yuqi Liu, Jinhui Ye, Pengguang Chen, Changsheng Lu, Shu Liu, Bei Yu, and Jiaya Jia. Vp-vla: Visual prompting as an interface for vision-language-action models. *arXiv preprint arXiv:2603.22003*, 2026.

Yifei Wei, Linqing Zhong, Yi Liu, Yuxiang Lu, Xindong He, Maoqing Yao, and Guanghui Ren. Libra-vla: Achieving learning equilibrium via asynchronous coarse-to-fine dual-system. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39799–39815, 2026.

Sicheng Xie, Haidong Cao, Zejia Weng, Zhen Xing, Haoran Chen, Shiwei Shen, Jiaqi Leng, Zuxuan Wu, and Yu-Gang Jiang. Human2robot: Learning robot actions from paired human-robot videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11078–11086, 2026.

Rongtao Xu, Jian Zhang, Minghao Guo, Youpeng Wen, Haoting Yang, Min Lin, Jianzheng Huang, Zhe Li, Kaidong Zhang, Liqiong Wang, et al. A0: An affordance-aware hierarchical model for general robotic manipulation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13491–13501. IEEE, 2025.

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, 2025a.

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

Yandan Yang, Shuang Zeng, Tong Lin, Xinyuan Chang, Dekang Qi, Junjin Xiao, Haoyun Liu, Ronghan Chen, Yuzhi Chen, Dongjie Huo, et al. Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*, 2026.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. In *International Conference on Learning Representations*, volume 2025, pp. 28213–28239, 2025.

En Yu, Haoran Lv, Jianjian Sun, Kangheng Lin, Ruitao Zhang, Yukang Shi, Yuyang Chen, Ze Chen, Ziheng Zhang, Fan Jia, et al. Dm0: An embodied-native vision-language-action model towards physical ai. *arXiv preprint arXiv:2602.14974*, 2026.

Tianle Zhang, Zhihao Yuan, Dafeng Chi, Peidong Liu, Dongwei Li, Kejun Hu, Likui Zhang, Junnan Nie, Ziming Wei, Zengjue Chen, et al. Joyai-ra 0.1: A foundation model for robotic autonomy. *arXiv preprint arXiv:2604.20100*, 2026.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

A Appendix

A.1 Hierarchical Semantic Supervision and Latent Representation Shaping

As described in the main text, the Hierarchical Semantic Latent is jointly shaped by four types of structured semantic supervision: Intent, Spatial, Motion, and Interaction. This section provides further details on the automatic construction of these supervision signals, the associated quality-control procedure, and the semantic projectors and training objectives. All supervision is generated offline by frozen Structured Manipulation Experts (SMEs) and used to shape the shared Semantic Latent through four lightweight trainable projectors: the Subtask, Grounding, Motion, and Contact Projectors.

A.1.1 Structured Manipulation Experts for Automatic Supervision

At timestep t, the structured supervision generated by the SMEs is denoted as

$$\mathcal{Y}_t = \left\{ \mathbf{y}_t^{\text{intent}}, \mathbf{y}_t^{\text{spatial}}, \mathbf{y}_t^{\text{motion}}, \mathbf{y}_t^{\text{interaction}} \right\}, \tag{9}$$

which describes the current manipulation intent, object-level spatial information, future motion, and local physical interaction, respectively. Although human and robot demonstrations differ in their observations and execution forms, the four supervision types are represented in semantically consistent spaces across embodiments, providing unified constraints for learning shared cross-embodiment representations. Figure 5 summarizes the four offline SMEs and their corresponding supervision-construction pipelines.

Task Progress Expert (TPE). TPE constructs the Intent supervision. We identify key interaction events from changes in hand or gripper opening and closing, and feed the start frame, end frame, and task instruction of each resulting manipulation segment into frozen Qwen3-VL to generate a local subtask description. Let the m-th manipulation segment be $\mathcal{S}_m = [t_m^s, t_m^e]$. The generated local subtask description is used as the Intent supervision for all $t \in \mathcal{S}_m$, such that frames within the same segment share the same local semantic target.

