

From Reasoning to Action: Consequence-Grounded Alignment for Vision-Language-Action Models

Anonymous Author(s)

Affiliation

Address

email

Abstract:

Reasoning informs decision-making by connecting intended actions with their expected consequences. Building on this principle, recent vision-language-action models incorporate embodied chains of thought (ECoTs) as an intermediate stage in action generation. However, such guidance remains implicit, as actions are conditioned on the generated ECoTs rather than explicitly optimized to align with them. To bridge this gap, we introduce the Consequence-grounded Reasoning-to-Action alignment method called CoRAcT, an offline training framework that explicitly aligns reasoning and action through reward-guided optimization in action noise space. Specifically, CoRAcT formulates two physically grounded ECoT components as verifiable signals: movement specifies the intended motion, whereas waypoints encode the expected spatial consequence of execution. By jointly modeling ECoT reasoning, action generation, and action-conditioned future prediction, a unified Mixture-of-Transformers explicitly couples each action with its articulated intent and predicted consequence. For action candidates generated from sampled noise, CoRAcT computes alignment rewards based on motion consistency with the described movement and agreement between predicted outcomes and the specified waypoints. These rewards drive group-relative optimization of the initial noise distribution, steering action generation toward consistency with the articulated intent and expected consequence. Extensive experiments across RoboCasa, VLABench, and real-world manipulation tasks demonstrate consistent gains in task success and generalization over strong ECoT-based VLA baselines. Further analyses reveal improved reasoning–execution alignment and show that our method can ground externally provided ECoTs in executable behavior, supporting zero-shot adaptation to novel scenarios by leveraging knowledge from powerful VLMs.

1 INTRODUCTION

Human decision-making often involves prospective reasoning that connects intended actions with their expected consequences. Such reasoning mediates between task intent and concrete behavior by relating action formulation to anticipated outcomes. Inspired by this capability, recent vision-language-action (VLA) models increasingly integrate intermediate reasoning into action generation in the form of embodied chains of thought (ECoTs) [1, 2, 3, 4, 5]. By structuring action generation around textual plans, visual subgoals, spatial targets, and action-level intentions, ECoTs provide an intermediate interface between task understanding and physical execution [6, 7]. As the paradigm evolves, embodied reasoning is increasingly integrated into the perception–action pipeline, extending beyond purely textual deliberation toward representations more closely tied to physical control. However, closer integration of reasoning and control does not guarantee consistency between reasoning and execution. Existing VLA models typically use ECoTs as conditioning context for action

generation, leaving this correspondence to be learned implicitly from demonstrations rather than enforced through an explicit optimization objective [4, 5, 8, 9]. In embodied control, this gap has concrete consequences: a model may articulate an appropriate motion yet generate an action that deviates from it, or anticipate a plausible spatial outcome that execution fails to realize. These discrepancies reveal a missing link between coherent intermediate reasoning and physical behavior that faithfully reflects both the stated intent and the expected outcome.

To bridge this gap, we introduce CoRACT, an offline training framework for consequence-grounded reasoning-to-action alignment. CoRACT explicitly optimizes the consistency between embodied reasoning and generated actions through reward-based learning in the action noise space. The key idea is to treat the motion intent and spatial expectations expressed in embodied reasoning as **verifiable constraints for action generation**. These signals provide concrete criteria for evaluating whether generated actions follow the articulated reasoning and whether their predicted consequences agree with the expected outcomes. By translating these consistency evaluations into alignment rewards, CoRACT establishes a direct connection between embodied reasoning and action optimization, enabling reasoning-aware improvement of robotic policies.

Specifically, CoRACT extracts two physically grounded signals from ECoT reasoning: movement for describing the intended motion and waypoints for representing the expected spatial consequence. To connect these reasoning signals to action generation and future consequences, we build a unified Mixture-of-Transformers (MoT) architecture that jointly models ECoT reasoning, action generation, and action-conditioned future prediction. This joint formulation associates each generated action with both its articulated intent and predicted consequence, allowing the same action candidate to be evaluated against the reasoning that motivates it and the future outcome it is expected to produce. CoRACT therefore measures whether a generated action follows the described movement and whether its predicted consequence agrees with the specified waypoint, and converts these two forms of consistency into alignment rewards. Building on these alignment signals, CoRACT optimizes action generation in the noise space. For each reasoning context, multiple action candidates are generated from sampled noise and evaluated for motion and consequence consistency. The resulting rewards drive group-relative optimization of a conditional noise predictor, guiding it toward noise samples that induce actions better aligned with both the intended motion and expected consequence. This improves the fidelity between ECoT reasoning and generated actions, strengthening reasoning-to-action alignment.

We evaluate CoRACT across RoboCasa, VLABench, and real-world manipulation tasks, demonstrating consistent improvements in task success and generalization over strong ECoT-based VLA baselines. Further analyses show stronger alignment between reasoning and execution, with the benefits extending beyond reasoning generated by the policy itself. In particular, CoRACT can ground ECoTs provided by powerful VLMs in executable actions, allowing their knowledge and decision-making capabilities to guide manipulation in novel scenarios. This capability enables zero-shot adaptation through external reasoning without additional policy training, extending ECoTs from an intermediate reasoning representation to a general interface for guiding physical behavior.

