AFTERIMAGE: OBJECT-GROUNDED ATTENTION FOR VISION–LANGUAGE–ACTION POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language–action (VLA) policies must retrieve the right visual evidence for control even as a scene's appearance changes. Most VLA training leaves this selection to action supervision, while recent approaches guide decoder attention with task-relevant labels. Where such guidance should intervene remains a fundamental question. We identify a cross-modal positional asymmetry in rotary position embeddings: broad positional interventions severely impair control, whereas changes confined to action-to-image attention can preserve task performance. This finding motivates AfterImage, powered by a Conditional Object-Region Attention (CORA) mechanism. AfterImage treats the action stage following image encoding as a place to relearn where to look. It separates the selection of image regions from the total attention allocated to images, so object guidance redistributes an existing visual budget. We turn this principle into an adaptation system that learns from region priors during training. Deployment requires no external perception model and incurs negligible parameter/computational overhead. Attention measurements and closed-loop transfer evaluations connect this intervention pathway to more task-directed visual retrieval and improved manipulation robustness.

1 INTRODUCTION

Vision–language–action (VLA) policies transform multi-view images, language instructions, and robot state into continuous action chunks. By coupling expressive vision–language backbones with flow-matching action experts, recent architectures have demonstrated impressive visuomotor manipulation capabilities (Brohan et al., 2023; Zitkovich et al., 2023; Kim et al., 2024; Black et al., 2024; Physical Intelligence, 2025). Yet, reliable control under real-world distribution shifts—such as altered camera angles, unfamiliar backgrounds, or visual distractors—requires more than scene understanding: the action generator must selectively retrieve task-relevant visual evidence while disregarding irrelevant clutter (Pumacay et al., 2024; Fei et al., 2025; Guo et al., 2026). While standard imitation and flow-matching objectives supervise *what* motions to execute, they leave this visual retrieval process entirely implicit (Chi et al., 2023; Lipman et al., 2023). Consequently, action tokens remain vulnerable to attention dispersion when scene appearances drift.

Supervising attention with object-centric regions provides an intuitive inductive bias to guide visual selection (Jia et al., 2026), but doing so in a pretrained VLA presents a fundamental challenge: *where can attention be steered without destabilizing the policy's learned representations or generative dynamics?* In modern unified architectures, cross-modal attention is tightly bound to rotary position embeddings (RoPE) (Su et al., 2021). Arbitrary fine-tuning risks corrupting these pretrained spatial coordinate frames, which are indispensable for continuous motor execution.

Frozen-policy ablations on LIBERO and DexJoCo (Wang et al., 2026) reveal a *cross-modal positional asymmetry*: visual self-attention is RoPE-sensitive, whereas low-frequency action-query-to-image-key ($\mathrm{A} \to \mathrm{I}$) interventions preserve task success. We therefore use this relation as the adaptation site.

Target-region supervision creates an *attention inflation* shortcut: the loss can decrease by raising all visual logits rather than improving spatial selection. We prevent this by decomposing attention into total visual mass and intra-image routing, then applying an exact Log-Sum-Exp (LSE) correction that preserves the former while learning the latter.

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

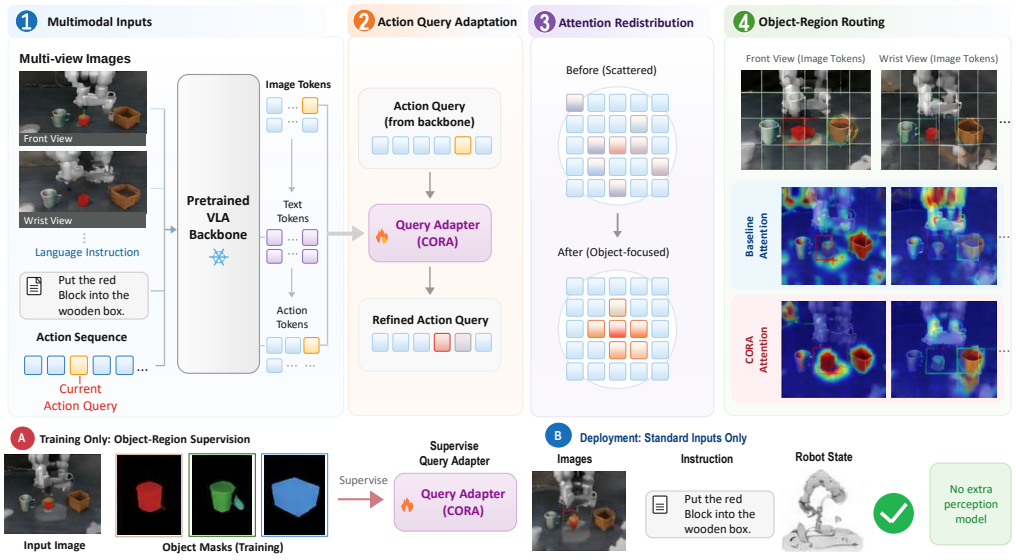


Figure 1: **Overview of CORA. Top:** A query adapter refines visual retrieval (1–2), redistributing attention toward task-relevant objects while strictly preserving the total visual attention budget via closed-form normalization (3–4). **Bottom: (A)** During training, object masks supervise the adapter with gradient isolation. **(B)** Deployment requires no external perception model and incurs negligible parameter/computational overhead.

We introduce **CORA** (Conditional Object-Region Attention), a zero-initialized query adapter confined to the low-frequency $\mathrm{A} \rightarrow \mathrm{I}$ subspace. An exact normalization preserves each row's visual attention mass, while a detached auxiliary path trains only the adapter from object-region labels. These labels are discarded at inference. Deployment requires no external perception model and incurs negligible parameter/computational overhead. Figure 1 illustrates the framework.

We validate our framework through extensive closed-loop evaluations spanning 10,030 multi-condition perturbation instances in LIBERO-Plus and 32,000 episodes in LIBERO-PRO across two generations of flow-based backbones (π_0 and $\pi_{0.5}$). Complementary Target Attention Ratio (TAR) measurements and qualitative difference maps confirm that consistent transfer improvements are causally accompanied by corrected visual routing toward task-relevant regions.

In summary, our contributions follow a progression from foundational discovery to mathematical and architectural innovation:

- We uncover an empirical cross-modal positional asymmetry in flow-based VLAs, showing that while visual self-attention is positionally rigid, $\mathrm{A} \rightarrow \mathrm{I}$ cross-attention possesses high geometric plasticity, thereby identifying an isolated, non-destructive manifold for surgical policy adaptation.

- We formally diagnose the pathology of attention inflation in auxiliary supervision and derive an exact, closed-form Log-Sum-Exp normalization that theoretically enforces local visual attention mass invariance, converting external object guidance into pure intra-scene attention redistribution.