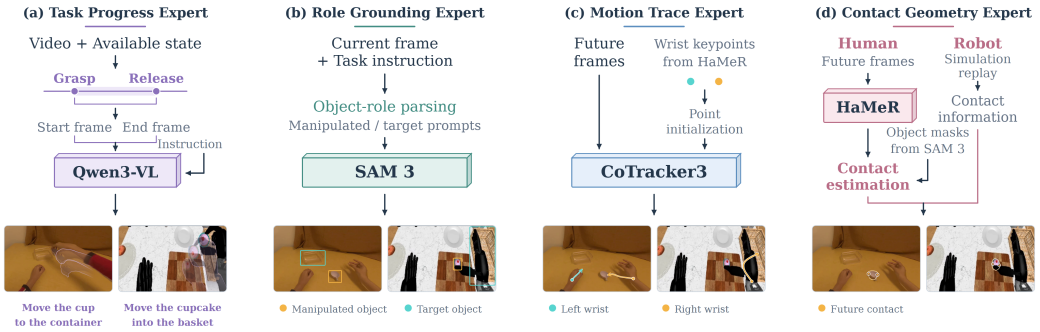

Figure 5: Overview of the four Structured Manipulation Experts for constructing Intent, Spatial, Motion, and Interaction supervision. All SMEs are frozen and used only for offline supervision generation.

Role Grounding Expert (RGE). RGE constructs the Spatial supervision. Given the task instruction, Qwen3-VL identifies the manipuland and target involved in the current manipulation and predicts the corresponding visual anchors. Frozen SAM 3 expands these anchors into object regions and produces frame-wise 2D bounding boxes together with visibility annotations. Specifically,

$$\mathbf{y}_t^{\text{spatial}} = \left\{ \mathbf{b}_t^{\text{manip}}, v_t^{\text{manip}}, \mathbf{b}_t^{\text{target}}, v_t^{\text{target}} \right\}, \tag{10}$$

where $\mathbf{b}_t \in \mathbb{R}^4$ denotes a 2D bounding box and v_t indicates whether the corresponding object is visible in the current view.

Motion Trace Expert (MTE). MTE constructs the Motion supervision. For human demonstrations, we obtain the left and right wrist locations from hand keypoints and use frozen CoTracker3 to track them over the following H frames:

$$\mathbf{y}_t^{\text{motion}} = \left\{ \mathbf{p}_{t+\tau}^L, \mathbf{p}_{t+\tau}^R \right\}_{\tau=1}^{H}, \qquad \mathbf{p}_{t+\tau}^h \in \mathbb{R}^2, \tag{11}$$

where $h \in \{L, R\}$ denotes the left or right wrist. For robot demonstrations, the corresponding trajectories are obtained from the end-effector states and converted into the same supervision representation.

Contact Geometry Expert (CGE). CGE constructs the Interaction supervision. For human demonstrations, we project the 3D hand keypoints recovered by HaMeR into the image plane and combine them with the manipuland mask provided by RGE to localize hand-object contact locations. These locations are used as contact anchors and propagated temporally to obtain future contact points and contact-state supervision for frames preceding the interaction. For robot demonstrations, the corresponding supervision is constructed from contact information obtained through simulation replay and converted into the same visual representation. The unified Interaction supervision is written as

$$\mathbf{y}_t^{\text{interaction}} = \left\{ \mathcal{C}_t^L, s_t^L, \mathcal{C}_t^R, s_t^R \right\}, \tag{12}$$

where $\mathcal{C}_t^h$ denotes the future contact locations expressed in the current image coordinates, and s_t^h denotes the corresponding contact state.

The SMEs are used only to construct training supervision offline. During policy inference, none of these experts are executed, nor are the corresponding intermediate semantics explicitly generated.

A.1.2 Data Quality Control and Low-Quality Sample Filtering

Because the structured supervision is generated automatically by offline experts, we apply hierarchical quality control and conservative filtering to reduce the effects of detection errors, tracking drift, and temporal noise. For recoverable local errors, we first rerun object localization and segmentation; a demonstration is discarded only when the corresponding supervision cannot be reliably recovered.