Overall, our main contributions are as follows:

- We introduce CoRACT, a consequence-grounded alignment framework that explicitly enforces consistency between embodied reasoning and physical execution.
- Building on joint training, we derive alignment rewards from motion consistency with movement descriptions and agreement between predicted outcomes and waypoints. These rewards drive group-relative policy optimization in action noise space, steering actions toward the articulated intent and expected consequences
- Experiments in simulation and the real world demonstrate higher task success and stronger generalization than strong ECoT-based VLA baselines. Further experiments show that improved reasoning-action alignment enables zero-shot adaptation to novel scenarios

88 2 RELATED WORKS

89 **Embodied Chain-of-Thought Reasoning.** To improve generalization and interpretability in vision-
90 language-action policies, embodied chain-of-thought (ECoT) introduces intermediate reasoning
91 representations that connect semantic task understanding with grounded robot control. Early ap-
92 proaches instantiate reasoning in language: ECoT generates grounded textual traces [1], while
93 EMMA-X augments them with look-ahead spatial guidance [2]. Subsequent work broadens the rea-
94 soning substrate beyond text through representations including future visual subgoals [3], reinforced
95 visual planning [10], and coarse action trajectories [11]. Complementary work shifts reasoning from
96 explicit generation into latent representations: ECoT-Lite studies how reasoning supervision shapes
97 action-relevant representations [12], while LaRA-VLA [13] and LaST₀ [14] internalize multimodal
98 and spatio-temporal CoTs in continuous latent states. ERVLA shows that ECoT supervision is not
99 uniformly beneficial: high-level semantic traces alone yield limited gains, whereas action-grounded
100 signals provide stronger guidance for control [15]. Despite this progress, most methods induce cor-
101 respondence between reasoning and action implicitly through conditioning supervision, without di-
102 rectly verifying whether policy-generated actions comply with the model’s CoT. In contrast, CoRAct
103 turns generated ECoTs from conditioning context into verifiable supervision, explicitly training the
104 policy to translate articulated reasoning into physical execution.

105 **Consistency between Reasoning and Execution.** Consistency between articulated reasoning and
106 physical execution is essential for reliable embodied decision-making. Prior work formulates this
107 correspondence as explicit language–action alignment, using contrastive grounding and preference
108 learning to align generated descriptions with visual observations and action trajectories [16]. Be-
109 yond representation-level alignment, model-based approaches couple structured reasoning with low-
110 level planning or action-conditioned future prediction [17, 18, 19]. Runtime verification selects
111 among policy samples through simulated consequences or learned action verifiers [20, 21, 22, 23].
112 These deployment-time methods require repeated policy sampling and external evaluation, often
113 through an interactive simulator or a VLM-based verifier. To move this burden into training, re-
114 cent work employs verifiable rewards, task-success signals, or learned critics to post-train reasoning
115 VLAs [24, 25, 26, 27]. However, these objectives primarily assess textual reasoning, visual ground-
116 ing, high-level trajectories, or episode-level success, rather than directly constraining correspon-
117 dence between structured CoT fields and fine-grained action chunks. They do not jointly impose
118 two sample-level constraints central to CoRAct: whether an action follows the generated CoT and
119 whether its predicted consequence realizes the intended spatial outcome. CoRAct addresses both
120 links by deriving complementary rewards for action adherence and consequence realization from
121 each structured CoT, while distilling multi-sample verification into a lightweight noise policy that
122 requires one action sample at deployment.

123 3 Method

124 CoRAct follows a two-stage training framework. First, we jointly model ECoT reasoning, action
125 generation, and action-conditioned future prediction within a Mixture-of-Transformers (MoT) ar-
126 chitecture (Sec. 3.1). Second, we freeze the backbone and optimize a conditional noise policy with
127 verifiable alignment rewards to improve reasoning–execution consistency (Sec. 3.2). At inference,
128 the learned policy enables efficient single-sample action generation (Sec. 3.3).

129 3.1 Joint Modeling of Reasoning, Actions, and Consequences

130 Consequence-grounded alignment requires connecting actions to articulated intent and evaluating
131 whether their predicted consequences agree with the expected outcomes. We make these require-
132 ments explicit through two ECoT components, *movement* and *waypoints*. Both are constructed
133 through rule-based annotation of demonstrations: movement is derived from relative end-effector
134 poses and gripper actions, while waypoints are obtained by projecting the future gripper position
135 into the camera view.

