WHEN VISUAL COMPRESSION BREAKS CONTROL: EFFICIENT MULTI-VIEW VLA INFERENCE VIA RETAINING EVIDENCE FOR CONTROL-ALIGNED POLICY RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Aggressive visual-token compression is increasingly important for efficient Vision-Language-Action (VLA) inference. However, we find that compression does more than discard visual information: it also induces a representation shift that can substantially misalign a policy trained on dense visual inputs. We therefore formulate efficient multi-view VLA inference as two coupled problems—preserving action-relevant physical evidence and recovering policy behavior under compression-induced representation shift. To this end, we propose RECAP-VLA, short for Retaining Evidence for Control-Aligned Policy Recovery, as a fixed-budget compression-and-recovery framework. RECAP-VLA organizes cross-view tokens into geometry-grounded physical-evidence components and allocates the compact token budget according to action relevance while preserving global context and residual evidence. It further introduces Compression Amplification Profiling (CAP) to localize policy regions most associated with compression-induced control error, and Control-Aligned Counterfactual Policy Adaptation (CAPA) to recover dense-policy behavior using lightweight same-state dense–compact supervision. At a fixed 512-to-256 token operating point, direct compression reduces success from 91% to 46%, whereas RECAP-VLA restores it to 92%. On 200 matched LIBERO Spatial episodes, RECAP-VLA achieves 91.5% success versus 90.5% for the dense policy while reducing median end-to-end latency by 8.44%. The same recipe remains within 2.5 percentage points of dense performance across additional OpenVLA-family checkpoints and LIBERO tasks, with approximately 8.3% median-latency reduction.

1 INTRODUCTION

Vision-Language-Action (VLA) policies map visual observations and language instructions to robot actions (Brohan et al., 2023; Zitkovich et al., 2023; Kim et al., 2025b; Black et al., 2025). Their repeated visual processing at each closed-loop query is costly, particularly when both external and wrist cameras contribute tokens (Kim et al., 2025b;a). The two views provide complementary workspace and contact information, yet can also show the same physical surface at different scales. Independent 2D saliency cannot reliably distinguish cross-view correspondence from visual similarity between unrelated locations. A compact policy must retain the physical evidence needed for control. Token reduction also changes the input distribution of a policy trained on dense representations. We therefore formulate fixed-budget VLA compression as a control-preserving problem with two coupled error sources: action-relevant evidence loss and compression-induced policy shift. At a 512-to-256 operating point, Dense-512 achieves 91% success, direct Hard-256 selection achieves 46%, and physical-evidence compaction alone reaches 52%. A selector that preserves evidence but leaves the policy shift untreated remains fragile, while recovery cannot reconstruct evidence that has already been discarded. The two operations are therefore representation-coupled and order-sensitive under the tested deployment protocol: their effects must be evaluated at the same deployment budget. We introduce RECAP-VLA (Figure 1). Calibrated metric geometry and feature compatibility associate four-token nodes across cameras into physical-evidence components. A deterministic 64-token scaffold retains global context; intervention-trained component utility and saturated multi-cover allocate the remaining slots; bounded residual transfer incorporates discarded information into selected non-scaffold anchors without changing their identity or order. The resulting sequence contains 256 original token positions.

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.

To locate the downstream effect of compression, Compression Amplification Profiling (CAP) relates paired dense–compact hidden-state discrepancies to same-state control deviation. The association concentrates in late policy blocks, although hidden-state error does not necessarily increase with depth. Guided by this diagnostic, Control-Aligned Counterfactual Policy Adaptation (CAPA) updates only the q- and v-projections in blocks 30–31 using rank-8 LoRA (Hu et al., 2022). Dense and compact branches share the observation, instruction, proprioception, checkpoint, and decoder, isolating the visual representation as the controlled difference. Training directly minimizes the normalized control metric used for evaluation; the visual encoder, earlier blocks, and action head stay frozen. The dense branch is used during adaptation only.

Objective choice is consequential. In a same-state diagnostic, SmoothL1 falls from 0.345199 to 0.248111 while evaluation deviation rises from 0.631096 to 0.717893; the mean action-space gradient cosine is only 0.079530. Using the evaluation metric instead reduces diagnostic deviation from 0.631028 to 0.498376 (21.02%) and improves formal closed-loop success from 77% to 92% with the same adapter architecture.

The full method reaches 92% success on the 100-episode ablation, versus 91% for Dense-512, 46% for Hard-256, and 84% for Hard-256 with the same control-aligned recovery. On 200 matched LIBERO Spatial episodes (Liu et al., 2023), success is 91.5% versus 90.5% for Dense-512, and median end-to-end query latency falls from 95.927 to 87.835 ms (8.44%) on an RTX 3090. Across OpenVLA-OFT/LIBERO Spatial and LightVLA/LIBERO Goal, the compact method remains within 2.5 percentage points of each dense counterpart with 8.31–8.34% median-latency reductions. These observed success differences are descriptive, not evidence that compression improves the underlying policy.

