DIAL: Decoupling Intent and Action via Latent World Modeling for End-to-End VLA

Yi Chen[1]   **Yuying Ge**[2†]   **Hui Zhou**[2]   **Mingyu Ding**[3]   **Yixiao Ge**[2]   **Xihui Liu**[1†]

[1]The University of Hong Kong   [2]XPENG Robotics   [3]University of North Carolina at Chapel Hill
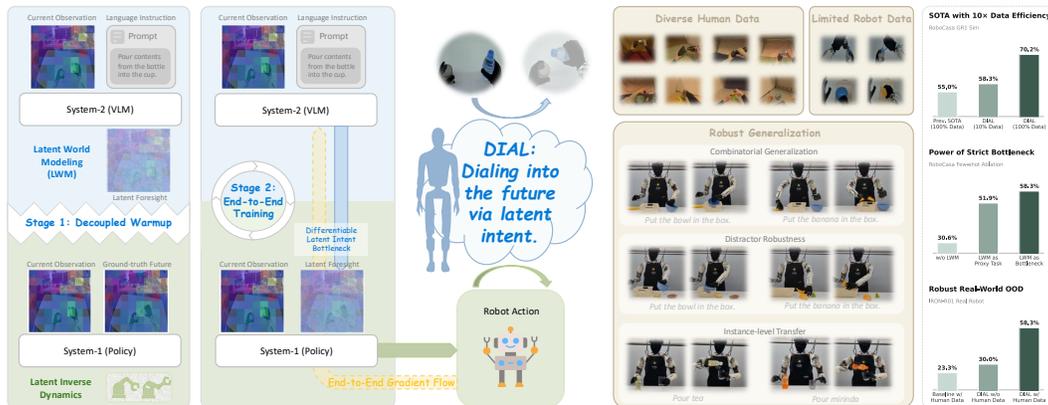
https://xpeng-robotics.github.io/dial

Figure 1: **Overview of the DIAL Framework.** DIAL bridges high-level decision making and low-level motor control through a **differentiable latent intent bottleneck**. **(Left)** System-2 (VLM) performs latent world modeling (LWM) to synthesize latent visual foresight within its native ViT feature space. This foresight serves as a structural bottleneck to convey the VLM's intent, which System-1 (Policy) then decodes into actions via latent inverse dynamics. A decoupled-to-unified training paradigm ensures stability, leveraging initial alignment in a consistent latent space to facilitate subsequent end-to-end refinement via action-aware gradients. **(Right)** Powered by this structural grounding, DIAL scales across heterogeneous human-robot data, achieving SOTA performance with $10\times$ higher data efficiency and robust zero-shot generalization to unseen real-world configurations.

Abstract

The development of Vision-Language-Action (VLA) models has been significantly accelerated by pre-trained Vision-Language Models (VLMs). However, most existing end-to-end VLAs treat the VLM primarily as a multimodal encoder, directly mapping vision-language features to low-level actions. This paradigm underutilizes the VLM's potential role in high-level decision making and introduces training instability, frequently causing degradation of its rich semantic representations. To address these limitations, we introduce **DIAL** (**D**ecoupling **I**ntent and **A**ction via **L**atent World Modeling), a framework that bridges high-level decision making and low-level motor execution through a differentiable latent intent bottleneck. Specifically, a VLM-based System-2 performs latent world modeling by synthesizing the latent visual foresight within the native feature space of the VLM's vision encoder; this foresight explicitly encodes the VLM's intent and serves as the structural bottleneck. A lightweight System-1 policy then decodes this predicted intent together with the current observation into precise robot actions via latent

---

† Corresponding authors.

inverse dynamics. To ensure optimization stability, we employ a two-stage training paradigm: a decoupled warmup phase in which System-2 learns to predict latent futures while System-1 learns motor control under ground-truth future guidance within a unified feature space, followed by seamless end-to-end joint optimization. This design enables action-aware gradients to refine the VLM backbone in a controlled manner while preserving its pre-trained knowledge. Extensive experiments on the RoboCasa GR1 Tabletop benchmark demonstrate that DIAL establishes a new state of the art, achieving superior performance with $10\times$ fewer demonstrations than prior methods. Furthermore, by leveraging heterogeneous human demonstrations, DIAL learns physically grounded manipulation priors and exhibits robust zero-shot generalization to unseen objects and novel configurations during real-world deployment on a humanoid robot.

1 Introduction

The development of generalist embodied agents has been significantly accelerated by pre-trained Vision-Language Models (VLMs)[1–4]. By internalizing massive semantic knowledge from the internet, VLMs provide a unified cognitive foundation capable of handling diverse multimodal tasks. Consequently, using these pre-trained models as cognitive backbones for robot policies has become a dominant trend [5–7]. However, effectively translating the abstract, high-level intent of a VLM into high-frequency, precise motor control remains a major challenge.

As summarized in Figure 2, existing approaches that leverage VLMs for robotic action generation face critical limitations. **Hierarchical Planners** [8–10] prompt the foundation model to generate high-level plans, typically text subtasks or code, to guide a separate low-level controller. While being interpretable and generalizable, this creates a non-differentiable wall that incurs high latency and prevents downstream action gradients from refining the VLM's physical understanding. In contrast, **End-to-End VLAs** [5, 11, 12] directly predict continuous actions. In practice, however, these approaches often treat the VLM primarily as a large multimodal encoder that extracts vision-language features, rather than allowing it to serve as a high-level decision maker that explicitly represents task intent. As a result, training under low-level action supervision can become unstable, frequently causing the VLM's semantic representations to collapse and overfit to spurious action patterns. Although auxiliary world modeling objectives [13–15] help by instilling physical dynamics and foresight, the absence of a strict structural bottleneck still permits the policy to rely on superficial correlations rather than truly translating the VLM's intent into precise motor commands.

This leaves the field in a structural dilemma: *How can we design an end-to-end VLA that strictly grounds the policy in the VLM's intent, seamlessly unifying cognitive generalization with execution precision?*

To resolve this dilemma, we present **DIAL** (**D**ecoupling **I**ntent and **A**ction via **L**atent-world-modeling), as shown in Figure 1. Inspired by the cognitive distinction between deliberate reasoning (System-2) and reflexive motor control (System-1), DIAL introduces *latent visual foresight* as a fully differentiable structural bottleneck between reasoning and execution. Rather than generating text or raw pixels, DIAL tasks the VLM (System-2) with predicting the future subgoal state entirely within the native feature space of its Vision Transformer (ViT) encoder, elevating it from a passive feature encoder to an active decision maker whose goal-directed foresight explicitly encodes the VLM's *latent intent* and structurally governs downstream execution. A lightweight flow-matching policy (System-1) then operates as a latent inverse dynamics model: it compares the current visual observation against this predicted intent to deduce the precise high-frequency motor commands needed to reach the anticipated goal.

This architectural decoupling yields three key advantages. First, it naturally supports a **stable decoupled warmup**: the VLM learns physical dynamics from diverse, action-free data while the policy independently masters sensorimotor control under ground-truth future guidance, preventing the gradient interference and representation collapse typical of naive joint training. Second, because the latent intent is continuous and provides a consistent interface between both systems, DIAL transitions smoothly into **end-to-end synergy**: action gradients flow back through the latent intent into the VLM, regularized by the same foresight reconstruction loss, encouraging the predicted intent to evolve into an actively "action-aware" representation without disrupting the VLM's pretrained knowledge. Third,
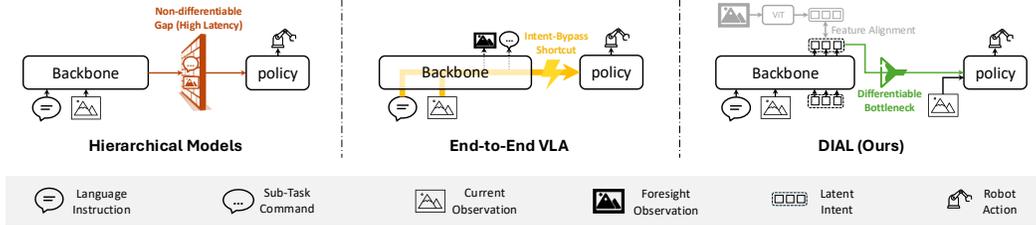
Figure 2: **Comparison of VLA Architectures. (Left) Hierarchical Models** decouple reasoning and execution via text or pixels, resulting in non-differentiable gaps and significant deployment latency. **(Middle) End-to-End VLAs** map multimodal features directly to actions. Even when auxiliary tasks are used, they are typically treated as optional context, which cannot strictly guarantee that actions are grounded in the VLM's intent. **(Right) DIAL (Ours)** introduces a differentiable latent bottleneck. By requiring System-1 to bridge the gap between current visual features and System-2's predicted latent foresight, DIAL ensures that execution is inherently anchored to the VLM's predictive intent.