Quality control focuses primarily on visual annotation quality and temporal consistency. For visual annotations, we examine the completeness and temporal stability of the manipuland, target, and hand

Table 5: Architecture of the hierarchical semantic projectors.

Projector	Architecture	Semantic Target
Subtask Projector	Linear (2048→768) + 1-layer causal Transformer	Local subtask
Grounding Projector	MLP (2048→512→8) + visibility head	Manipuland-target grounding
Motion Projector	MLP (2048→1024→64) + visibility head	Future bimanual wrist trajectory
Contact Projector	MLP (2048→512→8) + existence head	Hand-object contact geometry

masks, including persistent missing regions, abrupt changes in position or scale, and tracking results that are inconsistent with the observed manipulation motion. Samples that remain ambiguous after these checks are further reviewed using a frozen vision-language model. For temporal consistency, we verify key interaction events, active-hand tracking, and the temporal alignment among video observations, state trajectories, and structured supervision.

Demonstrations that fail these quality checks and cannot be reliably recovered are removed at the episode level. For otherwise valid demonstrations with locally missing supervision, the corresponding invalid targets are masked out using validity masks while the remaining reliable supervision is retained.

A.1.3 Semantic projectors and Training Objectives

The Hierarchical Semantic Latent does not have explicit latent-level ground truth. Instead, we decode four types of structured semantics from the same $\mathbf{z}_t^{\text{sem}}$ and jointly shape its representation using the corresponding prediction losses. To encourage the semantic information to be encoded primarily in the shared latent representation, all four prediction modules use lightweight architectures, as summarized in Table 5.

To handle missing annotations, invisible targets, and invalid timesteps in a unified manner, we define the masked average as

$$\langle \ell_i \rangle_{\mathbf{m}} = \frac{\sum_i m_i \, \ell_i}{\sum_i m_i + \epsilon}, \tag{13}$$

where $\mathbf{m}$ denotes the validity mask associated with the corresponding supervision.

Intent. The Subtask Projector autoregressively predicts the local subtask description conditioned on $\mathbf{z}_t^{\text{sem}}$. Given the target token sequence $\mathbf{y}_t^{\text{intent}} = (y_1, \ldots, y_M)$, the training objective is

$$\mathcal{L}_{\text{intent}} = \langle -\log p(y_j \mid y_{<j}, \mathbf{z}_t^{\text{sem}}) \rangle_{\mathbf{m}^{\text{tok}}}, \tag{14}$$

where $\mathbf{m}^{\text{tok}}$ masks padding tokens and samples without valid text supervision.

Spatial. The Grounding Projector predicts the 2D bounding boxes of the manipuland and target together with their visibility. Let $\mathbf{a}$ and $\mathbf{v}$ denote annotation validity and visibility, respectively.Let $r \in \{\text{manip}, \text{target}\}$ index the object role. The corresponding loss is The corresponding loss is

$$\begin{aligned} \mathcal{L}_{\text{spatial}} = & \left\langle \mathcal{L}_{\text{box}}\left(\hat{\mathbf{b}}_t^r, \mathbf{b}_t^r\right) \right\rangle_{\mathbf{a} \odot \mathbf{v}} \\ & + \langle \text{BCE}(\hat{v}_t^r, v_t^r) \rangle_{\mathbf{a}}, \end{aligned} \tag{15}$$

where

$$\mathcal{L}_{\text{box}} = \text{SmoothL1} + \mathcal{L}_{\text{GIoU}}. \tag{16}$$

Motion. The Motion Projector predicts the future 2D trajectories of both wrists over $H = 16$ frames, together with their visibility:

$$\begin{aligned} \mathcal{L}_{\text{motion}} = & \left\langle \text{SmoothL1}\big(\hat{\mathbf{p}}_{t+\tau}^h, \mathbf{p}_{t+\tau}^h\big) \right\rangle_{\mathbf{m}^{\text{traj}} \odot \mathbf{v}} \\ & + \left\langle \text{BCE}\big(\hat{v}_{t+\tau}^h, v_{t+\tau}^h\big) \right\rangle_{\mathbf{m}^{\text{traj}}}, \end{aligned} \tag{17}$$

where $h \in \{L, R\}$, $\tau = 1, \ldots, H$, and $\mathbf{m}^{\text{traj}}$ denotes the validity mask for trajectory supervision.