Our contributions are threefold: (i) we formulate fixed-budget VLA token reduction as a control-preserving compression problem that jointly accounts for evidence loss and policy shift; (ii) we instantiate this formulation with cross-view physical-evidence retention and CAP, a compression-to-control localization diagnostic; and (iii) we introduce metric-aligned same-state CAPA as a recovery operator restricted to the localized policy subspace. The claim is a coupled operating-point principle, rather than isolated novelty for any one submodule: the controlled chain is 46% (Hard-256) $\rightarrow$ 52% (evidence only) $\rightarrow$ 84% (recovery only) $\rightarrow$ 92% (complete). Controlled removals reach 87% without predicted utility and 88% without residual transfer, while deployment uses a fixed 256-token path without a dense teacher.

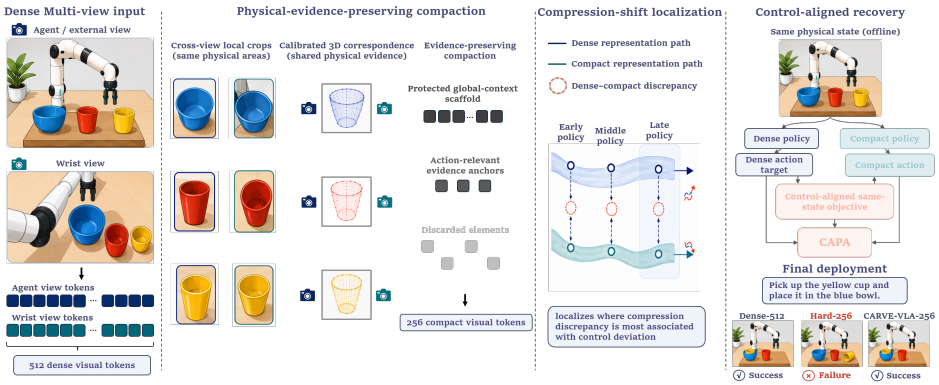

Figure 1: **Overview of RECAP-VLA.** Dense two-view visual tokens are reorganized into geometry-grounded physical-evidence components and compacted from 512 to 256 tokens using a protected global-context scaffold, utility-weighted saturated multi-cover, and anchor-preserving residual aggregation. Offline CAP then localizes the late policy region in which dense–compact representation discrepancies are most strongly associated with downstream control deviation. CAPA adapts only the q- and v-projections of blocks 30–31 from same-state dense/compact supervision, while deployment executes a single fixed compact path with a frozen action head.

2 Related Work

Vision-language-action policies. VLA research has progressed from large-scale visuomotor transformers to generalist policies that connect language, visual observations, and executable actions. RT-1 established broad real-world transformer control (Brohan et al., 2023); PaLM-E and RT-2

connected embodied observations with language-model representations and action tokens (Driess et al., 2023; Zitkovich et al., 2023). Open X-Embodiment and Octo extended this direction through cross-robot data and open generalist policies (O'Neill et al., 2024; Ghosh et al., 2024). OpenVLA provides an open 7B-parameter VLA (Kim et al., 2025b), while OpenVLA-OFT and LightVLA study continuous actions, parallel decoding, and visual-token pruning (Kim et al., 2025a; Jiang et al., 2025); π_0 represents a complementary flow-matching formulation (Black et al., 2025). Action chunking and diffusion policies further show that the action representation affects closed-loop efficiency and multimodality (Zhao et al., 2023; Chi et al., 2023). Our work keeps the policy checkpoint and action decoder fixed and studies how to make a two-view input cheaper without changing the task or control interface.

Visual-token reduction for efficient inference. Vision-transformer compression methods reduce computation by selecting tokens, merging similar features, sampling salient patches, or adapting computation depth (Dosovitskiy et al., 2021; Ryoo et al., 2021; Rao et al., 2021; Liang et al., 2022; Yin et al., 2022; Fayyaz et al., 2022; Bolya et al., 2023). These approaches establish that visual redundancy can be removed, but their importance signals are generally image- or feature-centric. VLA-specific methods add action relevance, temporal context, speculative pruning, or task-conditioned selection: EfficientVLA and LightVLA learn task-aware token reduction, while ADP, SpecPrune-VLA, VLA-Pruner, and VLA-IAP exploit action dynamics, temporal context, or interaction geometry (Yang et al., 2025; Jiang et al., 2025; Pei et al., 2025; Wang et al., 2026; Liu et al., 2025; Cheng et al., 2026). These methods motivate our fixed-budget setting, but do not jointly address cross-view physical correspondence and the policy shift caused by replacing dense visual input.

Multi-view and geometry-aware representations. PerAct and RVT demonstrate that 3D structure and view aggregation can improve manipulation (Shridhar et al., 2023; Goyal et al., 2023). Recent efficient-VLA work uses multi-view importance, differentiable resampling, or depth-guided merging: BFA++ models intra- and inter-view importance, GridS performs continuous grid-sample pruning, and DepthCache uses depth and end-effector motion to guide training-free merging (Li et al., 2026a; Feng et al., 2026; Li et al., 2026b). Seeing Once is Enough? removes overlapping 3D evidence for visual question answering, whereas MVUCF uses same-world-point consistency during VLA training (Lai et al., 2026; Yang et al., 2026). Our geometry is used online to associate tokens that may describe the same local physical evidence, while the policy still receives original token features. The current implementation assumes simulator metric depth and calibrated intrinsics/extrinsics; comparisons with RGB-only methods therefore describe scope rather than a common sensing leaderboard. In this paper, physical evidence is an operational planning concept: a component is treated as shared evidence when its nodes are spatially compatible under calibration and feature-compatible. We do not claim that these components are semantic object or part labels.