- We design a gradient-isolated, dual-path query adapter that internalizes dense object priors during training. Deployment requires no external perception model and incurs negligible parameter/computational overhead. The adapter yields consistent zero-shot robustness gains and verified causal attention shifts across multiple policy generations.

2 Related Work

Vision–Language–Action Policies and Continuous Control. Vision–Language–Action (VLA) models connect web-scale multimodal pretraining with robot manipulation (Driess et al., 2023; Zitkovich et al., 2023; Open X-Embodiment Collaboration, 2024; Kim et al., 2024). Recent architectures pair expressive backbones with continuous action experts using action chunking, diffusion, or flow matching (Zhao et al., 2023; Chi et al., 2023; Black et al., 2024; Physical Intelligence, 2025; Kim et al., 2025). While recent extensions incorporate spatial tokens, trajectory structures, or visual futures (Qu et al., 2025; Zheng et al., 2025; Zhao et al., 2025), adaptation typically treats all token interactions uniformly. Our approach is orthogonal to these advances in action decoding: rather than modifying the action generation head or denoising objectives, we focus specifically on the cross-modal attention operation through which action queries retrieve visual evidence from the scene.

Distribution Robustness and Object-Centric Manipulation. Systematic evaluations show that manipulation policies remain highly vulnerable to visual and task distribution shifts, such as altered camera views, backgrounds, and clutter (Pumacay et al., 2024; Fei et al., 2025; Zhou et al., 2025; Guo et al., 2026). Incorporating object-centric priors provides a strong inductive bias to counter such shifts (Shridhar et al., 2022; Goyal et al., 2023). One prominent class of methods relies on explicit perception models at inference, conditioning policies on masks, affordances, or relational keypoints (Stone et al., 2023; Huang et al., 2024; 2025; Zhu et al., 2025; Vo et al., 2025), which introduces runtime latency and alters the deployment interface. Closest to our work, GuidedVLA (Jia et al., 2026) supervises specialized attention heads internally during training. However, standard auxiliary objectives supervise region attention without constraining the total visual budget. In contrast, CORA enforces an exact local conservation constraint, ensuring that training-time object guidance purely redistributes attention within an invariant visual budget, operating strictly mask-free at deployment.

Parameter-Efficient Adaptation and Positional Geometry. Parameter-efficient fine-tuning (PEFT), such as LoRA and lightweight adapters, specializes pretrained transformers with minimal trainable parameters (Hu et al., 2022; Chen et al., 2022). However, modern multimodal backbones represent relative spatial and temporal coordinates via Rotary Position Embeddings (RoPE) (Su et al., 2021; Heo et al., 2024; Wang et al., 2024). Arbitrary fine-tuning risks corrupting these rigid spatial coordinate frames, which are crucial for physical control. While prior studies observe functional frequency bands in language models (Barbero et al., 2024), we investigate the functional sensitivity of RoPE across heterogeneous cross-modal interactions in control. By uncovering an empirical cross-modal positional asymmetry, CORA avoids generic backbone updates and instead surgically restricts query adaptation to a safe, low-frequency rotary subspace.

3 Visual Routing in Flow-Based Policies

Cross-modal attention formulation. At control time t, given observation $o_t = (\mathcal{X}_t, \ell_t, s_t)$, the flow-matching action expert generates trajectory chunks by querying multimodal contextual tokens via cross-modal attention (Vaswani et al., 2017). For a given action query, let z_i denote the pre-softmax logit corresponding to key token i. The attention probability over the joint sequence is given by

$$P_i = \frac{\exp(z_i)}{Z_{\mathcal{I}} + Z_{\neg\mathcal{I}}}, \tag{1}$$

where $Z_{\mathcal{I}} = \sum_{j\in\mathcal{I}} \exp(z_j)$ and $Z_{\neg\mathcal{I}} = \sum_{k\notin\mathcal{I}} \exp(z_k)$. Here, $\mathcal{I}$ indexes visual patch tokens across camera views, while $\neg\mathcal{I}$ denotes all non-visual tokens (e.g., language instructions and proprioception).

Decoupling visual budget and spatial routing. To separate inter-modal weighting from spatial selection, we decompose the visual attention probability into two factors:

$$P_i = M_{\mathcal{I}} \cdot \rho_i, \qquad M_{\mathcal{I}} = \frac{Z_{\mathcal{I}}}{Z_{\mathcal{I}} + Z_{\neg\mathcal{I}}}, \qquad \rho_i = \frac{\exp(z_i)}{Z_{\mathcal{I}}}, \quad i \in \mathcal{I}. \tag{2}$$

The scalar $M_{\mathcal{I}} \in (0, 1)$ represents the **aggregate visual budget**—the total attention mass allocated to visual tokens relative to other modalities. In contrast, ρ_i represents **intra-scene routing**, describing how this visual budget is distributed across spatial image regions.

Attention inflation in auxiliary supervision. Given object-region targets $m_i \in [0, 1]$, an auxiliary routing objective aims to maximize target attention, typically formulated as

$$\mathcal{L}_{\text{aux}} = -\log \sum_{i \in \mathcal{I}} P_i m_i = -\log M_{\mathcal{I}} - \log \sum_{i \in \mathcal{I}} \rho_i m_i. \tag{3}$$

This decomposition reveals that minimizing $\mathcal{L}_{\text{aux}}$ can be achieved either by improving spatial allocation over visual tokens (ρ_i) or by universally increasing the total visual attention mass ($M_{\mathcal{I}}$). The latter constitutes an unintended shortcut: visual logits can be globally amplified relative to non-visual tokens, thereby disturbing the pretrained cross-modal balance. We term this phenomenon **attention inflation**.

To eliminate this shortcut, CORA enforces local visual budget invariance ($M'_{\mathcal{I}} = M_{\mathcal{I}}$), restricting region supervision to purely redistributing intra-scene routing rather than expanding the overall visual budget.

4 Conditional Object-Region Attention

To guide visual attention toward task-relevant objects without triggering attention inflation, we introduce CORA (**C**onditional **O**bject-**R**egion **A**ttention). As shown in Figure 1, CORA follows three straightforward steps: (1) *Action-conditioned query adaptation*, which adjusts attention logits using a bounded residual in a safe low-frequency subspace; (2) *Budget-preserving normalization*, a closed-form offset that guarantees the total visual attention mass remains strictly unchanged; and (3) *Dual-path training*, which learns visual routing from region masks without affecting the backbone representations. Deployment requires no external perception model and incurs negligible parameter/computational overhead.