Table 6: Training and optimization configuration of Nexus-VLA.

Configuration	Value	Configuration	Value
Action horizon	16	Training steps	200K
Per-GPU batch size	32	Gradient accumulation	1
Global batch size	512	Learning rate	3×10^{-5}
Optimizer	AdamW	LR schedule	Cosine decay
Warmup ratio	0.05	Weight decay	1×10^{-5}
Human : Robot sampling	1 : 2	Semantic branch weights λ_k	1.0
Semantic weight λ_{sem}	0.1	Training precision	BF16

Table 7: Architecture configuration of Nexus-VLA.

Configuration	VLM Backbone	Action Head
Backbone	Qwen2.5-VL-3B-Instruct	Flow-Matching DiT
Transformer layers	36 (LLM) / 32 (ViT)	32
Residual width	2048 / 1280	1536
Attention heads $\times$ dim.	16×128 / 16×80	32×48
Positional encoding	M-RoPE	Learned 1D + sinusoidal time
FFN intermediate width	11008 / 3420	6144
Input	Image + text + $\mathbf{q}_{\text{sem}}$	$\mathbf{s}_t$, $\mathbf{A}_t^\tau$, τ, $\mathbf{H}_t$
Output	$\mathbf{H}_t$, $\mathbf{z}_t^{\text{sem}}$	$\hat{\mathbf{v}}_t^\tau$

Interaction. The Contact Projector predicts key contact locations for both hands together with the corresponding contact-existence states:

$$\mathcal{L}_{\text{interaction}} = \Big\langle \text{SmoothL1}\Big(\hat{\mathbf{c}}_t^{h,q}, \mathbf{c}_t^{h,q}\Big) \Big\rangle_{\mathbf{p}} + \big\langle \text{BCE}\big(\hat{s}_t^h, s_t^h\big) \big\rangle_{\mathbf{a}}, \tag{18}$$

where q indexes contact roles within each hand, and $\mathbf{p}$ and $\mathbf{a}$ denote the validity masks for contact points and Interaction annotations, respectively.

The four prediction losses jointly define the semantic shaping objective:

$$\mathcal{L}_{\text{sem}} = \sum_{k \in \mathcal{K}} \lambda_k \mathcal{L}_k, \qquad \mathcal{K} = \{\text{intent}, \text{spatial}, \text{motion}, \text{interaction}\}. \tag{19}$$

During training, we set all $\lambda_k = 1.0$ and $\lambda_{\text{sem}} = 0.1$, and jointly optimize the semantic shaping objective with the action-generation objective:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}}. \tag{20}$$

These semantic projectors are used only to shape the representation during training. At inference time, the policy directly uses the semantically shaped $\mathbf{z}_t^{\text{sem}}$ for action generation without explicitly decoding any of the intermediate semantics above.

B Model and Training Implementation Details

This section provides additional details on the Human-Robot joint training strategy and the main optimization settings.

B.1 Human-Robot Joint Training

To jointly leverage human and robot demonstrations within a single policy, we adopt a unified action representation and write the robot action as the concatenation of joint-space and wrist-space components:

$$\mathbf{a}_t = \Big[\mathbf{a}_t^{\text{joint}}; \mathbf{a}_t^{\text{wrist}}\Big]. \tag{21}$$

Table 8: Per-task success rates (%) on all 24 RoboCasa-GR1 tasks. Tasks follow the Close, Cuttingboard, Placemat, Tray, and Plate group ordering in Table 1. Best and second-best values are boldfaced and underlined, respectively; ties receive the same formatting.