DIAL enforces strict **structural grounding**. Unlike prior works that loosely append future features as auxiliary context, our inverse-dynamics design imposes a hard bottleneck: System-1 must resolve the discrepancy between current and predicted latent states to generate actions, effectively mitigating the shortcut learning that commonly afflicts existing VLAs.

We conduct extensive experiments to validate DIAL across both comprehensive simulations and real-world deployments. Latent visualizations further confirm that DIAL successfully grounds abstract linguistic instructions into a coherent, structurally aligned "visual roadmap" for the policy. In summary, our core contributions are:

- **Novel VLA Architecture:** We propose DIAL, an end-to-end framework that structurally bridges the cognitive generalization of VLMs with the execution precision of low-level policies. By utilizing latent visual foresight as a differentiable bottleneck, we ensure the generated actions are strictly grounded in the VLM's reasoning intent.

- **Decoupled-to-Unified Training Paradigm:** To ensure stable optimization, we introduce a targeted dual-warmup strategy. Using action-free data, the VLM warmup shift its abstract semantic knowledge toward physical world dynamics. Simultaneously, the policy independently learns to map low-level perception and specific visual goals into precise motor actions. This distinct dual-initialization prevents representation collapse and paves the way for seamless end-to-end fine-tuning.

- **State-of-the-Art Performance & Scalability:** DIAL achieves the highest reported performance on the RoboCasa GR1 Tabletop benchmark while using only 10% of robot demonstrations required by previous methods. It further shows strong scalability by successfully absorbing knowledge from diverse, cross-embodiment human data. Deployments on the IRON-R01-1.11 humanoid robot validate its reliable physical execution and impressive zero-shot transfer to novel scenarios.

2 Related Work

The integration of large pretrained foundation models [1–4, 16, 17] has driven a shift from task-specific policies to generalist robotic agents. By leveraging internet-scale pre-training, VLMs and Large Language Models (LLMs) provide embodied agents with robust semantic reasoning and instruction-following capabilities. Existing applications of these models in robotics can be broadly categorized into two dominant paradigms: hierarchical frameworks and end-to-end architectures. Within the latter, integrating predictive world modeling has recently emerged as a critical frontier to enhance physical grounding.

Hierarchical Abstraction vs. Low-Level Control. Hierarchical frameworks decouple high-level reasoning from low-level execution. Typically, LLMs or VLMs act as semantic planners, generating textual subtasks [8, 9, 18] or executable codes [19, 20] to drive separate downstream policies. Alternatively, some approaches employ heavy video diffusion models to predict pixel-level goal

images, followed by an inverse dynamics model to infer actions from observational histories [21]. However, these paradigms face critical challenges. Text-driven methods rely heavily on rigid human annotations and struggle with complex, high-frequency control due to deployment latency. Meanwhile, video-generation models incur prohibitive inference costs and lack the rich, internalized common-sense semantics inherent to VLMs. Fundamentally, the non-differentiable interface between the foundation model and the execution policy obstructs seamless collaboration, preventing the foundation model from acquiring the action-aware dynamics necessary for fine-grained manipulation.

End-to-End Vision-Language-Action Models. To bridge the gap between semantics and control, end-to-end VLA architectures directly map multimodal inputs to continuous robot actions. Early VLAs [5–7]) cast actions as discrete text tokens within the VLM's vocabulary, a paradigm further optimized by efficient action tokenization [22] and parallel decoding strategies [23]. Recently, a "dual-system" architecture has emerged as the dominant trend, pairing a VLM (System-2 for semantic understanding) with a lightweight, continuous action expert like diffusion or flow-matching based transformers (System-1 for precise execution) [24, 12, 25, 11, 26–29]. Despite empirical successes, an essential dilemma persists: end-to-end training frequently induces model collapse, degrading the VLM's pretrained knowledge. To mitigate catastrophic forgetting, current methods often truncate gradient flows [12] or freeze the VLM backbone entirely [11]. Consequently, the VLM is relegated to a passive representation encoder, leaving its core decision-making potential underexplored. While some works introduce Embodied Chain-of-Thought (CoT) [30, 31] to reactivate reasoning, these methods require costly annotations and introduce severe inference latency.

World Modeling within End-to-End VLAs. To address the lack of physical foresight in reactive policies and the weak coupling between high-level semantic reasoning and low-level action execution, recent work has increasingly integrated world modeling objectives. By anticipating future states, models are forced to develop a grounded understanding of physical dynamics. Early attempts [32, 33] utilize video pre-training to initialize policies with temporal priors. SEER [15] extends this by incorporating foresight-related features as an explicit condition for action generation. A prominent direction pursues unified autoregressive frameworks that jointly model sequences of discrete goal-image and actions tokens [14, 34, 35], enabling tighter integration of world prediction and decision-making. Rather than relying on expensive raw pixel-level generation, many recent methods achieve greater efficiency by predicting compact latent dynamics, such as through discrete latent actions or motion tokens that model inter-frame changes [36–38]. Alternatively, predicting continuous latents offers a more scalable approach: FLARE [13] leverages additional query tokens to align intermediate features with visual foresight, while UniCoD [39] employs a Mixture-of-Transformers (MoT) to unify world prediction and action execution. However, a fundamental limitation remains: most existing architectures treat visual foresight either as an auxiliary regularization task or simply append it to a long-context sequence. This loose coupling may be insufficient to enforce a strict causal dependency between the anticipated world states and the execution policy. Consequently, there remains a risk that models circumvent true physical understanding, degenerating instead into spurious shortcut learning.

3 Methodology

3.1 Model Overview

DIAL adopts a biologically inspired dual-system architecture to decouple high-level cognitive reasoning from low-level reactive execution, as illustrated in Figure 3. At each timestep t, given a language instruction l_t, the current visual observation o_t, and the robot's proprioceptive state q_t, DIAL generates an action chunk $A_t = [a_t, a_{t+1}, \ldots, a_{t+H-1}]$ with a horizon $H(= 16)$. The overall policy is formulated as a sequential composition of predictive intent generation and reflexive action decoding:

$$x_t = f_{\text{System-2}}(l_t, o_t), \quad A_t \sim \pi_{\text{System-1}}(\cdot \mid x_t, o_t, q_t) \tag{1}$$

Specifically, **System-2** (analogous to the "Brain") leverages a pretrained VLM to process multimodal inputs and synthesize a **latent intent** x_t. Instead of directly outputting discrete actions, System-2 is tasked with *latent world modeling* to envision the visual foresight of a future subgoal. Subsequently, **System-1** (the "Cerebellum") operates as a flow-matching-based reactive policy. It compares the
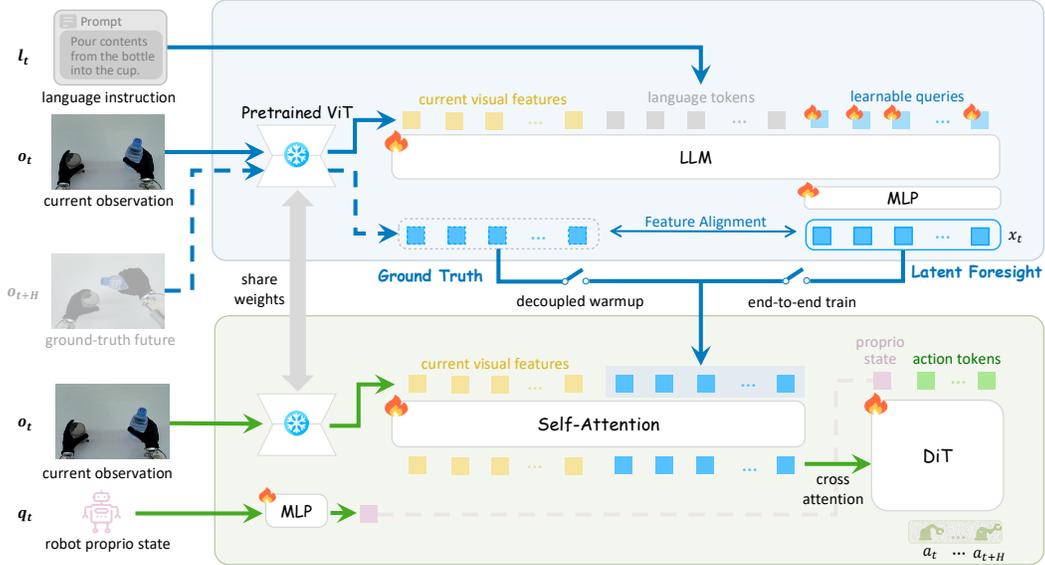
Figure 3: **The Dual-System Architecture of DIAL.** Built upon a pre-trained VLM, **System-2** (top) synthesizes a latent foresight (x_t) from language (l_t), current visual observation (o_t), and learnable queries via its LLM backbone and an MLP head. **System-1** (bottom) employs self-attention to fuse current and foresight visual features, serving as the cross-attention condition for a DiT-based action decoder. This decoder directly takes the projected proprioceptive state (q_t) and noisy action tokens to generate action chunks. To ensure feature consistency, both systems share the VLM's frozen pre-trained ViT. As indicated by the switches, the training transitions from a **decoupled warmup** (conditioned on ground-truth features of o_{t+H}) to **end-to-end optimization** (conditioned on x_t). Throughout both stages, an MSE loss is applied to align the latent foresight with ground-truth features.