4.1 Adapting Action Queries

Instead of altering the shared image keys or adding extra perception heads, CORA modifies how action tokens retrieve visual information by adding a small residual to the action query. Our diagnostic analysis (Section 5.5) shows that visual attention relies on high-frequency rotary dimensions for precise spatial coordinates, while lower frequencies primarily handle semantic retrieval. We therefore restrict query adaptation to a low-frequency band $\mathcal{B}$ consisting of K complete 2D rotary pairs ($2K$ dimensions).

For layer l, head h, and action query q, let $x_{lq} \in \mathbb{R}^D$ be the pre-attention hidden state. A lightweight projection matrix $W_{lh} \in \mathbb{R}^{D \times 2K}$ maps this state into the rotary band $\mathcal{B}$. To ensure training stability, we bound this residual relative to the base query norm:

$$\Delta \mathbf{q}_{lhq,\mathcal{B}} = \alpha \cdot \text{sg}\big(\|\mathbf{q}_{lhq,\mathcal{B}}\|_2\big) \cdot \frac{x_{lq} W_{lh}}{\sqrt{1 + \|x_{lq} W_{lh}\|_2^2}}, \tag{4}$$

where $\text{sg}(\cdot)$ stops gradients through the base query norm, and $\alpha \in (0, 1)$ limits the maximum relative change. The normalization term $v/\sqrt{1 + \|v\|_2^2}$ smoothly bounds the vector length to prevent runaway updates. We initialize all W_{lh} to zero, so the policy starts exactly with its pretrained behavior.

Next, we rotate $\Delta \mathbf{q}_{lhq,\mathcal{B}}$ with the base model's RoPE frequencies for band $\mathcal{B}$ (denoted by rotation matrix $R_{\Theta,\mathcal{B}}$). The logit increment d_{lhqi} for visual key $i \in \mathcal{I}$ is computed by taking the dot product with the corresponding rotated key dimensions:

$$d_{lhqi} = \frac{1}{\sqrt{d_k}} \left(R_{\Theta,\mathcal{B}}\, \Delta \mathbf{q}_{lhq,\mathcal{B}}\right)^\top \left(R_{\Theta,\mathcal{B}}\, \mathbf{k}_{lhi,\mathcal{B}}\right), \tag{5}$$

where d_k is the attention head dimension. Crucially, this logit increment is applied *only* to action-query-to-image-key pairs ($q \in \mathcal{Q}, i \in \mathcal{I}$). Visual self-attention and non-visual tokens remain untouched, ensuring that the policy's pretrained coordinate system is preserved.

4.2 Redistributing a Fixed Visual Budget

Simply adding d_i to image logits would increase the sum of exponentials over visual tokens, causing the model to allocate more total attention to images at the expense of text or state tokens (attention

inflation). To prevent this, we require the updated image logits $\{z'_i\}_{i\in\mathcal{I}}$ to preserve the original visual sum:

$$\sum_{i\in\mathcal{I}} \exp(z'_i) = \sum_{i\in\mathcal{I}} \exp(z_i) = Z_{\mathcal{I}}. \tag{6}$$

We achieve this by subtracting a shared constant offset c from all visual logits, setting $z'_i = z_i + d_i - c$ for $i \in \mathcal{I}$, while leaving non-visual logits unchanged ($z'_j = z_j$ for $j \notin \mathcal{I}$). Enforcing Eq. equation 6 yields:

$$\sum_{i\in\mathcal{I}} \exp(z_i + d_i - c) = \exp(-c) \sum_{i\in\mathcal{I}} \exp(z_i + d_i) = Z_{\mathcal{I}}. \tag{7}$$

Solving for c gives an exact, closed-form Log-Sum-Exp (LSE) correction:

$$c = \mathrm{LSE}_{i\in\mathcal{I}}(z_i + d_i) - \mathrm{LSE}_{i\in\mathcal{I}}(z_i). \tag{8}$$

Because the softmax denominator $Z_{\mathcal{I}} + Z_{\neg\mathcal{I}}$ is unchanged, the resulting attention probability becomes:

$$P'_i = M_{\mathcal{I}} \cdot \frac{\exp(z_i + d_i)}{\sum_{j\in\mathcal{I}} \exp(z_j + d_j)} = M_{\mathcal{I}} \cdot \rho'_i, \quad i \in \mathcal{I}, \tag{9}$$

while non-visual attention probabilities stay exactly the same ($P'_j = P_j, \forall j \notin \mathcal{I}$). In short, the base policy decides *how much* total attention to give the image ($M_{\mathcal{I}}$), and the adapter only learns *where* to redistribute that attention (ρ').

4.3 Learning from Object Regions

During training, object masks tell the model which visual tokens are task-relevant. Let $m_i \in [0, 1]$ be the target occupancy of image patch i. Instead of supervising each attention head individually—which could force every head to look at the same thing—we pool attention across the selected heads $\mathcal{H}$. The fraction of image attention landing on the target is:

$$r_{lq} = \frac{\sum_{h\in\mathcal{H}} \sum_{i\in\mathcal{I}} P'_{lhqi} m_i}{\sum_{h\in\mathcal{H}} M_{\mathcal{I},lhq}}. \tag{10}$$

Here, the denominator is the total visual attention across the chosen heads, which remains constant due to our budget conservation. Pooling before dividing allows different heads to specialize cooperatively on different parts of the scene.

Given a minibatch where a subset $\mathcal{V}$ contains visible targets, the routing loss averages the negative log-attention on target regions across supervised layers $\mathcal{S}$ and action queries $\mathcal{Q}$:

$$\mathcal{L}_{\text{route}} = -\frac{1}{|\mathcal{V}|\,|\mathcal{S}|\,|\mathcal{Q}|} \sum_{b\in\mathcal{V}} \sum_{l\in\mathcal{S}} \sum_{q\in\mathcal{Q}} \log r_{lq}^{(b)}. \tag{11}$$

The overall training loss combines the standard flow-matching action loss with the routing loss:

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_{\text{route}} \mathcal{L}_{\text{route}}, \tag{12}$$

where we set $\lambda_{\text{route}} = 0.05$ and supervise the first five action queries ($\mathcal{Q} = \{0, \ldots, 4\}$). If an image has no visible target, its routing loss is simply set to zero, while the flow loss continues to train action generation as usual.

4.4 Dual-Path Training and Lightweight Deployment

A practical risk of auxiliary supervision is that gradients from the region loss might disrupt the backbone features needed for smooth motor control. To prevent this, we decouple the gradient updates using a **dual-path training scheme**.

