PGOT: PHYSICALLY GROUNDED OPTIMAL TRANSPORT FOR DISTILLING VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language–action (VLA) models bring large-scale visual and semantic knowledge to embodied control, but their computational demands motivate distillation into compact counterparts for responsive deployment. Existing distillation objectives, however, fall short in capturing the physical semantics and decision structure of embodied control: categorical matching of token distributions treats physical commands as unstructured labels, while unselective supervision dilutes learning at critical decision points and passes imperfect teacher decisions directly to students. We introduce **Physically Grounded Optimal Transport (PGOT)**, a distillation framework that addresses these gaps through three complementary mechanisms. **Geometry** aligns action distributions through optimal transport of probability mass, grounding teacher–student discrepancies in physical action space. **Criticality** exploits the temporal structure of expert trajectories to identify pivotal decision points and concentrate learning on them. **Reliability** adjusts teacher guidance according to its physical agreement with expert demonstrations, favoring trustworthy supervision across states and action dimensions. Across UAV navigation and robotic manipulation, PGOT distills 7B/9B teachers into 0.8B students that retain teacher-level success with approximately 9–11$\times$ fewer parameters. Over the strongest distillation baselines, it improves success rate by 2.67 and 2.40 percentage points and three-out-of-three completion by 7.33 and 6.00 points, respectively, exceeding even the teachers on the latter. Analyses of learning dynamics and closed-loop behavior further reveal that physical transport supplies optimization directions largely distinct from categorical matching, while selective supervision improves outcomes at critical maneuvers and limits the inheritance of teacher errors, with the largest gains where the teacher itself is unreliable.

1 INTRODUCTION

Vision–language–action (VLA) models connect visual understanding and task interpretation to physical control, bringing knowledge acquired through large-scale pretraining to embodied tasks (Driess et al., 2023; Jiang et al., 2023). An established approach, exemplified by RT-2 and OpenVLA, represents quantized actions as tokens and predicts them autoregressively (Brohan et al., 2023a; Kim et al., 2025b). This interface adapts pretrained vision–language models (VLMs) to control through sequence prediction on expert demonstrations, directly reusing their visual and semantic representations. Yet closed-loop interaction requires timely decisions, making the computational demands of large backbones a practical obstacle (Jeon et al., 2026; Shi et al., 2026). Compact models reduce this cost but lose much competence when trained on demonstrations alone: in our UAV navigation experiments, a fine-tuned 0.8B model succeeds in 53.33% of executions, against 87.56% for a 7B teacher. Knowledge distillation offers a practical path to efficient VLA deployment, enabling smaller models to benefit from the richer supervision of large teacher policies.

Distillation for general-purpose language and multimodal tasks facilitates knowledge transfer through distribution matching (Gu et al., 2024; Wang et al., 2025a; Jung et al., 2025) and representation alignment (Jiao et al., 2020). Recent efforts extend distillation to embodied control by aligning action predictions and transferring action-relevant visual–language representations (Jeon

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.

et al., 2026; Ye et al., 2025), incorporating task-phase and motion semantics (Shi et al., 2026), and obtaining teacher feedback on student-generated trajectories to mitigate distribution shift (Zhong et al., 2026). These advances address important requirements of embodied learning, yet the physical semantics of actions and the unequal value of supervision remain insufficiently reflected in distillation objectives. Categorical matching quantifies probability discrepancies without accounting for their physical consequences. During long-horizon execution, individual decisions differ in their importance to task completion, while teacher guidance varies in reliability across situations. Ignoring this heterogeneity can dilute learning at pivotal decisions and propagate imperfect teacher judgments, whose errors compound as executed actions shape subsequent observations (Ross et al., 2011). Effective VLA distillation therefore calls for physically meaningful distribution alignment and selective supervision that reflects the demands of control.

We introduce **Physically Grounded Optimal Transport (PGOT)**, a distillation framework that grounds teacher–student alignment in action geometry and further adapts supervision to its importance and reliability (Figure 1). Its three complementary mechanisms determine how action discrepancies are measured, where learning is emphasized, and which teacher guidance is retained. **Geometry** recasts action-distribution alignment as optimal transport of probability mass, using the Wasserstein distance to characterize the physical implications of distributional mismatch. **Criticality** exploits the temporal structure of expert trajectories to identify pivotal decisions and amplify their contribution to student learning. **Reliability** prioritizes trustworthy guidance by attenuating teacher predictions that substantially diverge from demonstrated behavior, limiting the propagation of local teacher weaknesses. All three mechanisms operate on decoded physical actions rather than token identities, so physical grounding shapes both the alignment metric and the allocation of supervision. Requiring only output distributions and offline demonstrations, and making no assumptions about teacher–student architectures, PGOT applies across diverse embodied domains.