current observation o_t against the predictive intent x_t, grounding the high-level semantic goal into precise, high-frequency motor commands.

3.2 System-2: Predictive Intent Synthesis via Latent World Modeling

System-2 utilizes a pre-trained VLM (e.g., Qwen2.5-VL-3B [3]) as its cognitive backbone. To endow the model with spatially aware predictive capabilities without the computational overhead of pixel-level reconstruction, we append N learnable query tokens to the LLM's input sequence. The LLM processes these tokens alongside the visual patches of o_t and the instruction l_t. The output representations corresponding to these queries are then passed through an MLP projection head to synthesize the latent intent $x_t \in \mathbb{R}^{N \times d}$. The number of queries N is set to match the number of visual patches extracted by the ViT from a single observation, preserving spatial structure.

We explicitly constrain x_t to encapsulate visual foresight by aligning it with the future state of the environment. Specifically, x_t is trained to predict the visual representation of the observation o_{t+H} at H timesteps ahead. To ensure optimization stability and strictly align feature spaces, the target foresight representation is extracted using the *identical* pre-trained ViT encoder, $\text{Enc}_{\text{ViT}}(\cdot)$, shared with the VLM backbone. The latent world modeling objective is optimized via Mean Squared Error (MSE):

$$\mathcal{L}_{\text{world}} = \|x_t - \text{Enc}_{\text{ViT}}(o_{t+H})\|_2^2 \tag{2}$$

By minimizing this loss, System-2 learns to translate abstract semantic instructions into a structured, forward-looking latent representation, providing actionable predictive guidance for System-1.

5

3.3 System-1: Reactive Motor Control as Latent Inverse Dynamics

Given the predictive intent x_t from System-2, System-1 focuses purely on resolving the immediate physical requirements to reach that anticipated future.

To achieve this, System-1 utilizes an independent perceptual pathway powered by the *same* pre-trained ViT encoder shared with System-2. This architectural choice enforces strict feature-space consistency. By mapping both the current observation o_t and the future intent x_t into a unified latent manifold, System-1 can directly discern fine-grained spatial and dynamic discrepancies without cross-modal alignment overhead.

Architecturally, System-1 employs a lightweight 4-layer self-attention module to fully fuse the multi-modal visual context, taking the current visual features $\text{Enc}_{\text{ViT}}(o_t)$ and the predictive intent x_t as inputs. The resulting spatially-aware fused representation serves as the conditioning signal for a 16-layer Diffusion Transformer (DiT), integrated via cross-attention layers. Concurrently, the low-dimensional proprioceptive state q_t is projected into a dense feature token via an MLP and fed directly into the DiT as part of the input sequence, alongside the noisy action tokens.

We formulate the action generation as an optimal transport problem using *flow matching*. Given a ground-truth action chunk $A_t = [a_t, a_{t+1}, \ldots, a_{t+H-1}]$, a time variable $\tau \sim \mathcal{U}[0,1]$, and Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we define the interpolated path $A_t^\tau = \tau A_t + (1-\tau)\epsilon$. System-1 learns a velocity field V_θ to approximate the target vector field by minimizing:

$$\mathcal{L}_{\text{fm}}(\theta) = \mathbb{E}_{\tau,\epsilon}\left[\|V_\theta(A_t^\tau \mid x_t, \text{Enc}_{\text{ViT}}(o_t), q_t, \tau) - (A_t - \epsilon)\|_2^2\right] \tag{3}$$

Conceptually, System-1 functions as a latent inverse dynamics model. Unlike traditional inverse models that compute actions directly from high-dimensional, noisy raw pixels, our approach resolves state-transition dynamics entirely within a structured latent space. This hierarchical separation isolates low-level execution from high-level reasoning, ensuring robust and high-frequency motor control.

3.4 Optimization Strategy: From Decoupled Warmup to End-to-End Synergy

The hierarchical design of DIAL naturally supports a stable two-stage training paradigm, transitioning from independent module initialization to fully differentiable joint optimization. Throughout both stages, all parameters of System-1 remain fully trainable. For System-2, we freeze the pre-trained ViT encoder and the text embedding layer of the VLM, while the rest of the network (including the LLM blocks, learnable queries, and the MLP projection head) is fully updated.

Stage 1: Decoupled Warmup. We first pre-train System-1 and System-2 independently to prevent posterior collapse. System-2 is optimized solely via $\mathcal{L}_{\text{world}}$ (Eq. 2) to master physically-grounded semantic reasoning and visual foresight. Concurrently, System-1 is trained via $\mathcal{L}_{\text{fm}}$ (Eq. 3) by substituting x_t with the *ground-truth* future visual features $\text{Enc}_{\text{ViT}}(o_{t+H})$. This ensures the "Cerebellum" learns optimal sensorimotor control under perfect future guidance, while the "Brain" concurrently learns to imagine that future.

Stage 2: End-to-End Training. Following the warmup, we unify the pipeline. System-1 is now conditioned directly on the synthesized latent intent x_t generated by System-2. Crucially, unlike traditional inverse dynamics pipelines that are disjointed and difficult to optimize jointly, our latent formulation is *fully differentiable*. This enables seamless end-to-end training, allowing downstream action-generation gradients (via $\mathcal{L}_{\text{fm}}$) to backpropagate smoothly through x_t and directly into the trainable parameters of the VLM backbone.

This gradient feedback is the cornerstone of DIAL: it forces the intent x_t to become explicitly action-aware. Rather than remaining a pure visual prediction, x_t evolves into a task-oriented representation strictly optimized for downstream motor execution. The overall training objective seamlessly combines both losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{world}} + \mathcal{L}_{\text{fm}} \tag{4}$$

This joint optimization effectively marries high-level cognitive planning with low-level physical grounding, providing a highly scalable template for end-to-end Vision-Language-Action (VLA) models.

| Croissant To Box | Bottle To Cabinet | Cupcake To Pan | Can To Drawer | Pumpkin To Shelf |

Figure 4: Examples from the 24 RoboCasa GR1 Tabletop Tasks, including object rearrangement (e.g., Croissant to Box) and interaction with articulated fixtures (e.g., Bottle to Cabinet).

4 Experimental Setup

4.1 Benchmarks and Datasets

4.1.1 RoboCasa GR1 Tabletop Simulation

We conduct simulation experiments on the RoboCasa benchmark utilizing the GR1 robot. As illustrated in Figure 4, our evaluation suite comprises 24 tabletop tasks, each assessed over 50 episodes. This suite includes 18 "Pick-and-Place" rearrangement tasks, where the robot follows language instructions to move objects between containers, and 6 "Articulated tasks" that involve more complex interactions such as placing objects inside and subsequently closing cabinets, drawers, or microwaves. We represent the robot's state and actions using a 47-dimensional vector. This vector comprises 29 joint-space degrees of freedom (DoF)—including the dual arms (14), hands (12), and waist (3)—along with 18 dimensions for the end-effector (EEF) poses (3D position and 6D rotation for each wrist). While the robot is ultimately commanded via its 29-dimensional joint space, we include the EEF poses to align the robot's representation with human data, following standard practice [28].