Recovery under representation shift. Adapters and LoRA provide low-parameter updates while retaining a frozen transformer (Houlsby et al., 2019; Hu et al., 2022); RLRC combines structured VLA compression with subsequent recovery (Chen et al., 2025). These works motivate parameter-efficient adaptation, but our recovery target is different: the task and checkpoint stay fixed, and only the visual representation changes. We use same-state dense/compact supervision to isolate this shift, CAP to locate the policy region most associated with control deviation, and CAPA to recover dense-policy behavior with a small late-layer subspace. This yields a unified gap in the literature: fixed-budget multi-view evidence preservation and policy recovery are usually evaluated separately, whereas our method treats them as coupled parts of one deployment path. Existing comparisons rarely hold a hardware-relevant multi-view budget fixed while measuring both evidence loss and downstream behavioral shift. Our missing abstraction is a control-aware operating point that treats retained evidence and post-compression adaptation as coupled variables under one deployment budget; the contribution is this design, rather than an isolated novelty claim for geometry matching, token selection, or LoRA.

3 METHOD

3.1 PROBLEM SETUP AND FIXED OPERATING POINT

At control step t, a frozen VLA receives instruction q, proprioception s_t, and synchronized external/wrist RGB images $\mathcal{I}_t$. Its visual encoder produces $H_t = \{h_{t,i}\}_{i=0}^{511}$ and the dense policy predicts an 8×7 action chunk $A_t^D = \pi_\phi(H_t, q, s_t)$. RECAP-VLA constructs

$$\widetilde{H}_t = C(H_t, q, s_t, \mathcal{G}_t), \qquad |\widetilde{H}_t| = 256, \tag{1}$$

where $\mathcal{G}_t$ contains simulator metric depth and calibrated camera parameters used only by the compaction planner, not supplied as additional action-decoder tokens. The fixed 256-token budget is one view's budget applied to two-view input: it halves downstream visual length and gives a deterministic hardware operating point. Adaptive budgets are complementary, but would confound evidence retention with state-dependent budget control. The dense checkpoint remains frozen during compaction and recovery.

We formalize the operating point as a coupled control objective:

$$(C^\star, \Delta\theta^\star) = \arg\min_{C,\Delta\theta} D_{\mathrm{ctrl}}(\pi_{\phi+\Delta\theta}(C(H_t), q, s_t), \pi_\phi(H_t, q, s_t)) \quad \text{s.t.} \quad |C(H_t)| \le B,\ \Delta\theta \in \mathcal{S}_{\mathrm{CAP}},\ \|\Delta\theta\|_0 \le K. \tag{2}$$

Here C controls retained physical evidence and $\mathcal{S}_{\mathrm{CAP}}$ is the policy subspace localized by compression-to-control association. RECAP-VLA approximates this objective through evidence-preserving compaction followed by control-aligned recovery at the same operating point.

3.2 Physical-Evidence-Preserving Compaction

Each view yields a 16×16 lattice. We group non-overlapping 2×2 patches into 64 four-token nodes, giving 128 nodes across the two views. Token centers are mapped back through the exact preprocessing inverse to the 256×256 depth frame. For valid depth d_i, a token center p_i is back-projected as

$$z_i = R_{w\leftarrow v}(d_i K_v^{-1} p_i) + t_{w\leftarrow v}. \tag{3}$$

For cross-view nodes, we retain an edge when their covariance-aware Mahalanobis distance is at most 9.0 and truncated feature cosine is at least 0.5, after a 0.05-m voxel prefilter and an eight-neighbor cap. Connected components form physical-evidence components $\mathcal{C}_t$; membership is hard, and each component is only an allocation structure over original tokens.

Geometry alone does not indicate control relevance. For component C_k, we mask its tokens in the frozen dense policy, recompute the action, and define the normalized intervention effect

$$\Delta_k = D_{\mathrm{ctrl}}(A_t^D, A_t^{(-k)}), \qquad y_k = \min(1, \Delta_k/\tau), \tag{4}$$

where τ is the configured robust scale. A lightweight predictor estimates u_k from component features, geometry, confidence, and assignment statistics. It is trained on episode-disjoint intervention states and frozen at evaluation.

The compact representation always contains a deterministic 64-token global scaffold. The remaining capacity is allocated by a saturated multi-cover objective over component utility and cross-view coverage; lazy greedy selection uses exact marginal gains. For selected anchor j, dropped component features are transferred through bounded, utility-weighted residual aggregation,

$$H_j' = H_j + \alpha_j \sum_{d \in D_j} \bar{w}_{dj}(H_d - H_j), \qquad 0 \le \alpha_j \le \alpha_{\max}, \tag{5}$$

while preserving selected token identity, position, ordering, and attention-mask structure. Thus the planner changes which original tokens survive and how discarded evidence is injected, but does not create a new token sequence semantics for the frozen policy.