We evaluate PGOT on aerial navigation and robotic manipulation using CoNav-UAV and LIBERO-10, respectively (Song et al., 2026; Liu et al., 2023). Across both domains, 0.8B students retain teacher-level success with approximately 9–11$\times$ fewer parameters than their 7B/9B teachers. Compared with the strongest baselines, PGOT improves success rate (SR) by 2.67 and 2.40 percentage points and three-out-of-three completion (Pass^3) by 7.33 and 6.00 points, demonstrating gains in both task competence and execution consistency. Analyses of learning dynamics and closed-loop behavior further explain these gains: physical transport supplies optimization directions largely distinct from categorical matching, while selective supervision improves outcomes at critical maneuvers and limits the inheritance of distant teacher errors, raising repeated completion most in states the teacher does not solve consistently.[1] Our contributions are summarized as follows:

- We introduce PGOT, which recasts action-distribution alignment as physically grounded probability transport and makes teacher transfer selective according to decision importance and guidance reliability.
- We demonstrate effective transfer from 7B/9B teachers to 0.8B students across aerial navigation and robotic manipulation, retaining teacher-level success while improving task completion and execution consistency over competing distillation methods.
- We provide a mechanistic analysis of PGOT, revealing how its supervision shapes student learning and linking these effects to measurable improvements in physical interaction.

2 Related Work

VLA modeling paradigms. Vision–language-guided control spans UAV navigation with TravelUAV and AutoFly (Wang et al., 2025b; Sun et al., 2026) and robotic manipulation with token-based controllers such as RT-1 (Brohan et al., 2023b). RT-2 and OpenVLA fine-tune VLMs to generate quantized actions autoregressively (Brohan et al., 2023a; Kim et al., 2025b), and FAST introduces frequency-based action tokens (Pertsch et al., 2025). Continuous-action alternatives use diffusion in Diffusion Policy and Octo (Chi et al., 2023; Ghosh et al., 2024), parallel regression in OpenVLA-OFT, and flow matching in π_0 (Kim et al., 2025a; Black et al., 2024). Discrete action tokens keep the VLM's native output interface and expose an explicit distribution over each action dimension

[1] Video demonstrations and code are available on our anonymous project page: `https://pgot-7425.pages.dev/`.

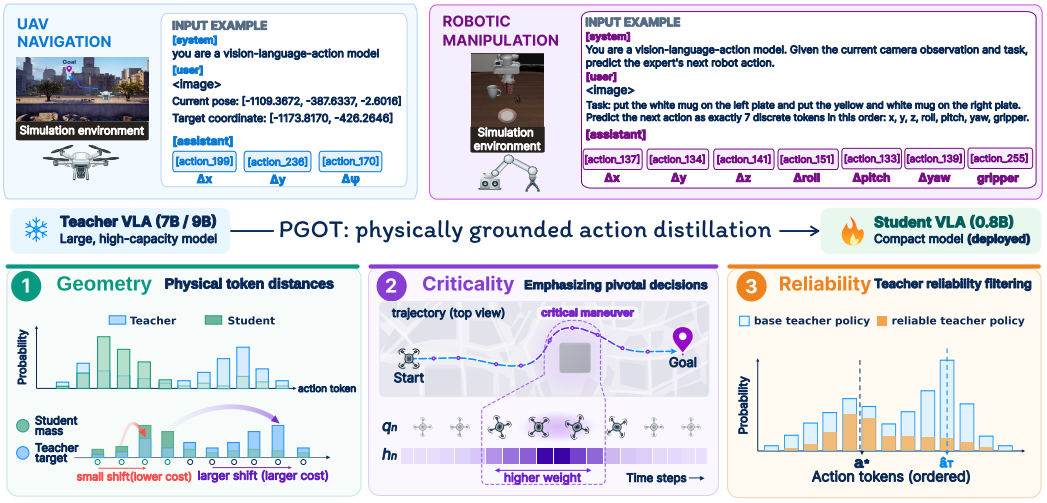


Figure 1: **PGOT grounds VLA distillation in physical action space. Top:** UAV navigation and robotic manipulation share one VLA interface that maps an observation and task context to one discrete action token per control dimension. **Center:** Arrows show knowledge transfer from the frozen teacher through PGOT to the compact, deployed student. **Bottom: Geometry** charges each shift of probability mass by the physical distance it moves; **Criticality** converts changes in expert actions (q_n) into sample weights (h_n) that rise at critical maneuvers; **Reliability** filters teacher probability far from the expert action a^*, attenuating guidance around a distant teacher mode $\hat{a}_T$.

that a teacher can transfer directly; PGOT builds on this interface and exploits the physical order that discretization preserves.