During the normal forward pass, the query residual updates the attention logits, directly influencing the predicted actions and receiving gradients from $\mathcal{L}_{\text{flow}}$. In parallel, to compute $\mathcal{L}_{\text{route}}$, we recompute the adapter logits using detached backbone features:

$$\bar{d}_{lhqi} = \mathrm{adapter}\big(\mathrm{sg}(x_{lq}), \mathrm{sg}(\mathbf{q}_{lhq,\mathcal{B}}), \mathrm{sg}(\mathbf{k}_{lhi,\mathcal{B}}); W_{lh}\big). \tag{13}$$

As a result, the model parameters θ (backbone, action expert, and projections) and adapter weights W receive clean, isolated gradients:

$$g_\theta = \nabla_\theta \mathcal{L}_{\text{flow}}, \qquad g_W = \nabla_W \mathcal{L}_{\text{flow}} + \lambda_{\text{route}} \nabla_W \mathcal{L}_{\text{route}}. \tag{14}$$

In other words, region supervision updates *only* the lightweight adapter W. The rest of the policy updates solely through the flow loss, adapting naturally to the improved visual routing.

Table 1: Zero-shot success rates (%) on LIBERO-Plus after fine-tuning on 1,722 LIBERO demonstrations. CORA improves overall success with both backbones. Stock and CORA use matched training settings and are evaluated at 20K steps with identical task-keyed action noise. Total is the micro-average over 10,030 perturbation instances, with one rollout each. Bold marks the higher rate within each backbone, and Δ denotes the difference in percentage points.

Backbone	Method	Camera	Robot	Language	Light	Background	Noise	Layout	Total
$\pi_{0.5}$	Stock baseline	52.66	**66.71**	**82.24**	**93.96**	91.26	68.58	**83.54**	75.44
$\pi_{0.5}$	CORA	**59.97**	66.39	81.78	**93.96**	**93.03**	**72.70**	82.69	**77.21**
	Δ (pp)	+7.32	-0.32	-0.46	+0.00	+1.77	+4.12	-0.85	**+1.76**
π_0	Stock baseline	11.01	7.68	23.49	42.12	28.90	14.12	33.05	21.71
π_0	**CORA**	**13.57**	**12.45**	**27.20**	**43.96**	**38.29**	**19.61**	**38.30**	**26.32**
	Δ (pp)	+2.56	+4.77	+3.71	+1.84	+9.39	+5.50	+5.25	**+4.61**

Deployment Cost. At deployment time, all region masks and loss computations are removed. The policy runs its normal action denoising steps, only adding the small query residual and the constant offset c. In total, CORA adds only six 1024×48 matrices across 3 layers and 2 heads—just $294{,}912$ parameters (about 1.18 MB). Deployment requires no external perception model and incurs negligible parameter/computational overhead.

5 Experiments

We evaluate CORA across two generations of flow-based VLAs (π_0 and $\pi_{0.5}$) to investigate three central questions: **(1)** Does budget-conserving attention routing improve zero-shot transfer under observation and task shifts? **(2)** Are closed-loop performance gains causally accompanied by corrected visual retrieval? **(3)** Why is query adaptation safe specifically within the low-frequency $A \to I$ rotary subspace?

5.1 Experimental Setup

Training and Baselines. We fine-tune both π_0 and $\pi_{0.5}$ on 1,722 LIBERO demonstration trajectories (40 tasks, 277k frames) for 20,000 steps (Liu et al., 2023). We compare CORA against the standard fine-tuned baseline (**Stock**), which uses identical demonstration sampling, action chunking, and optimization hyperparameters. CORA additionally uses stage-aware object masks (Jia et al., 2026) solely during training to supervise the query adapter.

Evaluation Benchmarks. We evaluate closed-loop zero-shot transfer without runtime masks or external perception models across two complementary benchmarks: (1) **LIBERO-Plus** (Fei et al., 2025), evaluating robustness against 7 categories of visual and environmental shifts across 10,030 individual rollouts; and (2) **LIBERO-PRO** (Zhou et al., 2025), testing generalization across 4 task-level variations across 32,000 episodes (40 conditions $\times$ 50 rollouts). Matched evaluations share identical initial robot/object poses and task-keyed action noise.

5.2 Robustness to Visual Perturbations (LIBERO-Plus)

Under visual distribution shifts, action queries in stock policies frequently disperse attention over irrelevant background clutter. As shown in Table 1, CORA effectively counters this dispersion, achieving consistent overall improvements across both policy generations: total success increases from 21.71% to 26.32% (+4.61 pp, +462 successful rollouts) for π_0, and from 75.44% to 77.21% (+1.76 pp, +177 successes) for $\pi_{0.5}$.

Crucially, the sharpest recoveries occur in visually demanding categories. On π_0, the largest gains are observed in Background (+9.39 pp), Noise (+5.50 pp), and Layout (+5.25 pp). On $\pi_{0.5}$, performance improves substantially under altered Camera perspectives (+7.32 pp) and sensor Noise (+4.12 pp). These shared benefits across direct appearance and viewpoint shifts validate our core hypothesis: enforcing visual budget conservation compels the policy to filter out visual distractors and retrieve task-critical visual evidence.

Table 2: Zero-shot success on LIBERO-PRO for π_0 and $\pi_{0.5}$ policies fine-tuned on LIBERO demonstrations. CORA improves Semantic and Object for both backbones. Each category contains 40 task conditions with 50 rollouts per condition. Cells show success rates (%) and successes/episodes; Overall pools 8,000 episodes per policy. Matched methods use the same initial states and task-keyed action noise. Bold marks the higher rate within each backbone, and Δ is the difference in percentage points.

Backbone	Method	Semantic	Object	Position	Task	Overall
$\pi_{0.5}$	Stock baseline	93.75 (1875/2000)	82.60 (1652/2000)	**26.55 (531/2000)**	**26.95 (539/2000)**	**57.46 (4597/8000)**
$\pi_{0.5}$	CORA	**95.40 (1908/2000)**	**84.55 (1691/2000)**	23.05 (461/2000)	24.80 (496/2000)	56.95 (4556/8000)
	Δ (pp)	+1.65	+1.95	-3.50	-2.15	-0.51
π_0	Stock baseline	36.15 (723/2000)	37.05 (741/2000)	2.70 (54/2000)	9.05 (181/2000)	21.24 (1699/8000)
π_0	CORA	**44.85 (897/2000)**	**42.55 (851/2000)**	**3.05 (61/2000)**	**12.35 (247/2000)**	**25.70 (2056/8000)**
	Δ (pp)	+8.70	+5.50	+0.35	+3.30	+4.46

Table 3: Target Attention Ratio on perturbed initial observations. Values are the mean fraction of image attention assigned to visible task-object regions (higher is better). Each column uses 80 observations from 40 task conditions and two initial states, shared across all four policies. Bold marks the higher ratio within each backbone.