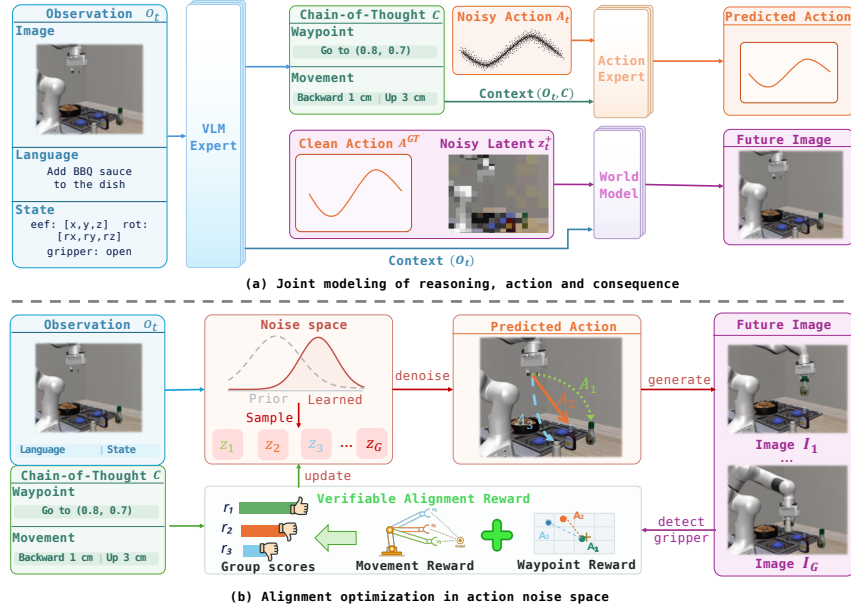


Figure 1: **Framework of CoRACT.** (a) CoRACT jointly models ECoT reasoning, action generation, and action-conditioned future prediction in a unified Mixture-of-Transformers, coupling actions with intent and consequence. (b) It freezes the backbone and optimizes the noise policy with movement and consequence rewards, steering action sampling toward better reasoning–execution alignment.

We jointly model ECoT reasoning, action generation, and future prediction within a Mixture-of-Transformers (MoT) backbone [28] comprising three Transformer experts. Let $o = (I, s, \ell)$ denote the multimodal input, consisting of visual observations I , robot state s , and language instruction ℓ . We denote the structured ECoT by C , an action chunk of horizon H by $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_H)$, and the subsequent visual observation by I^+ . The VLM expert processes the multimodal context and autoregressively predicts structured ECoT tokens, while the action and image generation experts model continuous action chunks and future visual latents through flow matching. This design maintains separate parameters for reasoning, action generation, and future prediction while allowing controlled interaction among them through attention mask design. The backbone is trained through two streams with distinct sample formats and conditioning structures, as follows:

Reasoning-conditioned action generation The first stream jointly models ECoT and action generation as

$$p_{\theta}(C, \mathbf{A} | o) = p_{\theta}(C | o) p_{\theta}(\mathbf{A} | o, C). \quad (1)$$

The VLM expert generates C autoregressively using a token-level cross-entropy objective. Conditioned on o and C , the action expert learns a conditional flow from action noise to the demonstrated action chunk.

Given a clean action chunk $\mathbf{A}^{gt}$, we sample $\epsilon_a \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t \sim \mathcal{U}(0, 1)$, and construct the noised action state $\mathbf{A}_t = \alpha_t \mathbf{A}^{gt} + \sigma_t \epsilon_a$. The action expert is optimized with

$$\mathcal{L}_{act} = \mathbb{E} \left[\left\| v_{\theta}^{act}(\mathbf{A}_t, t | o, C) - \mathbf{u}_t^{act} \right\|_2^2 \right], \quad (2)$$

where $\mathbf{u}_t^{act}$ denotes the target velocity induced by the chosen action flow path. Through the attention mask, action tokens attend to the multimodal context and ECoT prefix, with bidirectional attention within the action block. This stream establishes an implicit correspondence between articulated reasoning and generated actions, which is explicitly strengthened in the subsequent alignment stage.

Action-conditioned future prediction The second stream models the visual consequence of executing an action. We encode the subsequent observation into the latent space of the VAE used

159 by JanusFlow, $\mathbf{z}^+ = \mathcal{E}_{\text{VAE}}(I^+)$, and train the image generation expert to predict this future latent
 160 conditioned on o and the clean demonstrated action $\mathbf{A}^{\text{gt}}$.

161 Given visual noise $\epsilon_I \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we construct the noised future latent $\mathbf{z}_t^+ = \alpha_t \mathbf{z}^+ + \sigma_t \epsilon_I$. The
 162 image generation expert is optimized with

$$\mathcal{L}_{\text{world}} = \mathbb{E} \left[\left\| v_{\theta}^{\text{img}}(\mathbf{z}_t^+, t \mid o, \mathbf{A}^{\text{gt}}) - \mathbf{u}_t^{\text{img}} \right\|_2^2 \right], \quad (3)$$

163 where $\mathbf{u}_t^{\text{img}}$ denotes the target velocity.

164 ECoT tokens are excluded from this stream, ensuring that future prediction depends only on the
 165 current observation and clean demonstrated action, thereby preventing the image generation expert
 166 from accessing the articulated waypoint. At inference, integrating the learned flow produces a pre-
 167 dicted future latent $\hat{\mathbf{z}}^+$, which is decoded into a complete future image as $\hat{I}^+ = \mathcal{D}_{\text{VAE}}(\hat{\mathbf{z}}^+)$. The
 168 decoded prediction is subsequently used for waypoint-grounded consequence evaluation.

169 **Interleaved joint training** At each optimization step, we sample the reasoning–action stream with
 170 probability η and the action–world stream with probability $1 - \eta$. Let $\mathcal{D}_{\text{RA}}$ and $\mathcal{D}_{\text{AW}}$ denote the
 171 corresponding sample distributions. The overall objective is

$$\mathcal{L}_{\text{base}} = \eta \mathbb{E}_{\mathcal{D}_{\text{RA}}} [\mathcal{L}_{\text{CoT}} + \mathcal{L}_{\text{act}}] + (1 - \eta) \mathbb{E}_{\mathcal{D}_{\text{AW}}} [\mathcal{L}_{\text{world}}]. \quad (4)$$

172 Here, $\mathcal{L}_{\text{CoT}}$ is the autoregressive cross-entropy loss.

173 This interleaved training equips the MoT backbone with reasoning-conditioned action generation
 174 and action-conditioned consequence prediction. These capabilities are subsequently used to evaluate
 175 whether generated actions follow the described movement and whether their predicted consequences
 176 agree with the specified waypoints.