Knowledge distillation for language models. Knowledge distillation transfers teacher probabilities, generated sequences, or intermediate representations (Hinton et al., 2015; Kim & Rush, 2016; Jiao et al., 2020). MiniLLM uses reverse KL, ABKD balances probability-mass allocation through α–β divergence, and ToDi adapts divergence control at the token level (Gu et al., 2024; Wang et al., 2025a; Jung et al., 2025). GKD supervises student-generated sequences, DistiLLM combines skew KL with adaptive off-policy sampling (Agarwal et al., 2024; Ko et al., 2024), and DSKD aligns heterogeneous output spaces (Zhang et al., 2024). These objectives compare distributions over an unordered vocabulary, whereas PGOT measures discrepancy by the physical distance between decoded actions.

VLA distillation. VITA-VLA transfers an action expert into a VLM, while ActDistill combines hierarchical action supervision with dynamic layer routing (Dong et al., 2025; Ye et al., 2025). Shallow-π compresses flow-based VLAs through output and cross-attention distillation (Jeon et al., 2026); VLA-AD supplies offline task-phase and motion-direction descriptions as semantic guidance (Shi et al., 2026); VLA-OPD provides reverse-KL supervision on student-generated trajectories to address policy-induced distribution shift (Zhong et al., 2026); and RT-VLA distills driving and reasoning capabilities for real-time autonomous driving (Huang et al., 2026). These methods enrich the transferred signals and, in some cases, adapt where supervision is collected or how action dimensions are weighted. They do not explicitly use changes in expert actions to emphasize pivotal states and teacher–expert physical agreement to filter action-level guidance. PGOT uses both signals to make supervision selective.

3 Method

3.1 Preliminaries: Discrete Action Interfaces

PGOT distills VLA policies that emit actions as discrete tokens. The tokenizer clips each continuous action dimension d to a fixed range and quantizes it uniformly into one of $B = 256$ ordered bins

$\mathcal{B} = \{1, \ldots, B\}$; the detokenizer maps bin i back to its center $v_d(i)$. A command with D dimensions is thus a response of D action tokens, with $D = 3$ for UAV navigation and $D = 7$ for robotic manipulation. Each sample n pairs a context x_n (a visual observation with task information) with expert bins $\mathbf{a}_n^* \in \mathcal{B}^D$. Conditioned on x_n and the preceding expert tokens $a_{n,<d}^*$, teacher and student $M \in \{T, S\}$ assign

$$p_{M,n,d}(i) = \frac{\exp(z_{M,n,d,i})}{\sum_{j \in \mathcal{B}} \exp(z_{M,n,d,j})}, \tag{1}$$

where $z_{M,n,d,i}$ is the logit of bin i. Both models therefore define distributions over one shared support whose elements are physical commands: at deployment the student samples a bin and the detokenizer executes $v_d(i)$, so moving probability from bin i to bin j changes the executed command by $v_d(j) - v_d(i)$. Training minimizes $\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{PGOT}}$, where the base loss $\mathcal{L}_{\text{CE}}$ is cross-entropy against expert demonstrations, the hard-label supervision commonly retained alongside a distillation term (Jung et al., 2025; Jeon et al., 2026), and $\mathcal{L}_{\text{PGOT}}$ compares teacher and student action distributions. Appendix A gives the base loss and token-level details.

3.2 Overview

PGOT addresses the two gaps identified above with a single objective whose components all read decoded physical actions. **Geometry** is the alignment objective: it replaces categorical comparison with optimal transport whose ground cost is the physical distance between decoded bins, so each discrepancy is charged in proportion to the control error it implies. Transport alone still treats every supervised position and every teacher prediction as equally valuable, so the other two components calibrate it according to the value of supervision. **Criticality** decides *where* learning is emphasized, raising the weight of samples at which expert commands change, and **Reliability** decides *how much* the teacher is trusted, attenuating guidance whose implied command departs from the demonstrated one. Each is grounded in a physical quantity (a transport cost, a change in expert commands, and a teacher–expert command discrepancy), and each enters the objective at a distinct place (Equation 7).