Task	ZR-0	JoyAI-RA 0.1	ACE-Ego-0	Nexus-VLA
CupToDrawerClose	20.0	**48.0**	<u>36.0</u>	**48.0**
PotatoToMicrowaveClose	59.0	<u>70.0</u>	58.0	**80.0**
MilkToMicrowaveClose	45.0	**84.0**	56.0	<u>74.0</u>
BottleToCabinetClose	39.0	<u>84.0</u>	78.0	**88.0**
WineToCabinetClose	40.0	<u>54.0</u>	**56.0**	**56.0**
CanToDrawerClose	47.0	**90.0**	70.0	<u>76.0</u>
CuttingboardToBasket	82.0	<u>88.0</u>	84.0	**92.0**
CuttingboardToCardboardBox	79.0	46.0	<u>84.0</u>	**94.0**
CuttingboardToPan	<u>92.0</u>	<u>92.0</u>	<u>92.0</u>	**96.0**
CuttingboardToPot	<u>85.0</u>	80.0	84.0	**88.0**
CuttingboardToTieredBasket	**80.0**	36.0	72.0	<u>78.0</u>
PlacematToBasket	78.0	76.0	<u>86.0</u>	**90.0**
PlacematToBowl	<u>87.0</u>	52.0	72.0	**92.0**
PlacematToPlate	<u>88.0</u>	38.0	80.0	**90.0**
PlacematToTieredShelf	<u>46.0</u>	14.0	44.0	**48.0**
TrayToCardboardBox	81.0	<u>82.0</u>	78.0	**86.0**
TrayToPlate	81.0	88.0	<u>90.0</u>	**92.0**
TrayToPot	74.0	88.0	**98.0**	<u>90.0</u>
TrayToTieredBasket	<u>74.0</u>	62.0	<u>74.0</u>	**76.0**
TrayToTieredShelf	<u>50.0</u>	24.0	<u>50.0</u>	**74.0**
PlateToBowl	<u>82.0</u>	48.0	68.0	**88.0**
PlateToCardboardBox	<u>81.0</u>	38.0	70.0	**84.0**
PlateToPan	**89.0**	46.0	<u>70.0</u>	<u>70.0</u>
PlateToPlate	85.0	<u>88.0</u>	**98.0**	82.0
Average	69.3	63.2	<u>72.8</u>	**80.5**

Robot demonstrations contain a 29-dimensional joint action and an 18-dimensional bimanual wrist pose, resulting in a 47-dimensional action representation. For human demonstrations, we use frozen HaMeR to recover the left- and right-wrist poses, each represented by a 3D position and a 6D rotation, yielding the shared 18-dimensional wrist-space action. Only these shared dimensions participate in action learning, while the robot-specific joint-space dimensions are masked out using an action validity mask. This allows human and robot data to share the same action-generation module without requiring nonexistent robot joint actions to be constructed for human demonstrations.

To further control the contribution of different embodiments during joint training, we adopt hierarchical probabilistic sampling. For a data source $d \in \{\text{human}, \text{robot}\}$, let $\mathcal{D}_d$ denote the set of datasets belonging to that source, and let n_j denote the number of samples in dataset j. The sampling weight of dataset j is defined as

$$w_j = s_d \frac{n_j}{\sum_{k \in \mathcal{D}_d} n_k}, \qquad j \in \mathcal{D}_d, \qquad \sum_d s_d = 1. \tag{22}$$

Here, s_d controls the overall sampling proportion of each embodiment, while datasets within the same source are sampled in proportion to their sizes. In our experiments, the Human : Robot sampling ratio is set to $1:2$, preventing the relative contribution of the two data sources from being determined directly by their original dataset sizes.

B.2 Model Architecture and Training Configuration

For reproducibility, Table 7 summarizes the main architecture of Nexus-VLA, while Table 6 reports the training and optimization settings. Unless otherwise specified, all experiments use the