Robot-Only Training Regimes.   To evaluate effectiveness and data efficiency, we assess our method under two robot-only settings. The **Full Data** regime (Figure 7) utilizes 24,000 trajectories (1,000 per task) and involves 160,000 training steps. The **Few-Shot** regime (Figure 8) utilizes a 10% subset of 2,400 trajectories (100 per task) trained for 40,000 steps. For DIAL, both regimes follow a two-stage training schedule: the first half of the steps are dedicated to decoupled warmup, followed by end-to-end training for the remainder.

Learning from Human Data.   To further examine the scalability of DIAL, we investigate its ability to leverage large-scale human demonstrations to enhance generalization (Figure 8). We incorporate the `basic_pick_place` subset of the EgoDex dataset [40], which contains 27,419 trajectories of non-articulated pick-and-place interactions. To align human annotations with the robot's 47-dimensional state space, we extract the wrist EEF poses—the shared components between both embodiments—and pad the remaining dimensions. For human data, the state at time $t + 1$ is treated as the ground-truth action for the state at t. This large-scale dataset is combined with the few-shot robot set for a two-stage training process: a 40,000-step pre-training phase (split evenly between warmup and end-to-end training), followed by 20,000 steps of end-to-end fine-tuning exclusively on the few-shot robot data.

Generalization Scenarios.   To rigorously verify the generalization gains from human-centric priors, we constructed three Out-of-Distribution (OOD) testing scenarios using the assets from the RoboCasa benchmark: (1) *Unseen Appearance (18 Tasks)*, which introduces novel visual textures to familiar source-target container pairs; (2) *Unseen Combinations (14 Tasks)*, which requires manipulating seen objects within novel container pairings not encountered during training; and (3) *Unseen Object Types (32 Tasks)*, which tests the model's ability to generalize to entirely novel object categories across 32 different container combinations.

Figure 5: **Real-world Tasks and Data Sources.** Comparison between human demonstrations from the EgoDex dataset and corresponding robot executions (Pick & Place and Pouring) used for cross-embodiment learning on the IRON-R01-1.11 robot.
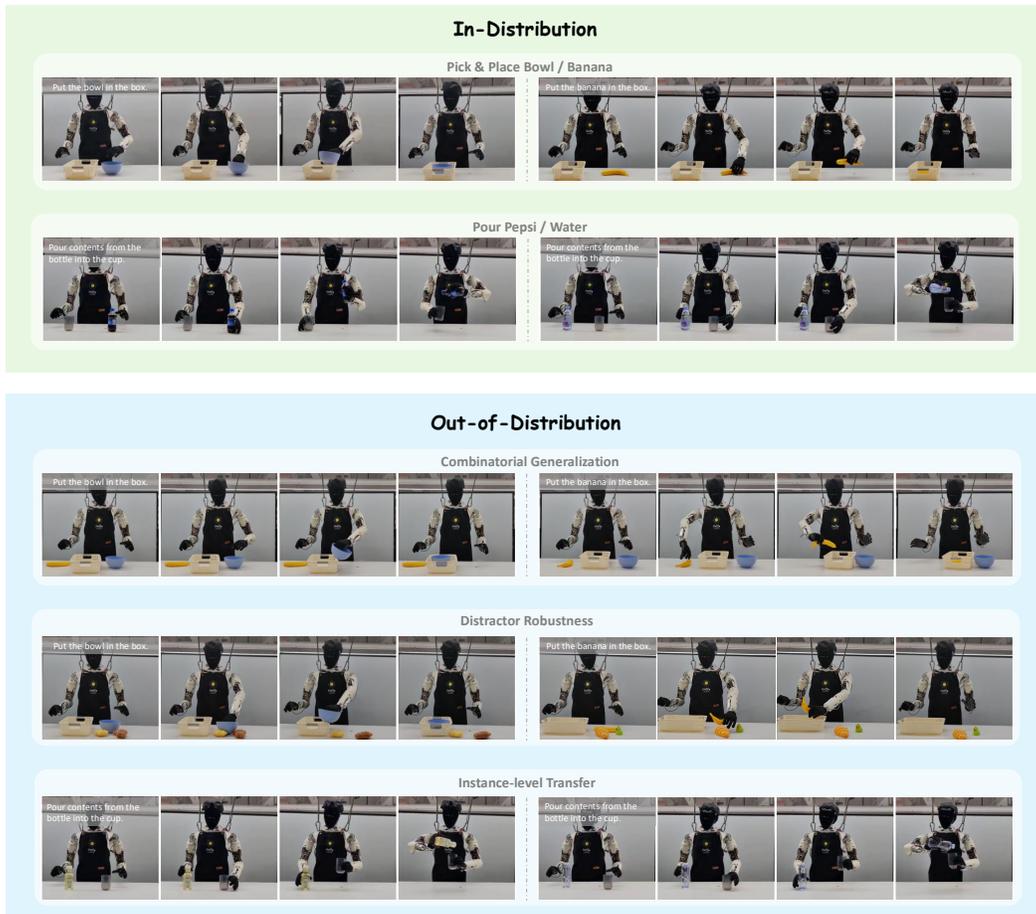

Figure 6: **Real-world Generalization Scenarios.** Comparison of in-distribution tasks and three OOD categories: combinatorial generalization (multiple seen objects), distractor robustness (unseen background items), and instance-level transfer (novel object types).

4.1.2 Real-World Experiments

We validate our method on the IRON-R01-1.11 robot using a 50-dimensional state and action space. This setup extends the simulation configuration (arms, hands, waist, and EEF poses) by incorporating an additional 3-DoF head. As shown in Figure 5, we design two real-world tasks analogous to representative EgoDex subsets:

- **Pick & Place:** Mimicking the EgoDex `basic_pick_place` subset (27,419 trajectories), the robot must pick up an object (e.g., a bowl or banana) and place it into a box.

8

- **Pouring:** Mimicking the EgoDex `pour` subset (3,205 trajectories), the robot must grasp a bottle with one hand and a cup with the other, performing a pouring motion.

Training Protocol. For each task, we collected 120 robot trajectories in a laboratory setting. All models follow a two-stage training paradigm: a large-scale pre-training phase of 160,000 steps on a mixture dataset comprising 32k proprietary factory-collected robot trajectories and 30k EgoDex trajectories, followed by a task-specific fine-tuning phase of 2,000 steps. To ensure a fair comparison, DIAL allocates the pre-training phase into 80,000 steps of decoupled warmup and 80,000 steps of end-to-end training, matching the total training steps of the baselines.

Generalization Scenarios. To further evaluate the model's robustness and generalization capabilities, we establish three OOD testing scenarios. (1) *Combinatorial Generalization*: In the pick-and-place task, both a banana and a bowl are simultaneously present in the workspace, requiring the model to correctly disambiguate and follow specific language instructions. (2) *Distractor Robustness*: We introduce unseen objects into the workspace as distractors while the target object (e.g., a banana or bowl) is present, testing the policy's resilience to visual clutter. (3) *Instance-level Transfer*: In the pouring task, the model is required to manipulate novel object instances, such as beverage or mineral water bottles with previously unseen geometries, sizes, or liquid colors.

4.2   Compared Methods

We evaluate DIAL against a diverse set of methods, ranging from established policy learning frameworks to state-of-the-art VLA architectures, as well as several controlled variants designed to isolate component contributions.

4.2.1   Representative Prior Arts

We first compare DIAL with standard policy learning frameworks, including: **Diffusion Policy** [41], which models action generation via U-Net denoising; **UWM** [42], a transformer unifying action and video diffusion processes; and **FLARE** [13], a flow-matching framework with future latent alignment. For VLA comparisons, we include **GR00T-N1.6** [26], an upgraded GR00T framework incorporating a larger DiT backbone and fine-tuned late-stage VLM layers.