177 3.2 Alignment Optimization in Action Noise Space

178 The joint model in Sec. 3.1 establishes reasoning-conditioned action generation and action-
 179 conditioned consequence prediction. However, base training induces correspondence between rea-
 180 soning and execution only through conditional generation, without explicitly favoring action sam-
 181 ples that more faithfully realize the articulated intent and expected consequence. We therefore freeze
 182 the MoT backbone and use alignment rewards to optimize a lightweight conditional policy over
 183 action-flow initializations, thereby explicitly improving sample-level reasoning-to-action alignment.

184 **Conditional noise policy** During alignment training, each pair (o, C^{gt}) is encoded by the frozen
 185 MoT backbone into a context representation $\mathbf{h}$. We replace the context-independent Gaussian prior
 186 used in base training with a context-conditioned diagonal Gaussian policy [29, 30]:

$$q_{\phi}(\epsilon_a \mid \mathbf{h}) = \mathcal{N}(\mu_{\phi}(\mathbf{h}), \text{diag}(\sigma_{\phi}^2(\mathbf{h}))). \quad (5)$$

187 For each context, we sample $\{\epsilon_{a,i}\}_{i=1}^G$ from $q_{\phi_{\text{old}}}(\cdot \mid \mathbf{h})$. The frozen action expert then denoises
 188 each initialization for 10 steps to produce an action candidate $\mathbf{A}_i$.

189 **Movement alignment reward.** We first assess whether each action candidate follows the move-
 190 ment specified in C^{gt} . Let $\mathbf{m}^k$ denote the corresponding target for $k \in \mathcal{K} = \{p, r, g\}$, representing
 191 translation, rotation, and gripper command, and let $\mathcal{M}_k(\mathbf{A}_i)$ extract the corresponding endpoint
 192 motion from the action chunk. After expressing the predicted and target motions under a common
 193 coordinate convention, we compute the normalized error $e_{i,k} = \|\mathcal{M}_k(\mathbf{A}_i) - \mathbf{m}^k\|_2 / \tau_k$ and define
 194 $R_i^{\text{MOV}} = -\sum_{k \in \mathcal{K}} \lambda_k e_{i,k}$, where τ_k normalizes the scale of each motion component and λ_k con-
 195 trols its contribution. This reward directly measures whether the sampled action realizes the motion
 196 intent expressed in the provided ECoT.

Consequence alignment reward To verify the spatial consequence, we construct waypoint annotations by projecting demonstrated future end-effector positions into the corresponding camera images using calibrated extrinsics and intrinsics. These annotations are used to train a Grounding DINO-based waypoint localizer D_ψ , which predicts the projected gripper waypoint from an image.

For each action candidate A_i , the image generation expert predicts a future latent conditioned on (o, A_i) , which is decoded by the VAE decoder into a future image $\hat{I}_i^+$. The frozen waypoint localizer extracts $\hat{u}_i = D_\psi(\hat{I}_i^+)$. Let u^{ECoT} denote the waypoint specified in C^{gt} , expressed in the same image coordinates. We compute the waypoint error $d_{u,i} = \|\hat{u}_i - u^{\text{ECoT}}\|_2$ and define the consequence reward as $R_i^{\text{vis}} = \exp(-d_{u,i}/\tau_u)$, where τ_u controls sensitivity to prediction error.

The movement and consequence terms provide distinct signals for candidate-level alignment:

$$R_i = R_i^{\text{MOV}} + \lambda_{\text{vis}} R_i^{\text{vis}}. \quad (6)$$

The resulting reward favors actions that follow the movement specified in the ECoT while producing predicted consequences consistent with its waypoint.

Group-relative noise-policy optimization For candidates generated under the same observation and ECoT, the combined rewards provide a relative measure of alignment quality. We use the group mean as a critic-free baseline and define the advantage of candidate i as $\hat{A}_i = R_i - \frac{1}{G} \sum_{j=1}^G R_j$. We omit variance normalization to preserve the scale of reward differences and avoid amplifying noise when the within-group variance is small [31].

For each sampled initialization, we define the importance ratio $\rho_i = q_\phi(\epsilon_{a,i} | \mathbf{h}) / q_{\phi_{\text{old}}}(\epsilon_{a,i} | \mathbf{h})$ and its clipped counterpart $\bar{\rho}_i = \text{clip}(\rho_i, 1 - \varepsilon_{\text{clip}}, 1 + \varepsilon_{\text{clip}})$. We optimize the conditional noise policy by maximizing

$$\mathcal{J}(\phi) = \mathbb{E} \left[\frac{1}{G} \sum_{i=1}^G \min(\rho_i \hat{A}_i, \bar{\rho}_i \hat{A}_i) + \beta \mathcal{H}(q_\phi(\cdot | \mathbf{h})) \right]. \quad (7)$$

where expectation is taken over offline contexts and sampled candidate groups. The clipping threshold $\varepsilon_{\text{clip}}$ constrains the policy update, while β controls entropy regularization.