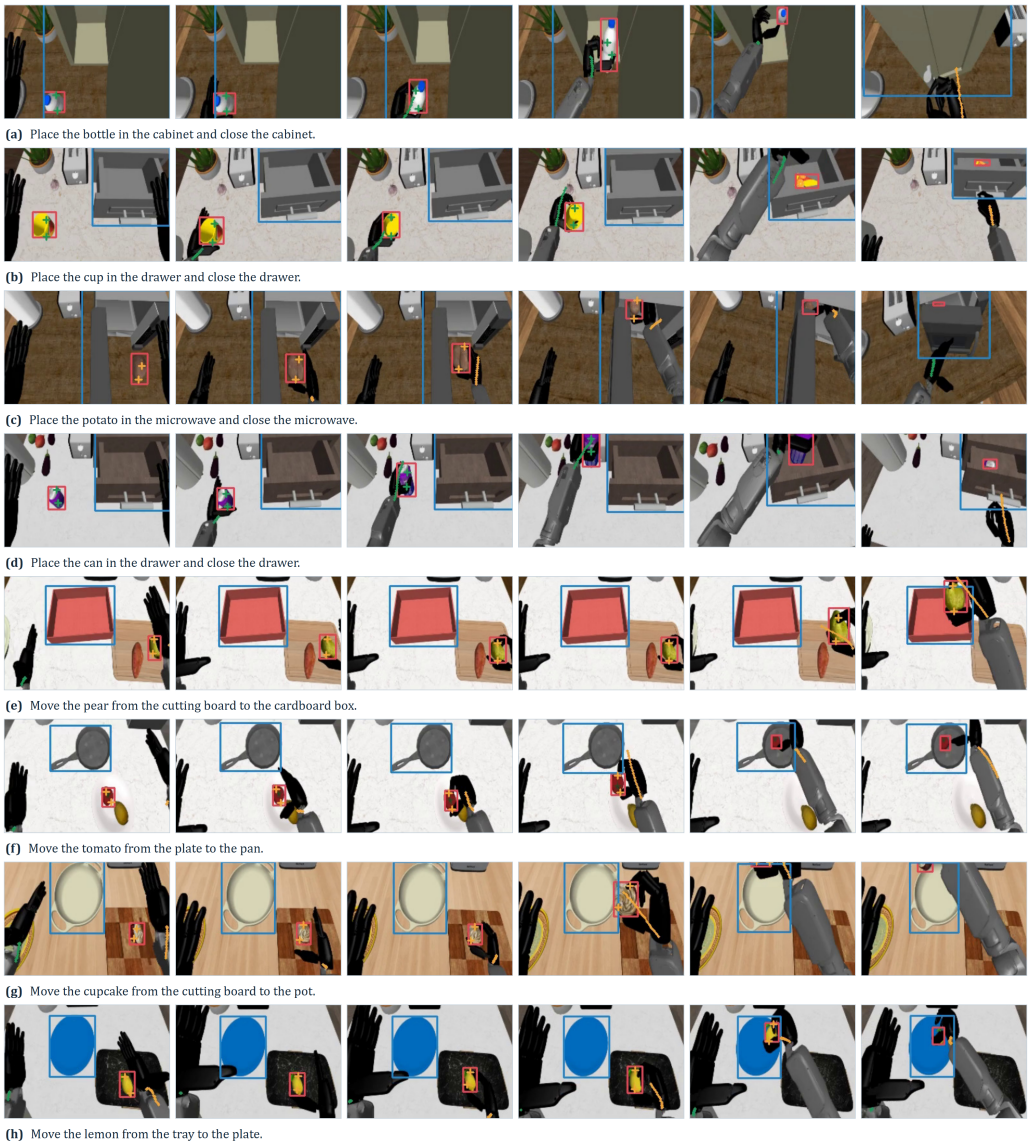

Figure 6: Qualitative visualization of eight RoboCasa-GR1 manipulation tasks. Each row shows six time-ordered frames, with colored overlays highlighting task-relevant objects, targets, and motion or interaction cues.

same model architecture and training configuration. The architectures and supervision targets of the semantic projectors are detailed in Section A.1.3.

C Supplementary Experimental Results

C.1 Additional RoboCasa-GR1 Results

RoboCasa-GR1 is our primary simulation benchmark and contains 24 humanoid manipulation tasks executed by the Fourier GR-1 robot. The task suite covers both object-transfer behaviors and interactions with articulated structures, including drawers, cabinets, and microwaves. In addition to the standard test distribution, we construct six controlled OOD settings to evaluate generalization under changes in visual appearance, spatial configuration, scene composition, and object identity.