Qwen3-based VLAs. To ensure a fair comparison against models equipped with cutting-edge VLM backbones, we evaluate several architectures implemented via *StarVLA* [43], all utilizing Qwen3-VL [4]:

- **GR00T-Qwen3** [11]: Adapts the GR00T dual-system, combining a frozen Qwen3-VL for vision-language representations with a flow-matching DiT for action generation.
- π-**Qwen3** [24]: Couples per-layer VLM representations with a flow-matching expert via KV caching.
- **FAST-Qwen3** [22]: Employs FAST action tokenization for efficient autoregressive prediction.
- **OFT-Qwen3** [23]: An optimized fine-tuning recipe featuring parallel action-chunk decoding.

4.2.2   Controlled Variants and Ablations

To rigorously isolate the impact of DIAL's specific architectural designs, we construct several controlled variants and ablation models, primarily based on the same Qwen2.5-VL-3B [3] backbone used in DIAL:

- **GR00T-Qwen2.5 / -FT:** Replicates the GR00T architecture using Qwen2.5 as the VLM backbone. We consider two settings: (i) a frozen variant, where both the ViT encoder and the LLM are fixed, and (ii) an LLM-tuned variant (-FT), which updates the core language modeling blocks while keeping the ViT encoder and text embedding layers frozen.
- **GR00T-Qwen2.5 + FLARE / SEER:** Extensions of the fine-tuned (-FT) baseline that incorporate future-aligned query tokens, inspired by FLARE [13] and SEER [15]. Specifically, the FLARE-style variant uses these tokens solely for latent regularization, whereas the SEER-style variant explicitly concatenates them with vision-language features to condition the DiT.
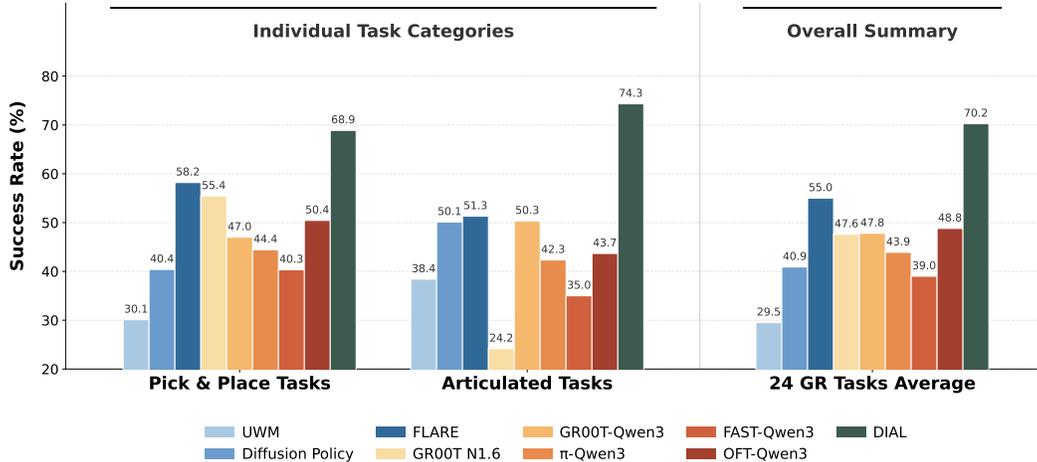
Figure 7: Results on RoboCasa GR1 Tabletop Simulation with full training data.

- **GR00T-Qwen2.5 + SEER-EV:** Augments the SEER variant with an independent perceptual path (Extra Vision) for System-1. This isolates the effect of decoupled perception but lacks DIAL's specialized bottleneck design.

- **DIAL-DINO:** Replaces the visual foresight target (for System-2) and the current observation encoder (for System-1) with an external DINO-v2 [44] encoder. Since System-2 still processes its current visual input via the VLM-native encoder, this variant tests the impact of forcing the high-level "Brain" and the low-level "Body" to communicate across mismatched latent spaces.

- **DIAL w/o Human Data:** Evaluates DIAL trained solely on robot data to quantify the gains from human data pre-training. Note that in certain evaluations where human data is entirely excluded for all baselines, the default "DIAL" entry represents this robot-only version. The explicit distinction is only made in Fig.9, Fig.10, and Fig.11.

5 Experiments

In this section, we empirically evaluate DIAL. To structure our analysis, we formulate five core research questions:

- **Q1 (Performance):** Does DIAL outperform state-of-the-art baselines in task success rate and sample efficiency on the public humanoid simulation benchmark?

- **Q2 (Architecture):** Which architectural designs are essential for grounding the VLM's high-level intent in low-level control?

- **Q3 (Scalability):** Can DIAL effectively scale by leveraging heterogeneous human demonstrations to enhance both in-distribution and OOD performance?

- **Q4 (Robustness):** Can DIAL achieve robust real-world generalization, and what role does the decoupled warmup play in achieving this?

- **Q5 (Interpretability):** Do the predicted latent foresights capture semantically meaningful, task-relevant dynamics in the VLM's native feature space?

5.1 Overall Performance Comparison

We benchmark DIAL against state-of-the-art policies on the comprehensive RoboCasa GR1 Tabletop simulation suite. As shown in Figure 7, DIAL achieves an average success rate of 70.2%, substantially outperforming the strongest baseline, FLARE (55.0%), as well as advanced VLA architectures such as GR00T-N1.6 (47.6%). This consistent margin establishes DIAL as the new state of the art on this benchmark.
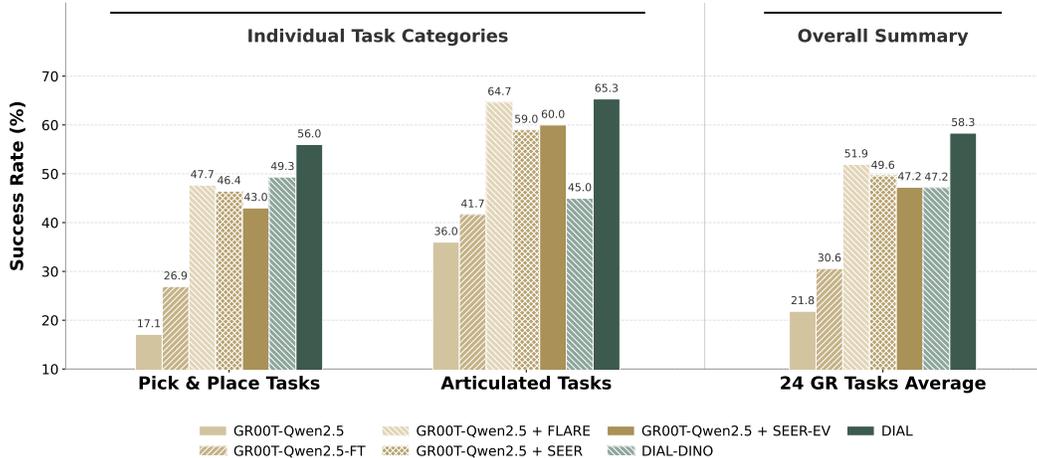
Figure 8: Results on RoboCasa GR1 Tabletop Simulation under the few-shot setting.

A closer examination across task categories further reveals DIAL's robustness. In Pick & Place tasks, which require precise grounding of semantic instructions into object rearrangement behaviors, DIAL attains a 68.9% success rate, significantly surpassing existing VLA variants and demonstrating stronger instruction-to-action alignment. In Articulated Tasks, where the robot must manipulate articulated objects, DIAL achieves an even higher 74.3% success rate. The strong performance across both categories highlights DIAL's balanced capability, maintaining consistently high effectiveness across distinct task types. We attribute this stability to DIAL's dual-system decoupling, which provides a structured interface between high-level intent and low-level control.

To assess data efficiency, we further evaluate DIAL under a strict few-shot setting. As shown in Figure 8, when trained with only 100 trajectories per task, DIAL achieves a 58.3% success rate. Notably, this performance already surpasses FLARE (55.0%) trained under the full-data regime with 1,000 trajectories per task (Figure 7). This $10\times$ reduction in demonstration requirements, while maintaining superior performance, highlights the strong inductive bias introduced by DIAL's structural bottleneck, enabling highly scalable and data-efficient robot learning.

5.2 Ablation on Bridging Mechanisms

To isolate the source of DIAL's improvements, we systematically decompose the architecture to evaluate the contributions of world modeling objectives, the System-1/System-2 interface design, and feature alignment. All ablations are conducted in the few-shot regime (100 trajectories per task), with results summarized in Figure 8.