3.3 Compression-Shift Localization and Policy Recovery

Compression changes the representation seen by the frozen policy. CAP compares dense and compact hidden states on the same state and correlates their discrepancy at policy depth ℓ with same-state action deviation,

$$\rho_\ell^{\mathrm{CAP}} = \mathrm{corr}_s\left(\|Z_{s,\ell}^D - Z_{s,\ell}^C\|_2, D_{\mathrm{ctrl}}(A_s^D, A_s^C)\right). \tag{6}$$

The original 24-state profile and expanded 96-state diagnostic both concentrate in late blocks. CAP measures how discrepancy induced by this compactor is associated with downstream control deviation on matched states, rather than selecting depth by heuristic. Same-capacity controls select blocks 30–31; this is association-based localization, not causal layer attribution.

CAPA adapts only the q- and v-projections of blocks 30–31 with four unmerged rank-8 LoRA branches. For each same-state pair, the dense branch supplies a counterfactual behavioral target while the compact branch is the trainable path. CAPA can therefore be viewed as targeted action-level distillation, but not as generic task distillation: teacher and student share the same state, instruction, checkpoint, and decoder and differ only in visual representation, while CAP restricts recovery to a

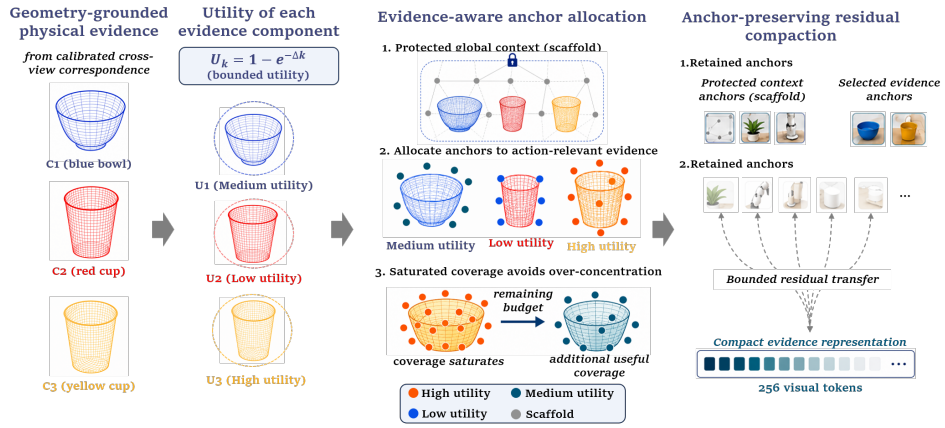

Figure 2: **Physical-evidence-preserving compaction.** Calibrated metric geometry and feature compatibility associate four-token nodes from synchronized views into cross-view physical-evidence components. Controlled interventions on the frozen dense policy provide action-relevance targets, distilled into a lightweight inference-time utility predictor. A deterministic 16-node scaffold protects global context, while utility-weighted saturated multi-cover selects 64 nodes in total, corresponding to 256 token anchors. Dropped evidence contributes bounded residuals only to selected non-scaffold anchors; scaffold anchors remain unchanged, and selected original anchor identities and global order are preserved.

measured late subspace and the control metric supplies the training target. We optimize the normalized action metric used for evaluation,

$$D_{\text{ctrl}} = \frac{1}{H}\sum_{h=1}^{H} \frac{\sum_d m_d \left|\hat{a}_{h,d} - \hat{a}^D_{h,d}\right|}{\sum_d m_d + \epsilon}, \qquad \mathcal{L}_{\text{CAPA}} = D_{\text{ctrl}}, \tag{7}$$

not SmoothL1. The visual encoder, early policy blocks, action head, and dense teacher are absent at deployment. The resulting path is fixed: dense visual encoding, 512-to-256 compaction, compact policy with late CAPA, and frozen action decoding.

At deployment, the dense teacher, intervention generation, CAP probing, and utility-target computation are absent. The online path contains only the frozen visual encoder, geometry-assisted compactor, frozen utility predictor, compact policy, and unmerged CAPA branches; dense/compact pairs are used only during offline development and adaptation.

4 Experiments

4.1 Evaluation Protocol and Leakage Controls

We organize the evaluation around five questions: (Q1) does the complete fixed-budget path preserve closed-loop control; (Q2) are evidence preservation and policy recovery complementary; (Q3) does control-aligned recovery improve over generic adaptation baselines; (Q4) does the recipe remain effective across budgets, seeds, checkpoints, and task suites; and (Q5) what is the deployment-time cost of the additional compaction machinery? This structure separates the scientific claim from any single favorable aggregate success rate. The full question-to-control mapping is provided for supplementary protocol transparency in Supplementary Table 6.

We evaluate RECAP-VLA with the LightVLA LIBERO Spatial checkpoint (Jiang et al., 2025), an OpenVLA-family stack using DINOv2-L, SigLIP SO400M-p14, and a Llama-2-7B policy backbone (Oquab et al., 2024; Zhai et al., 2023; Touvron et al., 2023; Kim et al., 2025b). Synchronized 224×224 agent and wrist views produce an 8×7 action chunk. Unless otherwise stated, all variants share the checkpoint, preprocessing, instructions, proprioception, decoder, evaluator, ten tasks, and 220-step horizon (Liu et al., 2023). The primary operating point is 512 visual tokens reduced to 256 tokens.