Only the conditional noise policy is updated, while the MoT backbone and waypoint localizer remain frozen. Alignment rewards reweight sampled action-flow initializations without backpropagating through action generation, future prediction, or waypoint localization, thereby favoring initializations that yield better-aligned actions.

3.3 Aligned Action Generation at Inference

At inference, the frozen VLM expert first autoregressively generates an ECoT C from the observation o , and the MoT backbone encodes (o, C) into the context representation $\mathbf{h}$. The learned noise policy transfers the reasoning–action alignment preferences acquired from candidate-level verification directly into the action sampling distribution. Specifically, we sample a single initialization $\epsilon_a \sim q_\phi(\cdot | \mathbf{h})$, which is denoised by the frozen action expert for 10 steps conditioned on (o, C) to produce an action chunk A . In this way, CoRAcT steers execution actions toward the articulated reasoning at the sampling stage to enable efficient deployment.

Table 1: In-distribution and generalization results on VLABench and RoboCasa. CoRAcT consistently outperforms prior methods across these settings, benefiting from strong VLM generalization and improved reasoning–action consistency.

Method	VLABench					RoboCasa Atomic				
	T1 ID	T2 Cat.	T3 C.S.	T4 Sem.	T6 Tex.	In-Dis.	OOD Obj.	OOD Tex.	OOD Light.	OOD Scen.
$\pi_{0.5}$	40.6	22.6	18.0	16.1	25.6	72.6	67.8	59.6	64.7	51.6
π_0	47.0	21.2	29.1	17.3	32.2	63.6	55.1	54.7	59.0	51.0
CoVer-VLA [20]	40.9	23.4	32.6	35.4	26.8	65.1	58.4	56.6	60.1	52.1
LaST ₀	40.9	21.1	34.4	35.0	26.6	69.0	63.7	63.7	66.5	55.5
CoRAcT	51.8	24.3	34.4	36.2	30.0	75.5	70.3	65.9	66.9	60.0

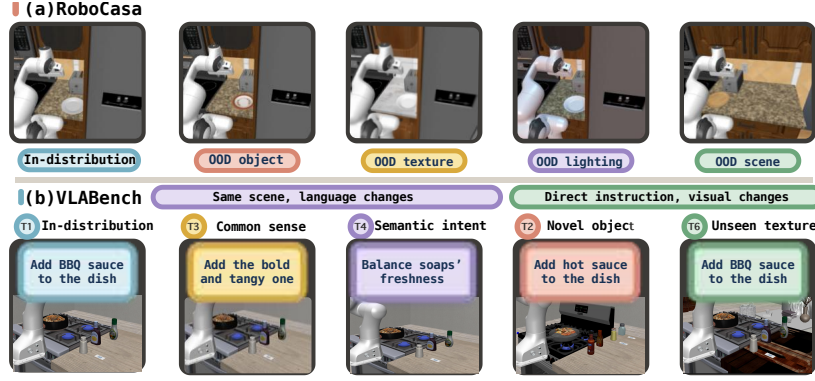


Figure 2: Generalization settings in RoboCasa and VLABench.

4 Experiments

We evaluate CoRACT from both task-level and mechanism-level perspectives. Experiments are organized to establish not only whether CoRACT improves robotic manipulation, but also whether the gains arise from stronger reasoning-action correspondence. In the experiments, we evaluate CoRACT from five perspectives: (1) in-distribution and out-of-distribution performance (Section 4.2), (2) contributions of individual components (Section 4.3), (3) fine-grained language following (Section 4.4), (4) closed-loop error recovery (Section 4.5), and (5) real-world reliability (Section 4.6).

4.1 Experimental Setup

We instantiate our policy from JanusFlow-1.3B [32], leveraging its unified multimodal understanding and generation capability for structured CoT reasoning and future prediction, and augment the backbone with an action-specialized Transformer expert following a Mixture-of-Transformers design. During training, the model autoregressively generates a structured numeric CoT containing translational motion, rotation, gripper state, and a future gripper waypoint. Conditioned on the generated CoT, the action expert predicts an 8-step chunk of 7-D Cartesian actions via conditional flow matching, while the world-model branch predicts a future third-person observation in the latent space of an SDXL VAE conditioned on the action chunk [33]. To prevent shortcut learning from the CoT, we mask CoT information from the world-model branch. For GRPO post-training, we freeze the base model and optimize only a 512-dimensional prefix-conditioned noise policy that controls the initialization of the frozen action flow. For each observation, we sample $G = 8$ candidate action chunks and decode each using 10 flow integration steps, with candidates ranked by CoT-action consistency and action-conditioned visual waypoint alignment. To avoid artificially small action deviations when applying GRPO to observations already seen during backbone training, we split each dataset into 90%/5%/5% subsets for backbone training, validation, and testing. We conduct GRPO post-training exclusively on the validation split and evaluate it on the held-out test split, ensuring a fair evaluation for subsequent quantitative analyses. For both RoboCasa and VLABench, we fine-tune the full backbone for 10 epochs in bfloat16 using AdamW with a learning rate of 2×10^{-4} , cosine decay, and weight decay of 0.01. GRPO uses task-balanced sampling and is trained for 2,500 updates with a learning rate of 10^{-5} . For evaluation, we test 17 RoboCasa tasks with 30 trials per task and 10 VLABench tasks with 50 trials per task, and report task success rate.