Effect of World Modeling Objectives. We begin with standard VLA baselines that primarily treat the VLM as a multimodal encoder without explicit world modeling supervision. The frozen GR00T-Qwen2.5 baseline achieves only 21.8% success, and full fine-tuning (-FT) improves performance modestly to 30.6%. This limited gain suggests that merely updating parameters in a tightly coupled architecture is insufficient. Without an explicit objective that grounds latent representations in future world states, the policy tends to overfit to proprioceptive cues while under-utilizing complex visual observations. These results highlight the necessity of incorporating structured predictive signals rather than relying solely on scale or fine-tuning.

Interface Design. We next examine how intent should be communicated from System-2 to System-1. Architectures with *loose coupling* treat the predicted intent as optional context. For instance, +SEER (49.6%) concatenates future tokens to the policy input, while +SEER-EV (47.2%) further provides System-1 with a direct visual perception path. Both variants fail to surpass the 50% success rate. Similarly, +FLARE, which introduces future prediction as an auxiliary training objective without enforcing its usage at execution time, achieves 51.9%.
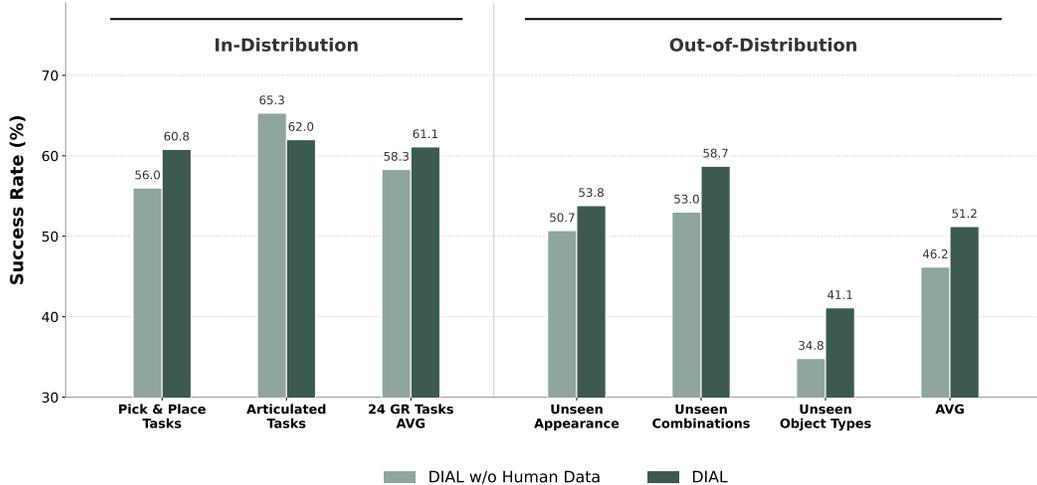
11

Figure 9: Impact of incorporating EgoDex `basic_pick_place` human demonstrations on few-shot performance in RoboCasa GR1 simulation tasks.

When the intent signal is not structurally enforced, the policy tends to bypass the high-level intent and instead exploit superficial shortcuts, since there is no architectural constraint requiring the intent to be faithfully translated into action. Notably, adding an extra visual pathway in SEER-EV even degrades performance ($49.6\% \rightarrow 47.2\%$), indicating that unconstrained access to raw features exacerbates the tendency to ignore cognitive foresight.

In contrast, DIAL introduces a *structural bottleneck* by formulating System-1 as a latent inverse dynamics model, where the synthesized intent serves as an indispensable target rather than auxiliary context. This design enforces that actions must be derived by bridging the current observation and the explicitly predicted future state. As a result, DIAL achieves a state-of-the-art **58.3%**, substantially outperforming all alternative interfaces in the low-data regime.

Feature Alignment. Finally, we investigate whether DIAL's gains depend on the native VLM feature space. In DIAL-DINO, we replace the internal ViT features with DINO-v2 representations. Despite their strong geometric priors, performance drops from 58.3% to 47.2%.

This degradation indicates that geometric richness alone is insufficient; what matters is latent consistency between reasoning and control. In DIAL-DINO, System-2 reasons in the VLM's native semantic space but must project its intent into a different feature manifold for execution, creating a semantic–physical misalignment. By contrast, DIAL's shared native ViT ensures that both systems operate within a unified latent space, eliminating cross-manifold translation and maximizing the fidelity of intent-to-action transfer.

Across all ablations, the results consistently demonstrate that effectively connecting a dual-system architecture requires more than auxiliary supervision or loose input concatenation. Superior performance is achieved only when (i) intent is grounded via explicit world modeling, (ii) the architecture enforces a structural dependency to prevent shortcut learning, and (iii) reasoning and control share a unified latent space. Together, these findings validate the necessity of DIAL's core design: explicit system decoupling bridged by a strict, representationally consistent predictive bottleneck.

5.3 Scalability via Human Data

We evaluate DIAL's ability to leverage cross-embodiment human demonstrations from the EgoDex `basic_pick_place` subset to improve both in-distribution performance and OOD generalization in RoboCasa GR1 simulation tasks. Results are detailed in Figure 9.

In-Distribution Performance. Incorporating human data yields a clear benefit for Pick & Place Tasks, with success rates rising from 56.0% to 60.8%. This indicates that System-2 effectively distills
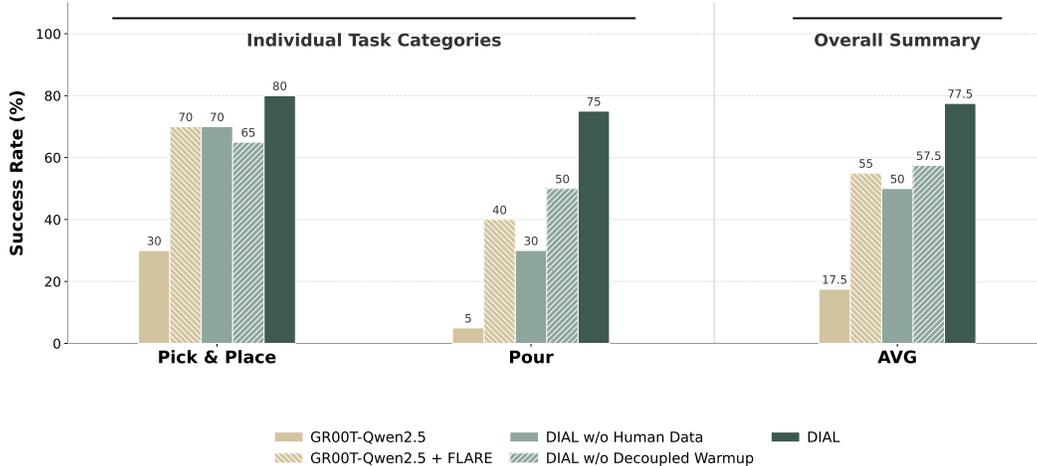
Figure 10: In-distribution experiment results on the IRON-R01-1.11 robot.

the core logic of pick-and-place actions—such as reaching, grasping, and spatial placement—from diverse human demonstrations. In contrast, Articulated Tasks show no improvement (62.0% with human data vs. 65.3% without), primarily due to a domain mismatch: the EgoDex `basic_pick_place` subset contains only pure rearrangement interactions and lacks demonstrations involving articulated objects. As a result, the human-centric prior provides little guidance for this task category.

Out-of-Distribution Generalization. The integration of human data has a pronounced impact on OOD performance, boosting zero-shot generalization across all three metrics. Success rates increase from 34.8% to 41.1% for *unseen object types*, from 53.0% to 58.7% for *unseen combinations*, and from 50.7% to 53.8% for *unseen appearances*. These improvements suggest that exposure to diverse human-object interactions enables the VLM (System-2) to acquire a more robust semantic understanding of manipulatable objects, focus on abstract goals rather than specific object-container pairings, and handle novel visual appearances more effectively. Overall, the average OOD success rate rises from 46.2% to 51.2%, highlighting DIAL's strength in cross-embodiment learning.

Overall, these results indicate that DIAL scales effectively: while the magnitude of gains depends on task coverage in the human dataset, integrating human demonstrations consistently enhances semantic reasoning and substantially improves OOD robustness.

5.4 Real-World Robustness and Stability

To further validate DIAL's physical viability and training stability, we deploy the model on the real-world IRON-R01-1.11 robot, with results shown in Figure 10 and Figure 11.