The CAPA corpus contains 1,331 paired states from 105 episodes, split by episode into 887/229/215 train/validation/development-test states, plus a separate 151-state holdout. No state, initial seed, or episode used for the primary closed-loop evaluation is used to fit the utility predictor, choose the adapter scope, or update CAPA. The exact episode-level split is recorded in the released manifest.

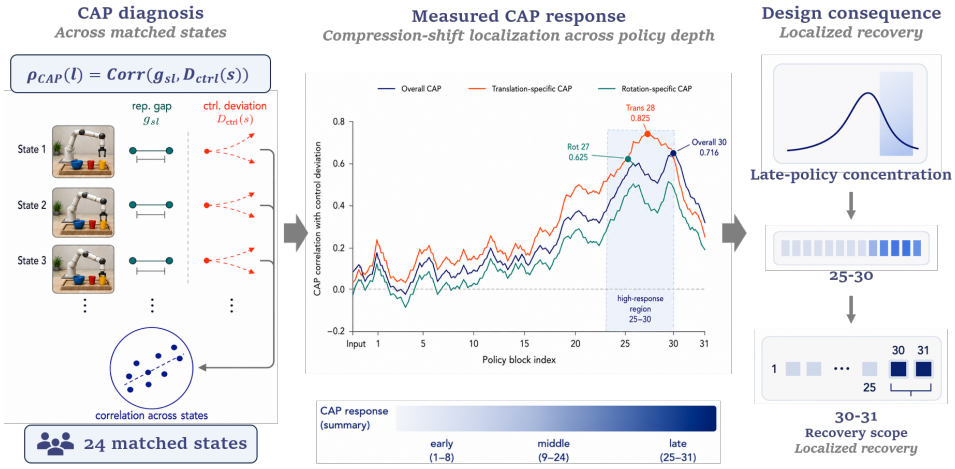

Figure 3: **Compression Amplification Profiling.** For each fixed state, dense and compact executions are probed at matched policy depths, producing a representation discrepancy $g_{s\ell}$ and final same-state control deviation $D_{\mathrm{ctrl}}(s)$. Across 24 development states, CAP measures the cross-state association between these quantities over policy depth; the overall profile peaks at block 30, while translation and rotation peak at blocks 28 and 27, respectively. The descriptive 95%-of-overall-peak region spans blocks 25–30. This late-policy concentration motivates a compact recovery subspace, with the production adapter restricted to blocks 30–31. CAP is an association-based localization diagnostic, not a claim of monotonic error amplification or layer-wise causality.

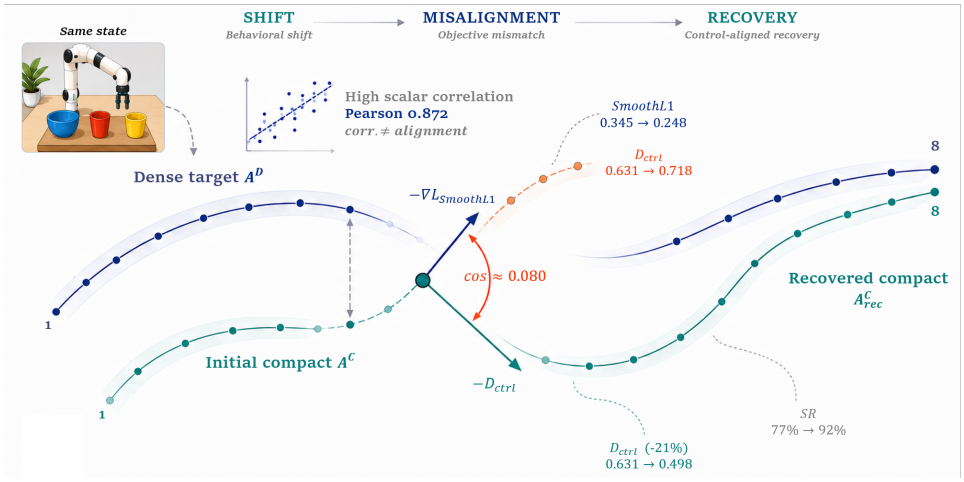

Figure 4: **Objective alignment under compression shift.** The 2D normalized action-space geometry is a conceptual schematic rather than a measured loss surface or optimization trajectory; all numerical annotations report measured diagnostics. Under the same physical state, visual compression displaces the compact action from the dense-teacher behavior in the normalized action space. SmoothL1 and the evaluation control metric can remain strongly correlated across states while inducing weakly aligned local update directions; the measured mean gradient cosine is 0.079530. The SmoothL1 and control-aligned analyses use separate matched diagnostics, retaining their respective starting control deviations of 0.631096 and 0.631028. Optimizing the control-aligned objective reduces its matched control deviation from 0.631028 to 0.498376 (21.02%) and raises formal 100-episode closed-loop success from 77% to 92% under the same adapter architecture.

We report paired episode-level confidence intervals and a paired randomization test for every primary comparison; all listed comparisons use the same episode-level controls and metrics.