4.2 Comparison Results

CoRACT explicitly aligns structured embodied reasoning with downstream action generation, encouraging the policy to execute actions consistent with its inferred intent rather than merely producing plausible CoTs. We compare against representative VLA baselines, including π_0 , $\pi_{0.5}$, CoVer-VLA [20], which explicitly supervises CoT-action correspondence through contrastive learning, and LaST₀ [14], which uses latent CoT representations to implicitly guide action generation. We evaluate both in-distribution performance and generalization on RoboCasa [34] and VLABench [35]. As demonstrated in Figure 2(b).

As shown in Table 1, CoRAct achieves the strongest overall performance across both benchmarks and their generalization settings, demonstrating the benefit of explicitly enforcing reasoning–action consistency over direct action learning or implicit reasoning guidance. Notably, RoboCasa emphasizes fine-grained manipulation across diverse atomic skills, while VLABench places greater emphasis on common-sense understanding, semantic reasoning, and visual generalization. CoRAct improves performance across both regimes, indicating that consequence-grounded reasoning–action alignment benefits both precise skill execution and higher-level embodied reasoning.

4.3 Ablation and Quantitative Analysis

We ablate the ECoT components and alignment objectives to identify the sources of CoRAct’s gains. As shown in Table 2, Movement CoT improves success from 39.2% to 44.3% on VLABench and from 68.6% to 72.9% on RoboCasa, while adding Waypoint CoT further raises performance to 47.5% and 73.5%, respectively. Explicit movement–action alignment increases success to 49.3% and 74.6%, and incorporating waypoint–consequence alignment yields the full CoRAct model, achieving 51.8% and 75.4%. Overall, CoRAct improves upon the baseline without ECoT or alignment by 12.6 and 6.8 percentage points, demonstrating the complementary benefits of structured reasoning and explicit grounding in actions and predicted consequences.

Table 2: Ablation results of CoT components and alignment objectives.

CoT Components		Alignment Objectives		Success Rate (%) ↑	
Move. CoT	Waypt. CoT	Move. Align.	Waypt. Align.	VLABench	RoboCasa
				39.2%	68.6%
✓				44.3%	72.9%
✓	✓			47.5%	73.5%
✓	✓	✓		49.3%	74.6%
✓	✓	✓	✓	51.8%	75.5%

4.4 Fine-Grained Language Following

To evaluate reasoning–execution consistency, Table 3 reports four metrics: positional (**Pos.**) and rotational (**Rot.**) errors between ECoT-specified movements and generated actions, gripper-state agreement (**Grip.**), and image-space waypoint error (**WP.**). Combining both alignment objectives achieves the best performance across all metrics, reducing position error from 0.401 to 0.171 cm and waypoint error from 5.998 to 0.542 px, while improving gripper agreement from 69.8% to 91.8%. These results demonstrate that CoRAct effectively grounds embodied reasoning in executable actions.

Table 3: Ablation of movement and waypoint alignment on quantitative reasoning–execution consistency metrics.

Alignment		Consistency Metrics			
Move. Align.	Waypt. Align.	Pos. cm ↓	Rot. deg ↓	Grip. % ↑	WP. px ↓
		0.401	0.534	69.789	5.998
✓		0.209	0.478	87.661	2.904
	✓	0.232	0.494	85.160	2.706
✓	✓	0.171	0.420	91.800	0.542

4.5 Closed-Loop Self-Correction

Although CoRAct improves reasoning–action alignment and generalization, its training on expert trajectories provides limited exposure to failure recovery. Consequently, unexpected execution states may induce unreliable ECoTs that hinder subsequent correction. To address this limitation, we introduce an external agent for closed-loop ECoT correction without additional policy training. As illustrated in Fig. 3, GPT-5.6 Sol revises erroneous VLM-generated reasoning and feeds the corrected movement descriptions back to CoRAct for action generation. In this setting, only movement descriptions are used as agent-guided ECoTs. Results in Table 4 demonstrate that agent-guided correction substantially improves task success, with CoRAct achieving 66.0% and 52.0% on Insert Flower and Chemistry Tube, respectively. These findings highlight the complementary benefits of reasoning–action alignment and agentic closed-loop correction.

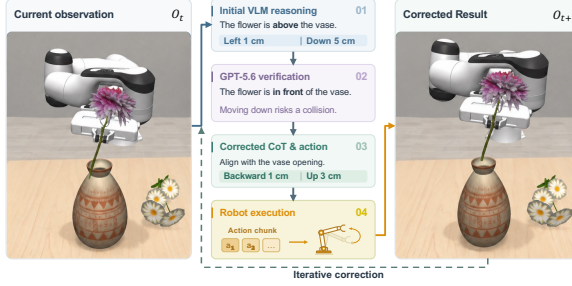


Figure 3: Agent-guided ECoT correction pipeline. GPT-5.6 Sol revises VLM-generated reasoning to guide CoAct’s action generation.

Table 4: Closed-loop self-correction on VLABench, demonstrating improved task success rates (%) through agent-guided CoT correction.

Components		Success Rate (%) ↑	
R–A Align.	Agent Corr.	Insert Flower	Chem. Tube
✓	✓	24.0	22.0
		26.0	30.0
		52.0	40.0
✓	✓	66.0	52.0



Figure 4: Real-world experimental setup.