Importance of Decoupled Warmup. During real-robot deployment, we observe that the decoupled warmup phase is crucial for the model's training stability. Ablation results indicate that removing this warmup leads to a severe drop in performance for both in-distribution tasks (average success rate drops from 77.5% to 57.5% in Figure 10) and out-of-distribution scenarios (from 58.3% to 30.0% in Figure 11). Without warmup, System-2 fails to form coherent visual foresight before System-1 overfits to early noise, destabilizing joint optimization. The warmup phase allows System-1 to reliably track *perfect* future states before confronting *predicted* intentions, providing a stable foundation for generalization.

Generalization in Complex Environments. Empowered by stable training, DIAL exhibits strong real-world generalization across several practical challenges (Figure 11). In *combinatorial generalization* tests, despite the fine-tuning dataset containing only single-object pick-and-place trajectories, DIAL successfully identifies and manipulates the target object among multiple familiar objects based on language instructions, whereas baseline methods largely lose their instruction-following capabilities. In scenarios with *unseen background distractors*, System-1 isolates the target by comparing
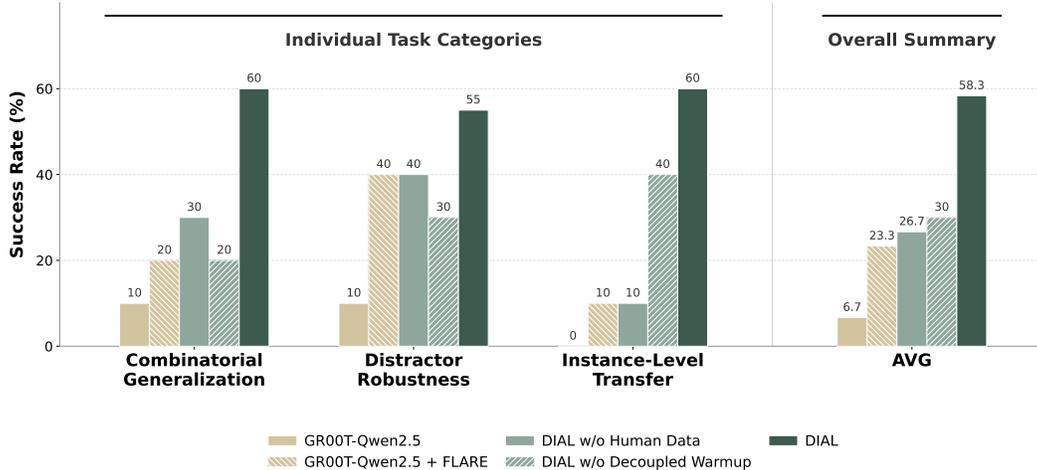
13

Figure 11: Out-of-distribution experiment results on the IRON-R01-1.11 robot across three generalization challenges: combinatorial generalization, distractor robustness, and instance-level transfer.

the latent encoding of the current observation with System-2's predicted visual foresight, effectively filtering out background noise and preventing shortcut learning. In the more demanding *instance-level transfer* setting, DIAL reliably grasps previously unseen bottles with different shapes and liquid colors to complete pouring tasks, demonstrating precise and adaptable control.

Role of Human Data.   We further find that incorporating human demonstrations during pre-training is essential for robust real-world performance. Removing human data significantly degrades OOD success rates (from 58.3% to 26.7% in Figure 11), highlighting the importance of diverse human priors for semantic reasoning and cross-embodiment generalization.

Overall, these real-robot results show that DIAL combines stable training with robust execution: decoupled warmup stabilizes joint optimization, human data enriches semantic understanding, and the system generalizes effectively across combinatorial, distractor, and instance-level challenges.

5.5   Interpreting Latent Foresight

To understand the semantic structure of the information passed between System-2 and System-1, we qualitatively analyze the learned latent representations using Principal Component Analysis (PCA). By projecting the high-dimensional features of the current observation, the ground-truth future, and the predicted foresight onto their first three principal components (mapped to RGB channels), we can visually interpret the model's internal foresight.

As illustrated in Figure 12, the spatial color distributions in the first three columns reflect the latent manifold structure. Taking the task *"Pick up the silver iPhone"* (Row 3) as an example, the Predicted Foresight closely mirrors the Ground-Truth Future, especially in task-relevant regions such as the target object and destination container (highlighted by circles). Conversely, these predicted features diverge significantly from the Current Observation precisely in the areas where physical manipulation is expected to occur. The final column further illustrates these anticipated changes through the per-patch cosine distance between the predicted foresight and current observation features, with warmer colors indicating greater deviation.

This structural alignment between the predicted and ground-truth futures, coupled with their deliberate divergence from the initial state (evident in both the PCA color maps and the change heatmaps), proves that System-2 does not merely reconstruct the current scene. Instead, it actively anticipates meaningful state transitions within the semantic space. These visualizations confirm that DIAL's predictive bottleneck successfully serves as a spatially aligned and semantically grounded bridge: generating a coherent "visual roadmap" that System-1 subsequently decodes into precise physical actions.
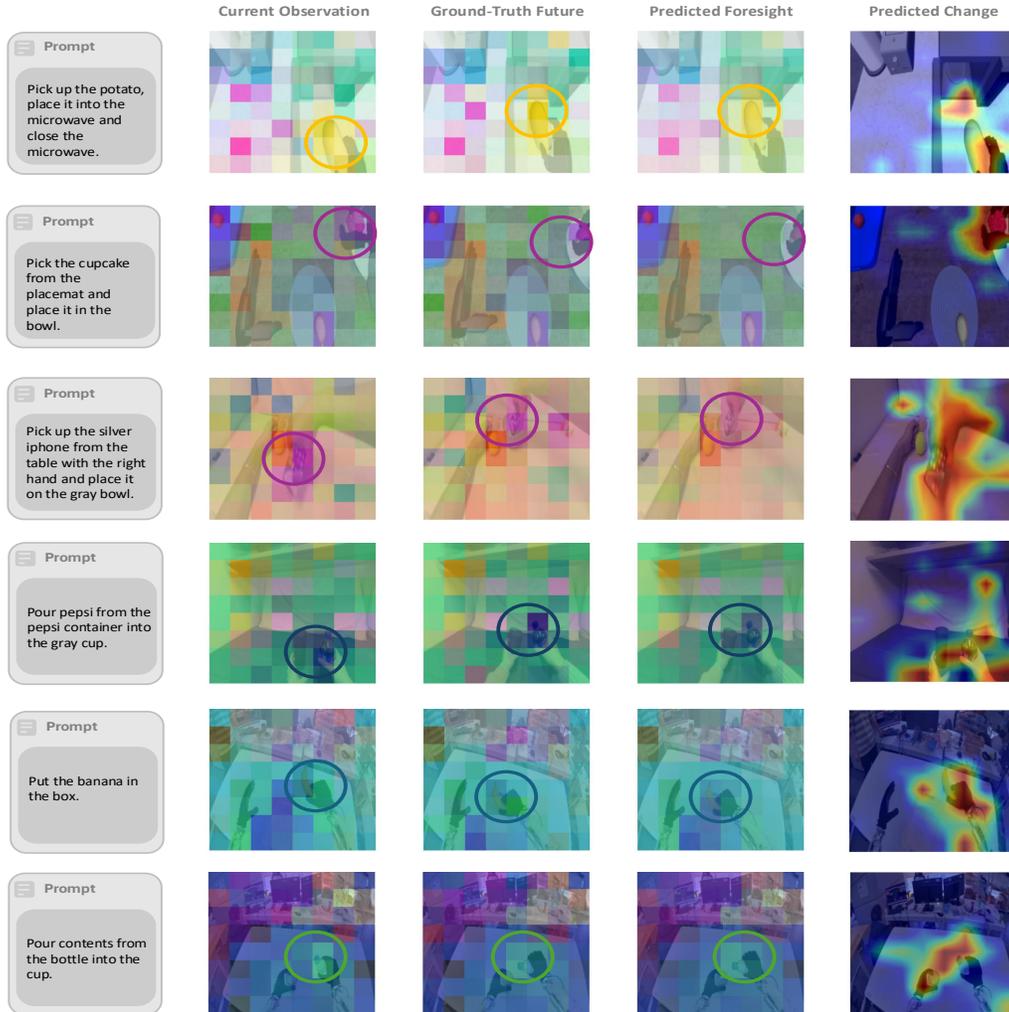
14

Figure 12: Visualization of latent representations for current observations, ground-truth futures, and predicted foresight, with colors encoding the first three PCA components mapped to RGB. The last column shows the per-patch cosine distance between predicted foresight and current observation features, highlighting regions where the model anticipates future change.