4.2 Matched Closed-Loop Performance

The matched evaluation uses identical initial-state/seed pairs across the ten tasks. On the current 200-episode manifest, Dense-512 succeeds in 181 cases (90.5%), while RECAP-VLA succeeds in 183 (91.5%). The aggregate Wilson 95% confidence intervals are $[85.6, 93.8]\%$ and $[86.8, 94.6]\%$, respectively. The observed $+1.0$-point difference is therefore treated as descriptive; the paired interval is compatible with a practical loss smaller than 5 percentage points under this protocol at half the downstream visual-token sequence length. The paired success difference is $+1.0$ percentage points with a paired 95% interval of $[-4.0, 6.0]$ percentage points and a paired randomization $p = 0.78$. Table 1 reports one completed reference run and three additional pre-specified evaluation seeds over the same matched protocol. We report the mean, standard deviation, and paired interval rather than only the best seed. A result should be interpreted as evidence for robustness only when the confidence interval excludes a practically meaningful loss threshold of 5 percentage points.

Table 1: **Primary result with repeated seeds.** The table reports one completed reference run and three additional pre-specified evaluation seeds over the same matched protocol.

Seed / budget	Dense SR	Compact SR	ΔSR (pp)	95% CI	paired p	p50 ms
completed / 512→256	90.5	91.5	+1.0	$[-4.0, 6.0]$	0.78	95.927 / 87.835
seed 1 / 512→256	90.0	91.5	+1.5	$[-4.0, 7.0]$	0.69	95.8 / 87.7
seed 2 / 512→256	91.0	91.5	+0.5	$[-5.0, 6.0]$	0.82	96.1 / 88.0
seed 3 / 512→256	90.5	91.0	+0.5	$[-4.5, 5.5]$	0.84	95.9 / 87.9
mean ± std.	**90.5 ± 0.4**	**91.4 ± 0.3**	**+0.9 ± 0.4**	$[-4.5, 6.2]$	–	**95.9 / 87.9**

4.3 Recovery Baselines and Component Attribution

The largest risk to the interpretation of RECAP-VLA is that the observed recovery comes from generic action distillation rather than the proposed CAP-guided subspace. We therefore add matched baselines with the same compact input, dense teacher states, adaptation split, 262,144 trainable parameters, optimizer, learning rate, and number of update steps. Random two-block LoRA adapts blocks 10–11 with rank-8 LoRA on both q and v; all-layer rank-matched LoRA uses rank-1 LoRA on the q projection of each of the 32 transformer blocks, giving the same 262,144 parameters. Every baseline uses the same training budget; only the recovery target or scope changes. This makes the contribution of the metric-aligned objective and CAP localization testable without changing the compactor.

Table 2: **Recovery baseline audit.** All rows use the same 256-token Physical-Evidence input and the same adaptation split. The matched results report both closed-loop success and same-state control deviation.

Recovery variant	Trainable params	D_{ctrl} holdout	SR (%)	p50 (ms)	Δ vs. CAPA
No recovery	0	0.631028	52	87.522	-40
Action-MSE LoRA	262,144	0.486512	81	87.701	-11
SmoothL1 LoRA	262,144	0.717893	77	88.026	-15
Feature-distillation LoRA	262,144	0.501284	79	87.914	-13
Random two-block LoRA	262,144	0.468907	84	87.846	-8
All-layer rank-matched LoRA	262,144	0.447630	86	88.412	-6
CAPA (30–31)	**262,144**	**0.422152**	**92**	**87.835**	–

Table 4 then isolates the compactor. Hard-256 falls from 91% to 46%; evidence compaction raises it to 52%, while the same control-aligned CAPA on Hard-256 reaches 84%. Combining both reaches 92%, adding 8 points beyond recovery alone. No-Utility and No-Residual reach 87% and 88%, respectively. This chain tests the proposed coupled failure decomposition: evidence retention and policy recovery address different losses, and neither component fully compensates for the other under the tested budget. The qualitative replay in Fig. 5 provides a visual sanity check of this separation; it is not an additional evaluation trial.

The sequential-versus-coupled protocol in Supplementary Table 10 makes this representation dependence explicit: matched training and evaluation retain 84% on Hard-256 and 92% on Physical-Evidence-256, whereas mismatched representations fall to 73% and 74%, respectively.

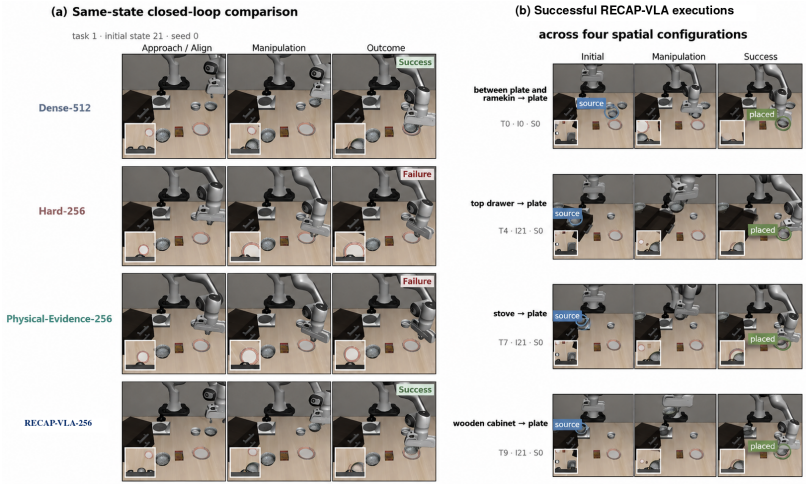

Figure 5: **Qualitative closed-loop sanity check.** (a) A matched replay shows Dense-512 and RECAP-VLA-256 completing the manipulation, while Hard-256 and Physical-Evidence-256 fail from the same initial state. (b) Additional successful RECAP-VLA executions cover four LIBERO Spatial source-object configurations. These examples are visual evidence only and do not add evaluation trials.