	LIBERO-Plus			LIBERO-PRO	
Model	Camera	Background	Noise	Semantic	Object
Stock π_0	0.1128	0.1234	0.1152	0.1127	0.1188
CORA π_0	**0.1286**	**0.1367**	**0.1352**	**0.1304**	**0.1370**
Stock $\pi_{0.5}$	0.0423	0.0571	0.0361	0.0614	0.0705
CORA $\pi_{0.5}$	**0.0537**	**0.0736**	**0.0498**	**0.0767**	**0.0844**

5.3 Generalization Across Task and Semantic Shifts (LIBERO-PRO)

While LIBERO-Plus evaluates visual perturbations on familiar tasks, LIBERO-PRO tests policy adaptation to novel semantic phrasing, unfamiliar object instances, and modified workspace configurations (Table 2).

Semantic and Object Transfer. On semantic instruction rephrasings and novel object geometries, CORA consistently outperforms the stock baseline across both backbones. Specifically, π_0 achieves gains of +8.70 pp on Semantic and +5.50 pp on Object variations, while $\pi_{0.5}$ improves by +1.65 pp and +1.95 pp, respectively. This demonstrates that object-grounded visual routing helps action queries ground the intended interaction targets despite unfamiliar visual textures or language descriptions.

Spatial and Task Trade-offs. In contrast, on the highly capable $\pi_{0.5}$ model, we observe slight performance drops in Position (-3.50 pp) and Task (-2.15 pp). This divergence reveals an informative property of our inductive bias: our training objective supervises attention strictly on the *manipulated object*, rather than broader spatial anchors such as goal receptacles or placement surfaces. When large workspace repositioning is required, strictly concentrating attention on the object can divert visual capacity from destination goals. Extending region priors to encompass relational spatial targets offers a promising avenue to resolve this trade-off.

5.4 Mechanistic Verification: Target Attention Ratio

To verify that these closed-loop gains stem causally from corrected visual routing rather than incidental training dynamics, we evaluate the *Target Attention Ratio* (TAR)—the proportion of action-to-image attention allocated to visible target objects across 400 perturbed initial states (Table 3).

Across all five evaluated perturbation groups, CORA consistently increases the target attention ratio by 1.33–2.00 pp for π_0 and 1.14–1.65 pp for $\pi_{0.5}$. The spatial difference maps in Figure 2 provide clear qualitative evidence for π_0: while the stock baseline disperses attention across background table surfaces and robot arms, CORA cleanly refocuses probability mass onto the manipulated objects,

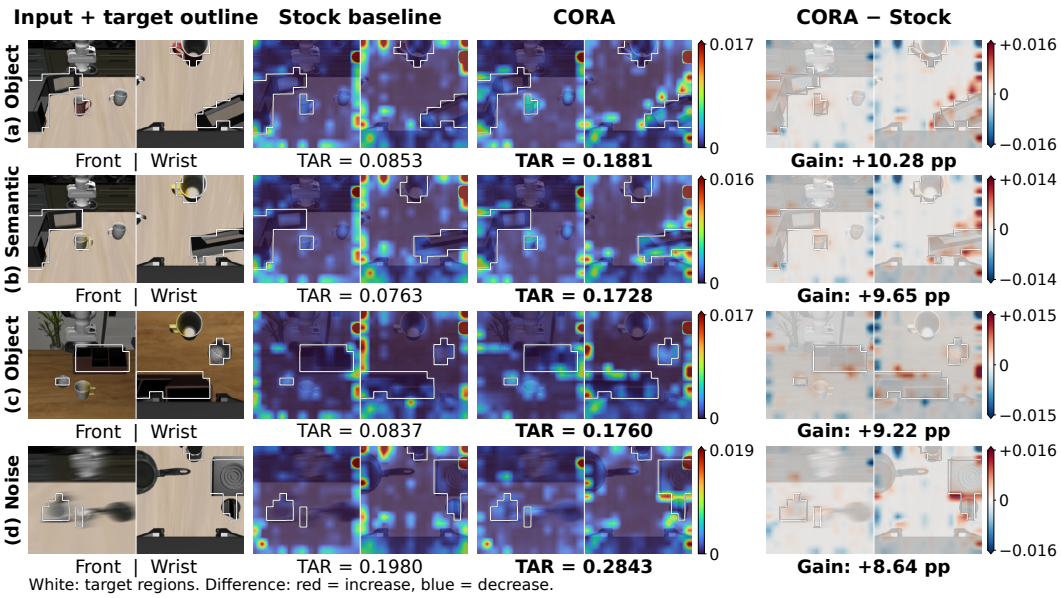


Figure 2: **CORA increases attention to task-relevant regions.** Attention maps for π_0 under Object, Semantic, and Noise perturbations. Each panel shows the front view (left) and wrist view (right) for the same observation. White outlines mark target regions. Stock and CORA share a color scale within each row, with warmer colors indicating higher attention. The final column shows CORA minus Stock: red denotes increased attention and blue denotes decreased attention. Target Attention Ratios (TAR) and gains use the original patch maps; Appendix E describes aggregation and display scales.

Table 4: Positional sensitivity of a frozen policy on LIBERO. Q/K identity removes the indicated RoPE rotations while keeping model weights fixed. Broad and image-wide interventions reduce success, whereas action-to-image (A→I) interventions remain close to Stock. Entries show success rates (%); Appendix Table 7 reports the corresponding successes/episodes. Each suite pools 10 tasks, three evaluation seeds, and 50 trials per task and seed; All Suites pools 6,000 episodes.

Condition	LIBERO-10	Goal	Object	Spatial	All Suites
Stock RoPE (baseline)	92.47%	97.67%	98.73%	98.80%	96.92%
All Q/K identity	21.27%	51.07%	49.93%	32.93%	38.80%
Image Q/K identity	28.07%	70.40%	65.27%	40.13%	50.97%
Action Q/K identity	88.27%	96.07%	96.00%	92.80%	93.28%
A→I Low-24 Q/K identity	91.20%	97.47%	98.93%	98.67%	96.57%
A→I High-24 Q/K identity	93.87%	97.73%	98.47%	98.07%	97.03%
A→I Full Q/K identity	92.53%	97.47%	98.47%	97.93%	96.60%

boosting local target attention by 8.64–10.28 pp. These measurements confirm that our training-time routing objective directly translates into corrected, object-directed retrieval at deployment.

5.5 Architectural Justification: Positional and Frequency Scope

Finally, we provide empirical justification for the design of CORA: *why can we adapt action queries without corrupting pretrained continuous control dynamics?* We diagnose this by ablating RoPE transformations on frozen policies without retraining (Tables 4 and 5).