6 Conclusion and Discussion

In this paper, we introduce **DIAL**, a novel framework for end-to-end VLA models that achieves a structural decoupling of cognitive decision making and motor execution through a differentiable latent intent bottleneck. By framing the VLM as a predictive latent world model (**System-2**) and the controller as a latent inverse dynamics model (**System-1**), DIAL ensures that every motor command is strictly grounded in the model's intent expressed by the latent visual foresight. Our extensive evaluations on the RoboCasa GR1 Tabletop benchmark and real-world deployments on the IRON-R01-1.11 humanoid robot demonstrate that DIAL establishes a new state-of-the-art in performance, achieving $10\times$ higher data efficiency than existing methods and exhibiting robust zero-shot generalization across novel objects and complex configurations.

Looking forward, there are several key directions to further scale the DIAL framework. Currently, our System-1 employs a relatively small DiT backbone; scaling this to larger parameter sizes could significantly enhance the precision and multi-modal handling of complex motor tasks. Furthermore, while our current implementation keeps the VLM-native ViT frozen to maintain stable feature alignment, future work will explore end-to-end fine-tuning of this vision backbone, potentially stabilized

by an EMA-based encoding strategy and latent token compression to further boost performance and efficiency. Most importantly, since System-2 is designed for latent world modeling, DIAL is uniquely positioned to scale by consuming massive, in-the-wild human videos that lack action labels. Leveraging such action-free data to pre-train visual foresight will likely be the next frontier in building truly generalist embodied agents.

Ultimately, we envision a shift toward a more integrated yet modular paradigm for embodied intelligence. A compelling frontier is to incorporate latent world modeling directly into the native pre-training tasks of foundational VLMs, instilling actionable physical priors and a dynamics-oriented understanding within the backbone from the outset. This would ensure the VLM's representations are inherently aligned with the physical requirements of downstream control. Furthermore, the decoupling of DIAL suggests a highly efficient iteration strategy: once a System-1 action expert is pre-trained to master motor control, new or updated VLM generations can be seamlessly coupled and aligned to it, enabling the rapid transfer of the latest cognitive advances to robotic embodiments. By treating latent foresight as the universal interface between reasoning and execution, DIAL paves the way for a new generation of versatile and scalable generalist agents.

References

[1] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024.

[2] Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Tuomas Rintamaki, Tyler Poon, Max Ehrlich, Tuomas Rintamaki, Tyler Poon, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. Eagle 2.5: Boosting long-context post-training for frontier vision-language models, 2025.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

[6] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.

[7] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024.

[8] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

[9] Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, Adrian Li-Bell, Danny Driess, Lachy Groom, Sergey Levine, and Chelsea Finn. Hi robot: Open-ended instruction following with hierarchical vision-language-action models, 2025.

[10] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation, 2024.

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

[12] 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.

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

[14] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1702–1713, June 2025.

[15] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. *arXiv preprint arXiv:2412.15109*, 2024.

[16] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas

Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiao-qing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[18] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning, 2024.

[19] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control, 2023.

[20] Gemini Robotics Team, Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, Michael Bloesch, Konstantinos Bousmalis, Philemon Brakel, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Christine Chan, Oscar Chang, London Chappellet-Volpini, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, Adrian Collister, David B. D'Ambrosio, Sudeep Dasari, Todor Davchev, Meet Kirankumar Dave, Coline Devin, Norman Di Palo, Tianli Ding, Carl Doersch, Adil Dostmohamed, Yilun Du, Debidatta Dwibedi, Sathish Thoppay Egambaram, Michael Elabd, Tom Erez, Xiaolin Fang, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Ruiqi Gao, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Oliver Groth, Agrim Gupta, Roland Hafner, Steven Hansen, Leonard Hasenclever, Sam Haves, Nicolas Heess, Brandon Hernaez, Alex Hofer, Jasmine Hsu, Lu Huang, Sandy H. Huang, Atil Iscen, Mithun George Jacob, Deepali Jain, Sally Jesmonth, Abhishek Jindal, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Matija Kecman, J. Chase Kew, Donnie Kim, Frank Kim, Junkyung Kim, Thomas Kipf, Sean Kirmani, Ksenia Konyushkova, Li Yang Ku, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Tuan Anh Le, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Guy Lever, Jacky Liang, Li-Heng Lin, Fangchen Liu, Shangbang Long, Caden Lu, Sharath Maddineni, Anirudha Majumdar, Kevis-Kokitsi Maninis, Andrew Marmon, Sergio Martinez, Assaf Hurwitz Michaely, Niko Milonopoulos, Joss Moore, Robert Moreno, Michael Neunert, Francesco Nori, Joy Ortiz, Kenneth Oslund, Carolina Parada, Emilio Parisotto, Amaris Paryag, Acorn Pooley, Thomas Power, Alessio Quaglino, Haroon Qureshi, Rajkumar Vasudeva Raju, Helen Ran, Dushyant Rao, Kanishka Rao, Isaac Reid, David Rendleman, Krista Reymann, Miguel Rivas, Francesco Romano, Yulia Rubanova, Peter Pastor Sampedro, Pannag R Sanketi, Dhruv Shah, Mohit Sharma, Kathryn Shea, Mohit Shridhar, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Rachel Sterneck, Ian Storz, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Saran Tunyasuvunakool, Jake Varley, Grace Vesom, Giulia Vezzani, Maria Bauza Villalonga, Oriol Vinyals, René Wagner, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Chengda Wu, Markus Wulfmeier, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Jimmy Yan, Sherry Yang, Skye Yang, Yuxiang Yang, Hiu Hong Yu, Wenhao Yu, Wentao Yuan, Yuan Yuan, Jingwei Zhang, Tingnan Zhang, Zhiyuan Zhang, Allan Zhou, Guangyao Zhou, and Yuxiang Zhou. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer, 2025.

[21] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation, 2023.

[22] 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.

[23] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

[24] 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.

[25] Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z. Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, and Sergey Levine. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better, 2025.

[26] NVIDIA GEAR Team, Allison Azzolini, Johan Bjorck, Valts Blukis, et al. Gr00t n1.6: An improved open foundation model for generalist humanoid robots. `https://research.nvidia.com/labs/gear/gr00t-n1_6/`, December 2025.

[27] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

[28] Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, Hao Niu, Wenxuan Ou, Wanli Peng, Zeyu Ren, Haixin Shi, Jiawen Tian, Hongtao Wu, Xin Xiao, Yuyang Xiao, Jiafeng Xu, and Yichu Yang. Gr-3 technical report, 2025.

[29] Andy Zhai, Brae Liu, Bruno Fang, Chalse Cai, Ellie Ma, Ethan Yin, Hao Wang, Hugo Zhou, James Wang, Lights Shi, Lucy Liang, Make Wang, Qian Wang, Roy Gan, Ryan Yu, Shalfun Li, Starrick Liu, Sylas Chen, Vincent Chen, and Zach Xu. Igniting vlms toward the embodied space, 2025.

[30] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *8th Annual Conference on Robot Learning*, 2024.

[31] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, Winson Han, Wilbert Pumacay, Angelica Wu, Rose Hendrix, Karen Farley, Eli VanderBilt, Ali Farhadi, Dieter Fox, and Ranjay Krishna. Molmoact: Action reasoning models that can reason in space, 2025.

[32] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *The Twelfth International Conference on Learning Representations*, 2024.

[33] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation, 2024.

[34] Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026.

[35] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. Worldvla: Towards autoregressive action world model, 2025.

[36] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19752–19763, October 2025.

[38] Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, Jianyu Chen, and Jiang Bian. villa-x: Enhancing latent action modeling in vision-language-action models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[39] Jianke Zhang, Yucheng Hu, Yanjiang Guo, Xiaoyu Chen, Yichen Liu, Wenna Chen, Chaochao Lu, and Jianyu Chen. Unicod: Enhancing robot policy via unified continuous and discrete representation learning, 2025.

[40] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[41] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

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

[43] starVLA Contributors. Starvla: A lego-like codebase for vision-language-action model developing. GitHub repository, 1 2025.

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.