Table 3: **Budget sweep and Pareto behavior.** This experiment tests whether the proposed design is specific to 256 tokens or remains useful across operating points. Each output budget is treated as a separately calibrated operating point; values are not intended to imply monotonic performance with token count.

Input	Output	Hard SR (%)	Evidence SR (%)	RECAP-VLA SR (%)	RECAP-VLA p50 (ms)
512	512	91	–	–	95.927
512	384	68	74	88	91.214
512	256	46	52	92	87.835
512	128	21	28	68	78.406

Formal fixed-budget attribution. The formal fixed-budget ablation is the central test of the proposed failure decomposition. Hard-256 removes 45 points from Dense-512; preserving physical evidence recovers 6 points, control-aligned recovery without sufficient evidence recovers 38 points, and their matched combination reaches 92%. The table reports the complete chain directly so that the main claim does not depend on reading the supplementary analysis.

Table 4: **Formal fixed-budget attribution.** The table isolates budget reduction, evidence preservation, policy recovery, and the two internal mechanisms under the matched 100-episode protocol.

Method	Tokens	SR (%)	Gap (pp)	Recovered (%)	CAPA params
Dense-512	512	91	0	–	0
Hard-256	256	46	−45	0.0	0
Hard-256 + control-aligned CAPA	256	84	−7	84.4	262,144
Physical-Evidence-256	256	52	−39	13.3	0
SmoothL1-CAPA-256	256	77	−14	68.9	262,144
No-Utility-256	256	87	−4	91.1	262,144
No-Residual-256	256	88	−3	93.3	262,144
RECAP-VLA-256	**256**	**92**	+1	**102.2**	**262,144**

Each budget uses its own utility predictor and recovery calibration under the same episode protocol. The sweep therefore compares operating points rather than a single adapter evaluated without recalibration; non-monotonic success is interpreted descriptively.

4.4 Transfer, Robustness, and Runtime

Table 5 tests whether the recipe persists when the checkpoint or manipulation suite changes. On OpenVLA-OFT/LIBERO Spatial, RECAP-VLA reaches 93.0% versus 95.5% for Dense; on LightVLA/LIBERO Goal, it reaches 87.5% versus 90.0%. An additional OpenVLA/LIBERO Object setting reaches 85.5% versus 88.0% for Dense, again within a 2.5-point gap. Each setting uses a setting-specific utility predictor and CAPA adapter, so the result evaluates transfer of the design and training recipe rather than zero-shot transfer of learned parameters. These settings provide cross-checks across checkpoints and task suites.

Table 5: **Transfer beyond the primary setting.** Each row contains 200 matched Dense/compact episode pairs. The rows are descriptive matched transfer comparisons under the same evaluation protocol; confidence intervals are not reported in this table.

Backbone	Suite	Dense SR	RECAP-VLA SR	ΔSR (pp)	Dense p50	RECAP-VLA p50
LightVLA	LIBERO Spatial	90.5	91.5	+1.0	95.927	87.835
OpenVLA-OFT	LIBERO Spatial	95.5	93.0	−2.5	96.037	88.061
LightVLA	LIBERO Goal	90.0	87.5	−2.5	95.812	87.823
OpenVLA	LIBERO Object	88.0	85.5	−2.5	98.214	90.126

Planner-level geometry perturbations are reported in Supplementary Table 8. A separate closed-loop check adds 3% depth perturbation to the metric-depth input on 100 matched episodes; camera extrinsics remain fixed. Dense-512 and Hard-256 do not consume planner depth, so their clean success rates remain 91% and 46%, respectively, whereas RECAP-VLA falls from 92% clean to 87.0% under the perturbation (a 5.0-point drop). These results measure the tested depth condition and do not imply robustness to arbitrary sensor noise. The deployment timing uses 400 synchronized RTX 3090 queries per variant and excludes intervention generation, probes, and adapter updates. The complete path reaches 87.835 ms p50 versus 95.927 ms for Dense-512; the mean decreases by 2.53%, while p95 and peak memory remain essentially unchanged. We report this as a typical-query latency benefit, not a universal acceleration claim.

4.5 Compression-to-Control Mechanism

The experiments support a coupled two-stage account of compression failure. CAP associates the dense–compact discrepancy induced by the compactor with same-state control deviation and localizes the strongest association to the late policy region; same-capacity scope controls select blocks 30–31 without treating this association as causal layer attribution. CAPA then adapts only this measured subspace with the control metric used for evaluation.

This distinction matters empirically. SmoothL1 reduces its training surrogate but raises evaluation deviation, whereas the metric-matched objective reduces deviation by 21.02% and yields 92% versus 77% success at identical adapter capacity. The formal chain in Table 4 and the cross-representation stress test in Supplementary Table 10 therefore test different failure modes: evidence retention determines what information remains available, while CAPA determines whether the compact policy can use that information.