4.6 Real-Robot Experiments

We evaluate three real-world tasks over 30 trials to assess long-horizon manipulation and semantic understanding (Figure 4). Tabletop organization requires removing an apple from a plate and placing juice on it. Image-guided word assembly involves selecting and arranging letters from a set of 26 to form the word indicated by a reference image. Language-guided fruit selection requires picking three specified fruit types from a bowl.

To accommodate master–slave teleoperation, we adapt ECoT to joint-space control: *movement* specifies per-joint motion, while *waypoint* represents the flange center’s projection in the top-camera image. Dedicated heads predict both components to reduce autoregressive generation latency. At inference, their outputs are assembled into a structured CoT using predefined rules to guide action generation. Table 5 shows improvements across all three tasks, demonstrating the effectiveness of our method in long-horizon and semantics-intensive real-world manipulation.

5 Conclusion

We present CoAct, a consequence-grounded framework that explicitly aligns embodied reasoning with physical execution. By leveraging movement and waypoint consistency rewards, CoAct optimizes a conditional noise policy through group-relative learning while keeping the backbone frozen. Experiments on RoboCasa, VLABench, and real robots demonstrate improved task success, generalization, and reasoning–execution consistency. Our results highlight ECoT as a verifiable interface for robotic action generation.

Table 5: Real-world evaluation on three long-horizon and semantic-intensive manipulation tasks.

Method	Tabletop organization	Image-guided word assembly	Language-guided fruit selection
$\pi_{0.5}$	33.3	26.6	56.6
LaST ₀	40.0	33.3	50.0
CoVer-VLA	33.3	26.6	50.0
CoAct	56.6	46.6	63.3

References

- [1] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine. Robotic control via embodied chain-of-thought reasoning. In P. Agrawal, O. Kroemer, and W. Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 3157–3181. PMLR, 06–09 Nov 2025.
- [2] Q. Sun, P. Hong, T. D. Pala, V. Toh, U.-X. Tan, D. Ghosal, and S. Poria. Emma-X: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14199–14214, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi:10.18653/v1/2025.acl-long.695.
- [3] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1702–1713. IEEE, 2025.
- [4] W. Huang, C. Chen, H. Qi, C. Lv, Y. Du, and H. Yang. Motvla: A vision-language-action model with unified fast-slow reasoning. *arXiv preprint arXiv:2510.18337*, 2025.
- [5] Q. Shou, F. Zhu, S. Chen, P. Yan, Z. Yan, Y. Miao, X. Pang, Z. Hong, R. Shi, H. Huang, et al. Halo: A unified vision-language-action model for embodied multimodal chain-of-thought reasoning. *arXiv preprint arXiv:2602.21157*, 2026.
- [6] X. Ma, L. Xing, H. Zhang, W. Li, and S. Lu. Unifying perception and action: A hybrid-modality pipeline with implicit visual chain-of-thought for robotic action generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22380–22390, 2026.
- [7] T. Lu, H. Zhang, Z. Diao, J. Wang, S. Xu, X. Lin, G. Zhong, Z. Ye, P. Wang, Z. Wu, et al. Thinkingvla: Interleaved vision and language reasoning for robotic manipulation. *arXiv preprint arXiv:2606.17937*, 2026.
- [8] D. Kim, S. Park, W. Song, S. Kim, T. Kim, H. Jang, J. Shin, J. Kim, and Y. Seo. Roboalign: Learning test-time reasoning for language-action alignment in vision-language-action models. *arXiv preprint arXiv:2603.21341*, 2026.
- [9] X. Sun, Z. Xu, C. Cao, Z. Liu, Y. Sun, J. Pang, R. Zhang, Z. Yang, K. Pang, D. He, et al. Atomvla: Scalable post-training for robotic manipulation via predictive latent world models. *arXiv preprint arXiv:2603.08519*, 2026.
- [10] C.-P. Huang, Y.-H. Wu, M.-H. Chen, F. Wang, and F.-E. Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *Advances in Neural Information Processing Systems*, 38:82782–82802, 2026.
- [11] L. Zhong, Y. Liu, Y. Wei, Z. Xiong, S. Liu, and G. Ren. Acot-vla: Action chain-of-thought for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8152–8162, 2026.
- [12] W. Chen, S. Belkhale, S. Mirchandani, K. Pertsch, D. Driess, O. Mees, and S. Levine. Training strategies for efficient embodied reasoning. In J. Lim, S. Song, and H.-W. Park, editors, *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 365–391. PMLR, 27–30 Sep 2025.
- [13] S. Bai, J. Lyu, W. Zhou, Z. Li, D. Wang, L. Xing, X. Zhao, P. Wang, Z. Wang, C. Chi, et al. Latent reasoning vla: Latent thinking and prediction for vision-language-action models. *arXiv preprint arXiv:2602.01166*, 2026.