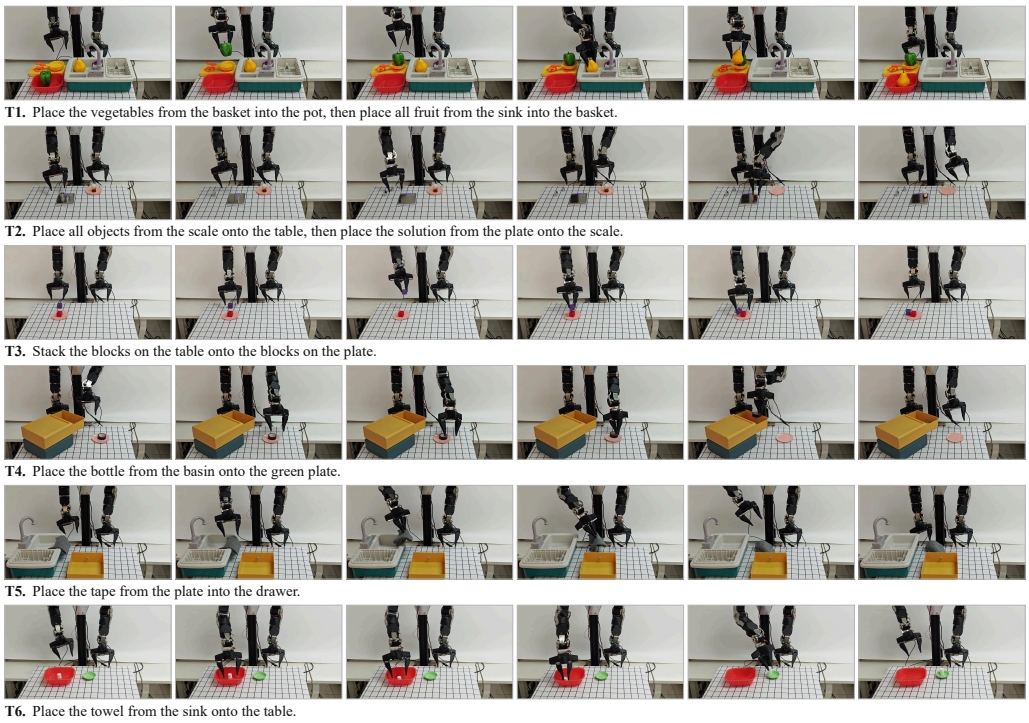

Figure 7: Real-robot executions under the standard setting. Rows T1–T6 correspond to the six task instructions printed below the image sequences.

Specifically, **Illumination** and **Visual Noise** perturb the visual observations while preserving the underlying task structure. **Cross-Configuration** evaluates novel spatial arrangements of task-relevant entities, while **Object-Receptacle** introduces object-target compositions not observed in the standard setting. We further consider two object-level shifts: **Ego-Object** uses objects absent from robot training but observed in the human demonstrations, whereas **Fully Unseen** uses objects that appear in neither the human nor robot training data. These settings complement the standard benchmark by progressively testing robustness to perceptual, spatial, compositional, and object-level distribution shifts.

Table 8 expands the five task-group results in Table 1 into success rates for all 24 RoboCasa-GR1 tasks. Nexus-VLA achieves the highest rate, including ties, on 18 tasks and the second-highest rate on five more. Its average success rate is 80.5%, 7.7 percentage points above ACE-Ego-0, the strongest competing method by overall average. The per-task results show that this improvement spans articulated-object tasks and several object-transfer groups rather than being driven by a single task family.The gains are especially clear on PlacematToBowl (92.0% versus 72.0% for ACE-Ego-0) and TrayToTieredShelf (74.0% versus 50.0%). Nexus-VLA also attains the best result on four of the six Close tasks, including ties, although JoyAI-RA 0.1 has a slightly higher Close-group average (71.7% versus 70.3%). The advantage is therefore not uniform: ZR-0 performs better on PlateToPan, and ACE-Ego-0 performs better on PlateToPlate. These cases identify task-level variation that is obscured by the aggregate benchmark score.