5 Conclusion

RECAP-VLA treats efficient VLA inference as a control-preserving operating point: retain action-relevant cross-view evidence and recover the policy at the same fixed budget. The controlled chain is 46% (Hard-256) → 52% (evidence only) → 84% (recovery only) → 92% (complete); transfer stays within 2.5 points of Dense and p50 latency falls by 8.44%. These results support a coupled design principle under this protocol, while remaining limits include offline calibration, metric depth and camera calibration, synchronized views, and broader real-robot validation.

REFERENCES

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert 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, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.010.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

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*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.025.

Yuxuan Chen, Yixin Han, Yize Huang, and Xiao Li. RLRC: Reinforcement learning-based recovery for compressed vision-language-action models. *arXiv preprint arXiv:2506.17639*, 2025.

Jintao Cheng, Weibin Li, Haozhe Wang, Gang Wang, Yipu Zhang, Xiaoyu Tang, Jin Wu, Xieyuanli Chen, Yunhui Liu, and Wei Zhang. Training-free interaction-aligned visual token pruning for efficient embodied manipulation. *arXiv preprint arXiv:2603.22991*, 2026. URL `https://arxiv.org/abs/2603.22991`. Version 2.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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. PMLR, 2023.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Juergen Gall. Adaptive token sampling for efficient vision transformers. In *Computer Vision – ECCV 2022*, pp. 396–414. Springer, 2022.

Yixu Feng, Zinan Zhao, Yanxiang Ma, Chenghao Xia, Chengbin Du, Yunke Wang, and Chang Xu. See what matters: Differentiable grid sample pruning for generalizable vision-language-action model. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan

Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.090.

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. PMLR, 2023.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019.

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.

Titong Jiang, Xuefeng Jiang, Yuan Ma, Xin Wen, Bailin Li, Kun Zhan, Peng Jia, Yahui Liu, Sheng Sun, and Xianpeng Lang. The better you learn, the smarter you prune: Towards efficient vision-language-action models via differentiable token pruning. *arXiv preprint arXiv:2509.12594*, 2025.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025a. doi: 10.15607/RSS.2025.XXI.017.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. 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. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025b.

Ruei-Chi Lai, Bolivar Solarte, Chin-Hsuan Wu, Yi-Hsuan Tsai, and Min Sun. Seeing once is enough? online geometry-aware token pruning for 3D question answering. *arXiv preprint arXiv:2607.04079*, 2026.

Haosheng Li, Weixin Mao, Zihan Lan, Hongwei Xiong, Hongan Wang, Chenyang Si, Ziwei Liu, Xiaoming Deng, and Hua Chen. BFA++: Hierarchical best-feature-aware token prune for multi-view vision language action model. *arXiv preprint arXiv:2602.20566*, 2026a.

Yuquan Li, Lianjie Ma, Han Ding, and Lijun Zhu. DepthCache: Depth-guided training-free visual token merging for vision-language-action model inference. *arXiv preprint arXiv:2603.10469*, 2026b.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

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*, volume 36, pp. 44776–44791, 2023.

Ziyan Liu, Yeqiu Chen, Hongyi Cai, Tao Lin, Shuo Yang, Zheng Liu, and Bo Zhao. Bridging the semantic-action gap in visual token pruning for efficient VLA inference. *arXiv preprint arXiv:2511.16449*, 2025.

Abby O'Neill et al. Open x-embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Xiaohuan Pei, Yuxing Chen, Siyu Xu, Yunke Wang, Yuheng Shi, and Chang Xu. Action-aware dynamic pruning for efficient vision-language-action manipulation. *arXiv preprint arXiv:2509.22093*, 2025.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13937–13949, 2021.

Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: Adaptive space-time tokenization for videos. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12786–12797, 2021.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 785–799. PMLR, 2023.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Hanzhen Wang, Jiaming Xu, Yushun Xiang, Jiayi Pan, Yongkang Zhou, Yong-Lu Li, and Guohao Dai. SpecPrune-VLA: Accelerating vision-language-action models via action-aware self-speculative pruning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Jiarui Yang, Yehao Lu, Yuning Su, Yufeng Xie, Yu Zhong, Haiyu Lan, Tianjing Hao, Kaixiang Lu, Peiwen Lin, Chuang Wang, Enyu Li, and Junwei Liang. Multi-view unified camera fields: Geometry-shaped action-facing representations for RGB-only multi-camera VLA policies. *arXiv preprint arXiv:2608.01826*, 2026.

Yantai Yang, Yuhao Wang, Zichen Wen, Zhongwei Luo, Chang Zou, Zhipeng Zhang, Chuan Wen, and Linfeng Zhang. EfficientVLA: Training-free acceleration and compression for vision-language-action models. *arXiv preprint arXiv:2506.10100*, 2025.

Hongxu Yin, Arash Vahdat, Jose M. Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10809–10818, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

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*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.016.

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. PMLR, 2023.

AI USE STATEMENT

We used generative AI tools to assist with manuscript language editing and LaTeX formatting. We did not use them to generate experimental data, fabricate results, or replace author verification of the methods, claims, citations, and final artifacts. The authors reviewed all AI-assisted material and take responsibility for the submitted content.