- [14] Z. Liu, J. Liu, H. Chen, J. Yu, Z. Guo, C. Hou, C. Gu, X. Mi, R. Zhang, K. Wu, et al. Last $\{0\}$: Latent spatio-temporal chain-of-thought for robotic vision-language-action model. *arXiv preprint arXiv:2601.05248*, 2026.
- [15] N. Sun, Y. Zhang, Y. Yang, W. Zhao, P. Li, J. Guo, W. Song, P. Ding, R. Suo, Y. Su, et al. Revisiting embodied chain-of-thought for generalizable robot manipulation. *arXiv preprint arXiv:2606.03784*, 2026.
- [16] T. Wulff, F. Tavella, R. S. Maharjan, M. Adikari, and A. Cangelosi. Grounding hierarchical vision-language-action models through explicit language-action alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 9269–9281, June 2026.
- [17] Y. Ma, Y. Cao, W. Ding, S. Zhang, Y. Wang, B. Ivanovic, M. Jiang, M. Pavone, and C. Xiao. dvlm-ad: Enhance diffusion vision-language-model for driving via controllable reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1050–1061, 2026.
- [18] S. Tan, K. Chitta, Y. Chen, R. Tian, Y. You, Y. Wang, W. Luo, Y. Cao, P. Krahenbuhl, M. Pavone, et al. Latent chain-of-thought world modeling for end-to-end driving. *arXiv preprint arXiv:2512.10226*, 2025.
- [19] Y. Wu, R. Tian, G. Swamy, and A. Bajcsy. From foresight to forethought: Vlm-in-the-loop policy steering via latent alignment. *arXiv preprint arXiv:2502.01828*, 2025.
- [20] J. Kwok, X. Zhang, M. Xu, Y. Liu, A. Mirhoseini, C. Finn, and M. Pavone. Scaling verification can be more effective than scaling policy learning for vision-language-action alignment. *arXiv preprint arXiv:2602.12281*, 2026.
- [21] Y. Wu, A. Li, T. Hermans, F. Ramos, A. Bajcsy, and C. Pérez-D’Arpino. Do what you say: Steering vision-language-action models via runtime reasoning-action alignment verification. *arXiv preprint arXiv:2510.16281*, 2025.
- [22] J. Kwok, C. Agia, R. Sinha, M. Foutter, S. Li, I. Stoica, A. Mirhoseini, and M. Pavone. Robomongo: Scaling test-time sampling and verification for vision-language-action models. In J. Lim, S. Song, and H.-W. Park, editors, *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 3200–3217. PMLR, 27–30 Sep 2025.
- [23] M. Dai, L. Liu, Y. Bai, Y. Liu, Z. Wang, R. Su, C. Chen, L. Lin, and X. Wu. Rover: Robot reward model as test-time verifier for vision-language-action model. *arXiv preprint arXiv:2510.10975*, 2025.
- [24] A. Ye, Z. Zhang, B. Wang, X. Wang, D. Zhang, and Z. Zhu. Vla-r1: Enhancing reasoning in vision-language-action models. *arXiv preprint arXiv:2510.01623*, 2025.
- [25] C. Yin, Y. Lin, W. Xu, S. Tam, X. Zeng, Z. Liu, and Z. Yin. Deepthinkvla: Enhancing reasoning capability of vision-language-action models. *arXiv preprint arXiv:2511.15669*, 2025.
- [26] Y. Wang, W. Luo, J. Bai, Y. Cao, T. Che, K. Chen, Y. Chen, J. Diamond, Y. Ding, W. Ding, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025.
- [27] M. Foutter, M. Cercola, L. Wild, Y. Wang, M. Li, D. Gammelli, and M. Pavone. Do vision-language-action models mean what they say? on the role of faithfulness in embodied reasoning. *arXiv preprint arXiv:2607.04681*, 2026.
- [28] W. Liang, L. Yu, L. Luo, S. Iyer, N. Dong, C. Zhou, G. Ghosh, M. Lewis, W.-t. Yih, L. Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

- 435 [29] Y. Li, X. Ma, J. Xu, Y. Cui, Z. Cui, Z. Han, L. Huang, T. Kong, Y. Liu, H. Niu, et al.
436 Gr-rl: Going dexterous and precise for long-horizon robotic manipulation. *arXiv preprint*
437 *arXiv:2512.01801*, 2025.
- 438 [30] A. Wagenmaker, M. Nakamoto, Y. Zhang, S. Park, W. Yagoub, A. Nagabandi, A. Gupta, and
439 S. Levine. Steering your diffusion policy with latent space reinforcement learning. *arXiv*
440 *preprint arXiv:2506.15799*, 2025.
- 441 [31] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al.
442 Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv*
443 *preprint arXiv:2402.03300*, 2024.
- 444 [32] Y. Ma, X. Liu, X. Chen, W. Liu, C. Wu, Z. Wu, Z. Pan, Z. Xie, H. Zhang, X. Yu, et al. Janus-
445 flow: Harmonizing autoregression and rectified flow for unified multimodal understanding
446 and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*
447 *(CVPR)*, pages 7739–7751. IEEE, 2025.
- 448 [33] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rom-
449 bach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *Inter-*
450 *national Conference on Learning Representations*, volume 2024, pages 1862–1874, 2024.
- 451 [34] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu.
452 Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint*
453 *arXiv:2406.02523*, 2024.
- 454 [35] S. Zhang, Z. Xu, P. Liu, X. Yu, Y. Li, Q. Gao, Z. Fei, Z. Yin, Z. Wu, Y.-G. Jiang, et al.
455 Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with
456 long-horizon reasoning tasks. In *2025 IEEE/CVF International Conference on Computer Vi-*
457 *sion (ICCV)*, pages 11142–11152. IEEE, 2025.