Because the Hierarchical Semantic Latent is an internal representation, Figure 6 visualizes task-relevant semantic cues alongside the executed motion rather than depicting the latent as an explicit symbol sequence. The eight representative RoboCasa-GR1 rollouts cover both articulated interactions and object transfers. Across six time-ordered frames per task, the marked object and destination remain associated with the approach, grasp, transport, placement, and closing stages as applicable. This qualitative view is consistent with the coarse-to-fine semantic structure used to shape the latent. The visual overlays serve interpretation; action generation does not require explicit intermediate semantic outputs at inference time, as described in Section 3.1.4.

C.2 Additional Real-World Results

Figure 7 presents representative real-robot executions for the six standard tasks. Each row shows six frames in temporal order, with the full instruction below the sequence. The examples include multi-stage transfers, block stacking, placement into a drawer, and moving objects between a sink or basin and a tabletop target. The frames make the execution progression visible beyond the success rates reported in Table 4.

To further evaluate robustness under controlled distribution shifts, we construct four real-world OOD settings while preserving the underlying task objective and success criterion. Each setting primarily stresses a different aspect of the hierarchical manipulation semantics.

Object Distractor introduces additional objects that are visually or semantically confusable with task-relevant objects, requiring the policy to maintain the correct object roles under clutter. For example, we place an additional tape roll on the table when the instruction specifies the tape on the plate, introduce plates with distractor colors when the target is the green plate, or add another towel outside the sink. The intended manipuland and target remain unchanged.

Spatial Layout changes the relative positions of the manipuland, target, and surrounding objects while preserving their identities and the task instruction. This setting evaluates whether the policy can re-ground task-relevant entities and execute the same manipulation under novel spatial configurations.

Motion Constraint introduces obstacles along the nominal motion path without changing the task goal or task-relevant objects. The robot must therefore adapt its wrist trajectory and use an alternative feasible motion path to complete the same manipulation.

Unseen Object replaces the manipulated object with an object absent from the robot training demonstrations while preserving the task structure. This setting evaluates whether manipulation behavior transfers to novel object appearances and geometries rather than relying on memorized object instances.

Following the task ordering in Figure 7, Figure 8 shows representative examples of the four OOD settings across the same six real-world tasks. Columns T1–T6 correspond to the six standard tasks, while rows represent Object Distractor, Spatial Layout, Motion Constraint, and Unseen Object.

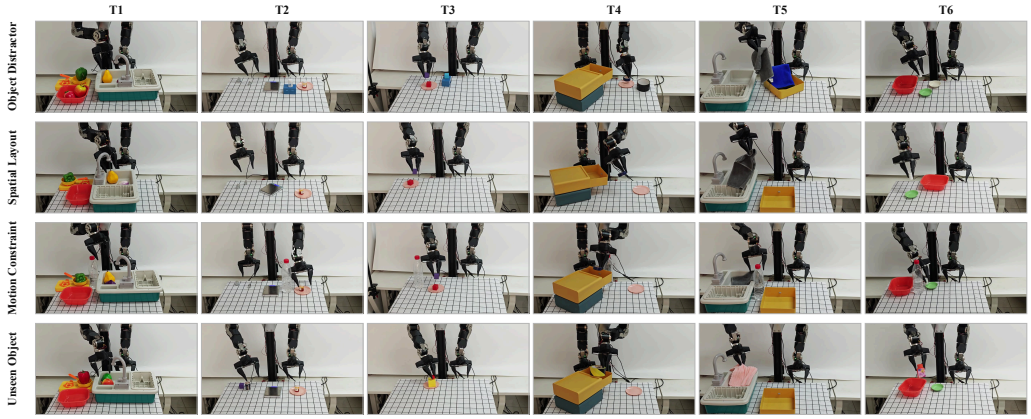

Figure 8: Representative real-robot scenes under four OOD conditions. Rows indicate the distribution shift, and columns T1–T6 correspond to the standard tasks in Figure 7. One frame is shown for each condition-task combination.