Cross-Modal Positional Asymmetry (LIBERO). As shown in Table 4, disrupting RoPE across all tokens causes catastrophic performance collapse (success falls from 96.92% to 38.80%), and ablating visual self-attention drops it to 50.97%. In stark contrast, ablating RoPE exclusively within action-query-to-image-key ($A \rightarrow I$) attention preserves 96.60% success. This reveals an essential asymmetry: visual self-attention establishes a rigid coordinate frame indispensable for

Table 5: Action-to-image positional interventions on frozen DexJoCo policies. Full-128 removes all 128 rotary pairs; Low-24 removes the lowest 24. Full-band removal substantially reduces aggregate success, while Low-24 retains rates close to Stock. Each task–condition entry is a success rate (%) over three evaluation seeds with 50 episodes per seed.

	Rand Obj			Rand Full		
Task	Stock	Full-128	Low-24	Stock	Full-128	Low-24
Bimanual Assembly	3.33	0.00	3.33	0.67	0.00	0.00
Bimanual Hanoi	16.00	0.00	22.00	33.33	0.67	28.00
Bimanual Microwave Cook	56.00	7.33	54.00	70.67	12.00	71.33
Bimanual Photograph	54.00	0.00	58.00	21.33	0.00	24.00
Bimanual Unlock iPad	26.67	0.00	34.00	24.67	0.00	28.00
Fold Glasses	68.67	57.33	66.00	44.67	34.00	42.67
Hammer Nail	86.67	68.67	83.33	21.33	16.67	20.00
Click Mouse	58.67	58.67	68.00	52.67	52.67	52.67
Pick Bucket	88.67	80.67	85.33	73.33	30.67	70.00
Pinch Tongs	20.00	14.00	12.00	19.33	14.67	20.00
Water Plant	83.33	23.33	85.33	78.00	38.00	76.00
11-task aggregate	**51.09**	**28.18**	**51.94**	**40.00**	**18.12**	**39.33**

spatial reasoning, whereas $A \to I$ cross-attention functions primarily as a flexible semantic routing channel.

Frequency Selectivity in Dexterous Control (DexJoCo). We extend this diagnostic to 11 bimanual dexterous manipulation tasks in DexJoCo (Table 5). While removing all 128 rotary pairs in $A \to I$ attention severely degrades aggregate success from 51.09% to 28.18% under object randomization, restricting phase removal to the lowest 24 pairs fully preserves baseline performance (51.94% vs. 51.09%).

Summary. Together, these findings confirm that the low-frequency $A \to I$ rotary subspace provides an isolated, non-destructive manifold for adaptation, validating the bounded query architecture of CORA.

6 Discussion, Limitations, and Conclusion

CORA adapts low-frequency action-to-image queries while exactly preserving total visual attention mass. Across LIBERO-Plus and LIBERO-PRO, it improves visual-shift robustness and semantic/object transfer across both policy generations, while Target Attention Ratio measurements associate these gains with more task-directed visual retrieval. Deployment requires no external perception model and incurs negligible parameter/computational overhead.

Three limitations remain. First, object-only supervision can underweight target receptacles or placement surfaces during large spatial shifts, as reflected by the $\pi_{0.5}$ position results. Relational targets could address this gap. Second, our evaluations are limited to multi-camera simulation; physical deployment must tolerate noisy, open-vocabulary masks during training. Third, fixing the visual budget isolates spatial routing but prevents the policy from adapting how much visual evidence it uses. Future work should therefore evaluate noisy labels on hardware and jointly learn where and how much to attend.

AI use statement

Generative AI tools were used solely for grammar and language polishing of the manuscript. The authors developed all methodologies, conducted the experiments, and take full responsibility for the technical content and results.

Ethics statement

This work relies exclusively on publicly available datasets and open simulation benchmarks, involving no human participants or private data. As with general robotics research, deploying learned policies on physical platforms requires standard laboratory safety protocols.

Reproducibility statement

Appendix A derives the attention normalization. Appendices B and C describe the model configuration, optimization settings, and region-label processing. Appendix D gives the training algorithm, and Appendix E summarizes the evaluation protocols.

References

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*, 2024. URL https://arxiv.org/abs/2410.06205.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H. Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.025.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.026.

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. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488, 2023. URL https://proceedings.mlr.press/v202/driess23a.html.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025. URL https://arxiv.org/abs/2510.13626.

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. RVT: Robotic view transformer for 3d object manipulation. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 694–710, 2023. URL https://proceedings.mlr.press/v229/goyal23a.html.

Jianing Guo, Zhenhong Wu, Chang Tu, Yiyao Ma, Xiangqi Kong, Zhiqian Liu, Jiaming Ji, Shuning Zhang, Yuanpei Chen, Kai Chen, Qi Dou, Yaodong Yang, Xianglong Liu, Huijie Zhao, Weifeng Lv, and Simin Li. On robustness of vision-language-action model against multi-modal perturbations. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=cS6xizdYD5.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2403.13298.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Haifeng Huang, Xinyi Chen, Yilun Chen, Hao Li, Xiaoshen Han, Zehan Wang, Tai Wang, Jiangmiao Pang, and Zhou Zhao. RoboGround: Robotic manipulation with grounded vision-language priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2504.21530.

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. ReKep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024. URL https://arxiv.org/abs/2409.01652.

Xiaosong Jia, Bowen Yang, Zuhao Ge, Xian Nie, Yuchen Zhou, Cunxin Fan, Yufeng Li, Yilin Chai, Chao Jing, Zijian Liang, Qingwen Bu, Haidong Cao, Chao Wu, Qifeng Li, Zhenjie Yang, Chenhe Zhang, Hongyang Li, Zuxuan Wu, Junchi Yan, and Yu-Gang Jiang. GuidedVLA: Specifying task-relevant factors via plug-and-play action attention specialization. In *Robotics: Science and Systems*, 2026. URL https://arxiv.org/abs/2605.12369.

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. *arXiv preprint arXiv:2406.09246*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Robotics: Science and Systems*, 2025. URL https://arxiv.org/abs/2502.19645.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

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

Open X-Embodiment Collaboration. Open X-embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Physical Intelligence. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. THE COLOSSEUM: A benchmark for evaluating generalization for robotic manipulation. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.133.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Jiayuan Gu, Zhigang Wang, Yan Ding, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring spatial representations for vision-language-action models. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.011.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. CLIPort: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 894–906, 2022. URL https://proceedings.mlr.press/v164/shridhar22a.html.

Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Sean Kirmani, Brianna Zitkovich, Fei Xia, Chelsea Finn, and Karol Hausman. Open-world object manipulation using pre-trained vision-language models. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3397–3417, 2023. URL https://proceedings.mlr.press/v229/stone23a.html.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Khoa Vo, Taisei Hanyu, Yuki Ikebe, Trong Thang Pham, Nhat Chung, Minh Nhat Vu, Duy Nguyen Ho Minh, Anh Nguyen, Anthony Gunderman, Chase Rainwater, and Ngan Le. Clutter-resistant vision-language-action models through object-centric and geometry grounding. *arXiv preprint arXiv:2512.22519*, 2025. URL https://arxiv.org/abs/2512.22519.

Hanwen Wang, Weizhi Zhao, Xiangyu Wang, Siyuan Huang, He Lin, Boyuan Zheng, Rongtao Xu, Gang Wang, Yao Mu, He Wang, Lue Fan, Hongsheng Li, Zhaoxiang Zhang, and Tieniu Tan. DexJoCo: A benchmark and toolkit for task-oriented dexterous manipulation on MuJoCo. *arXiv preprint arXiv:2605.16257*, 2026. URL https://arxiv.org/abs/2605.16257.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. URL https://arxiv.org/abs/2409.12191.

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*, pp. 1702–1713, 2025.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.016.

Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. TraceVLA: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=b1CVu9l5GO.

Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. LIBERO-PRO: Towards robust and fair evaluation of vision-language-action models beyond memorization. *arXiv preprint arXiv:2510.03827*, 2025. URL https://arxiv.org/abs/2510.03827.

Minjie Zhu, Yichen Zhu, Jinming Li, Zhongyi Zhou, Junjie Wen, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. ObjectVLA: End-to-end open-world object manipulation without demonstration. *arXiv preprint arXiv:2502.19250*, 2025. URL https://arxiv.org/abs/2502.19250.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo,

Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183, 2023. URL `https://proceedings.mlr.press/v229/zitkovich23a.html`.

A Local attention conservation

For a fixed layer, head, action query, and hidden-state input, let $Z_{\mathcal{I}} = \sum_{i\in\mathcal{I}} \exp z_i$ be the original image partition and $\widetilde{Z}_{\mathcal{I}} = \sum_{i\in\mathcal{I}} \exp(z_i + d_i)$ the partition after the query update. Equation 8 sets $c = \log \widetilde{Z}_{\mathcal{I}} - \log Z_{\mathcal{I}}$. The corrected image partition is therefore

$$\sum_{i\in\mathcal{I}} \exp(z_i + d_i - c) = \exp(-c)\widetilde{Z}_{\mathcal{I}} = Z_{\mathcal{I}}. \tag{15}$$

Write $Z_{\neg\mathcal{I}}$ for the partition over non-image keys. Their logits are unchanged, so the full partition remains $Z_{\mathcal{I}} + Z_{\neg\mathcal{I}}$. For an image key,

$$\begin{aligned} P'_i &= \frac{\exp(z_i + d_i - c)}{Z_{\mathcal{I}} + Z_{\neg\mathcal{I}}} \\ &= \frac{Z_{\mathcal{I}}}{Z_{\mathcal{I}} + Z_{\neg\mathcal{I}}} \frac{\exp(z_i + d_i)}{\widetilde{Z}_{\mathcal{I}}} = M_{\mathcal{I}} \operatorname{softmax}_{\mathcal{I}}(z + d)_i. \end{aligned} \tag{16}$$

Thus, the update preserves the image budget and changes its distribution over image keys. Each non-image probability is also unchanged for this local operation. The transformer then propagates the updated attention output to subsequent layers.

B Model and optimization settings

Table 6 summarizes the shared CORA configuration. The selected heads use six bias-free matrices of shape 1024×48, adding $3 \times 2 \times 1024 \times 48 = 294{,}912$ parameters. Each matrix is initialized to zero and stored in FP32. The rotary dimensions are complete pairs from the original head representation, using the original frequencies.

Table 6: CORA configuration shared by π_0 and $\pi_{0.5}$. Backbone-specific normalization and action transforms follow the corresponding Stock policy.

Component	Value
Layers / heads	15, 16, 17 / 0, 1
Rotary pairs	104–127 (24 complete low-frequency pairs)
Max relative residual	0.25
Added parameters	294,912 FP32
Action horizon / supervised prefix	10 / 5
Route weight	0.05
Global batch / FSDP devices	16 / 2
Optimizer schedule	10K warmup, peak/final LR 5×10^{-5}
EMA	disabled

The $\pi_{0.5}$ suffix begins with action queries. In π_0, a continuous state token precedes the action queries, so the adapter uses an action-query offset of one. This offset aligns the same five supervised action queries across the two backbones.

Algorithm 1: One CORA training step

Input: minibatch (o, a, m); non-adapter parameters θ; adapter matrices W; selected layers $\mathcal{S}$, heads $\mathcal{H}$, rotary pairs $\mathcal{B}$, image keys $\mathcal{I}$; route weight λ_{route}.
Output: updated (θ, W).

1: Sample flow time τ and noise ϵ; construct the noisy action a_τ and encode o.
2: Set $\mathcal{L}_{\text{route}} \leftarrow 0$.
3: **for** transformer layer $l = 1, \ldots, L$ **do**
4: Compute the base Q_l, K_l, V_l and logits Z_l; set $\widetilde{Z}_l \leftarrow Z_l$.
5: **if** $l \in \mathcal{S}$ **then**
6: Compute $\Delta\mathbf{q}_{lhq,\mathcal{B}}$ for $h \in \mathcal{H}$ and each action query q using Equation 4.
7: Rotate $\Delta\mathbf{q}$ with the corresponding base RoPE frequencies and compute d_{lhqi} for $i \in \mathcal{I}$.
8: Replace only the selected action-to-image entries of $\widetilde{Z}_l$ using the conserving correction in Equation 8.
9: Recompute $\bar{d}_l$ from $\text{sg}(x_l, Q_l, K_l)$ and the live W_l; apply Equation 8 to $\text{sg}(Z_l)$ and $\bar{d}_l$.
10: Compute ℓ_l by averaging $-\log r_{lq}$ from Equation 10 over the first five action queries with visible targets.
11: $\mathcal{L}_{\text{route}} \leftarrow \mathcal{L}_{\text{route}} + \ell_l/|\mathcal{S}|$.
12: **end if**
13: Advance the transformer block using $\text{softmax}(\widetilde{Z}_l)V_l$.
14: **end for**
15: Compute $\mathcal{L}_{\text{flow}}$ from the predicted action flow and set $\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_{\text{route}}\mathcal{L}_{\text{route}}$.
16: Set $g_\theta \leftarrow \nabla_\theta \mathcal{L}_{\text{flow}}$ and $g_W \leftarrow \nabla_W \mathcal{L}$; update (θ, W) with the configured optimizer.

C Region targets and image processing

We use the stage-aware object-region annotations provided with the LIBERO demonstrations (Jia et al., 2026). The front and wrist views each provide a 16×16 array with values in $[0, 1]$. These targets are aligned with the vision-token grid and indicate the coverage of the task-relevant region at each patch. The third, padded camera slot has a zero target.

Spatial augmentation is disabled for both Stock and CORA to preserve the alignment between the annotated patch targets and image tokens. Color augmentation affects RGB images only. The label loader checks array shape, finiteness, and value range. A training example contributes to the region objective when it contains a visible target; all examples contribute to the action objective.

D Training algorithm

Algorithm 1 describes one training step. The forward and auxiliary paths share the adapter matrices $W = \{W_{lh}\}$. In the algorithm, W_l denotes the matrices at layer l, and sg denotes stop-gradient. The auxiliary path has the same attention probabilities as the forward path but detaches its base inputs. Its gradients therefore reach the current adapter directly.

During inference, each denoising step executes the forward intervention in Lines 3–8 and 12–14. Region labels and Lines 9–11 are omitted. Zero initialization gives $d = 0$ and $c = 0$, so the initial adapted policy equals the base policy.

E Evaluation details

LIBERO-Plus. The seven groups contain 1,599 Camera, 1,550 Robot, 1,537 Language, 1,142 Light, 1,076 Background, 1,601 Noise, and 1,525 Layout instances, for a total of 10,030. We use the official classification and compute the overall score from total successes. Each instance receives one rollout with evaluation seed 0. Matched methods receive identical task-keyed explicit action noise.

The π_0 comparison contains 2,178 Stock and 2,640 CORA successes. Its paired outcomes include 737 Stock-only and 1,199 CORA-only successes (two-sided exact McNemar $p = 6.46 \times 10^{-26}$). For $\pi_{0.5}$, the corresponding totals are 7,567 and 7,744, with 712 Stock-only and 889 CORA-only successes ($p = 1.07 \times 10^{-5}$). For π_0, six category differences remain significant under a seven-way

Bonferroni correction; Light has $p = 0.280$ before correction. For $\pi_{0.5}$, Camera and Noise remain significant under the same correction.

LIBERO-PRO. Each of the four evaluated perturbation categories contains 10 conditions from each of LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-10. The 50 rollouts per condition use initial-state indices 0–49 and evaluation seed 0. This yields 2,000 episodes per category, 8,000 per policy, and 32,000 across the four policies. Matched methods receive identical task-keyed explicit action noise.

Target Attention Ratio. Table 3 uses a shared set of 400 initial observations. For each of Camera, Background, and Noise, we sample 10 LIBERO-Plus task conditions per suite without replacement, using NumPy RNG seed $20260925 + s$ for suite index $s \in \{0, 1, 2, 3\}$. For Semantic and Object, we use all 40 LIBERO-PRO conditions per category. Each condition contributes stored initial states 0 and 1. Observations are captured after ten standard dummy actions and before the first policy action. All four policies receive identical images, states, instructions, and initial Gaussian action noise for each observation.

The diagnostic targets are task-level unions of visible simulator instances listed as objects of interest in the task definition; fixture regions are mapped to their segmented parent object. RGB images and masks share the front- and wrist-view orientation. Masks are resized to 224×224 with nearest-neighbor interpolation, then averaged over 14×14 pixel patches to obtain fractional occupancy on the 16×16 token grid. Both views contribute to the ratio, and the padded camera slot is excluded. All 400 observations contain visible targets. These masks are used solely for measurement.

We pool shared-softmax probability mass across heads 0 and 1 before computing Equation 10, using each policy's own attention probabilities. We then average the ratios over layers 15–17, action queries 0–4, all ten denoising steps, and the 80 observations in each category. Layer and head indices are zero-based. This uses the same attention relations as the region objective while measuring attention to task objects in the transfer environments.

Attention visualization. Each spatial map in Figure 2 uses the same head pooling and layer, query, and denoising-step averaging as the ratio. Maps sum to one jointly over the two camera views, and the displayed TAR is their inner product with the fractional target mask. For display, we bilinearly interpolate each 16×16 camera grid separately. Stock and CORA share a linear color scale capped at the joint 99th percentile within each row. Difference maps subtract the original Stock map from the CORA map and use a symmetric linear scale capped at the 99th percentile of the absolute differences. Colorbar triangles indicate saturation. Colors have opacity 0.72 over partially desaturated images; white contours mark target occupancy 0.5. Ratios and gains are computed before interpolation or color clipping.

Frozen-policy diagnostics. The LIBERO positional experiment evaluates each condition over three evaluation seeds, 10 tasks per suite, and 50 trials per task and seed. Each suite contains 1,500 episodes and the four-suite aggregate contains 6,000. The DexJoCo experiment uses three evaluation seeds and 50 episodes per task and seed for each of its 11 tasks under both `Rand Obj` and `Rand Full`. These experiments change attention at inference while keeping the policy weights fixed. Table 7 provides the complete LIBERO counts underlying Table 4.

Table 7: Full LIBERO positional-sensitivity results corresponding to Table 4. Q/K identity removes the indicated RoPE rotations while keeping model weights fixed. Each entry reports the success rate with successes/episodes in parentheses. Each suite pools 10 tasks, three evaluation seeds, and 50 trials per task and seed; All Suites pools 6,000 episodes.

Condition	LIBERO-10	Goal	Object	Spatial	All Suites
Stock RoPE (baseline)	92.47% (1387/1500)	97.67% (1465/1500)	98.73% (1481/1500)	98.80% (1482/1500)	96.92% (5815/6000)
All Q/K identity	21.27% (319/1500)	51.07% (766/1500)	49.93% (749/1500)	32.93% (494/1500)	38.80% (2328/6000)
Image Q/K identity	28.07% (421/1500)	70.40% (1056/1500)	65.27% (979/1500)	40.13% (602/1500)	50.97% (3058/6000)
Action Q/K identity	88.27% (1324/1500)	96.07% (1441/1500)	96.00% (1440/1500)	92.80% (1392/1500)	93.28% (5597/6000)
A→I Low-24 Q/K identity	91.20% (1368/1500)	97.47% (1462/1500)	98.93% (1484/1500)	98.67% (1480/1500)	96.57% (5794/6000)
A→I High-24 Q/K identity	93.87% (1408/1500)	97.73% (1466/1500)	98.47% (1477/1500)	98.07% (1471/1500)	97.03% (5822/6000)
A→I Full Q/K identity	92.53% (1388/1500)	97.47% (1462/1500)	98.47% (1477/1500)	97.93% (1469/1500)	96.60% (5796/6000)