3.3 Geometry: transport in physical action space

Forward KL (FKL) matches bin probabilities without expressing their physical separation: moving mass to an adjacent bin and to a distant one can incur the same penalty. Geometry instead aligns the two distributions by optimal transport with the decoded distance as ground cost (Villani, 2009; Frogner et al., 2015; Peyré & Cuturi, 2019). For an ordered dimension d, moving probability from bin i to bin j costs

$$c_d(i, j) = |v_d(i) - v_d(j)|/s_d, \qquad s_d = \max_i v_d(i) - \min_i v_d(i), \tag{2}$$

while a binary actuator such as the gripper uses the categorical cost $c_d(i, j) = \mathbb{I}[\kappa_d(i) \neq \kappa_d(j)]$, with κ_d mapping a bin to its nearest endpoint state. On an ordered support this transport has a closed form. Sorting decoded values as $\bar{v}_{d,1} \leq \cdots \leq \bar{v}_{d,B}$, reordering both distributions accordingly, and writing $F_M^{n,d}(k) = \sum_{i \leq k} \bar{p}_{M,n,d}(i)$,

$$\mathcal{L}_{\text{OT}}^{n,d} = \sum_{k=1}^{B-1} |F_S^{n,d}(k) - F_T^{n,d}(k)| \frac{\bar{v}_{d,k+1} - \bar{v}_{d,k}}{s_d} \tag{3}$$

is the exact one-dimensional Wasserstein-1 distance (Vallender, 1974), computable in $O(B)$. Each term multiplies the probability mass that must cross an action interval by the physical width of that interval, so mass misplaced by the student is penalized by how far it would move the executed command. Geometry applies to all ordered dimensions; the gripper receives only the filtered FKL term below.

3.4 Criticality: emphasizing pivotal decisions

Long-horizon control spends most decisions on routine motion, while task outcomes hinge on a few moments, such as a maneuver around an obstacle or the closing of a gripper. Criticality uses local changes in the expert's commands as an offline indicator of such decision points, so that the student learns most where demonstrated behavior itself changes. With $\mathbf{u}_n$ the decoded expert action

normalized by each dimension's tokenizer range and $\mathcal{N}(n)$ its immediately preceding and following frames in the same trajectory,

$$q_n = \sum_{m \in \mathcal{N}(n)} \|\mathbf{u}_n - \mathbf{u}_m\|_1, \qquad h_n = 1 + \gamma \,\mathrm{clip}\left(\frac{q_n}{\max(Q_{.95}, \epsilon)}, 0, 1\right), \tag{4}$$

where $Q_{.95}$ is the training-set 95th percentile of positive q_n and ϵ a small positive constant. Normalizing by a high percentile makes the weight insensitive to each domain's action scale, and the bound $h_n \in [1, 1+\gamma]$ caps the transition-related increase at γ.

3.5 Reliability: expert-consistent teacher transfer

A teacher is not uniformly reliable: in some states its prediction departs from the demonstration, and transferring it teaches behavior the expert contradicts. Reliability scales teacher guidance by its agreement with the demonstrated action, measured with the same physical cost as Geometry and applied at each term's natural granularity. FKL decomposes over bins, so the contribution of bin i, $\mathcal{L}^{n,d}_{\mathrm{FKL}}(i) = p_{T,n,d}(i) \log[p_{T,n,d}(i)/p_{S,n,d}(i)]$, receives the weight

$$w_{n,d}(i) = \exp[-\alpha c_d(i, a^*_{n,d})], \tag{5}$$

which filters teacher mass far from the demonstrated action. For the student gradient, this equals FKL toward the renormalized expert-conditioned teacher, scaled by the retained teacher mass (Appendix A), so a teacher concentrated far from the expert also contributes less overall. Transport compares whole distributions and thus receives one weight based on the teacher's most probable action,

$$\widehat{a}_{T,n,d} = \arg\max_i p_{T,n,d}(i), \qquad r_{n,d} = \exp[-\beta_r c_d(\widehat{a}_{T,n,d}, a^*_{n,d})], \tag{6}$$

with decay strengths $\alpha, \beta_r \geq 0$ controlling how quickly trust falls with physical disagreement.

Complete objective. For a minibatch containing N action positions, the distillation loss is

$$\mathcal{L}_{\mathrm{PGOT}} = \frac{1}{N} \sum_{n,d} \underbrace{h_n}_{\text{Criticality}} \left[\sum_{i \in \mathcal{B}} \underbrace{w_{n,d}(i)}_{\text{Reliability}} \mathcal{L}^{n,d}_{\mathrm{FKL}}(i) + w_1 \eta_d \underbrace{r_{n,d}}_{\text{Reliability}} \underbrace{\mathcal{L}^{n,d}_{\mathrm{OT}}}_{\text{Geometry}} \right], \tag{7}$$

where w_1 weights Geometry and η_d is one for ordered dimensions and zero for the gripper. Geometry supplies the physical alignment loss, Criticality scales all distillation terms of a sample, and Reliability scales teacher guidance within and across bins; colors match Figure 1.

4 Experiments

We evaluate PGOT on coordinate-conditioned UAV navigation and instruction-conditioned robotic manipulation to answer three questions. **RQ1 (effectiveness):** does PGOT improve task performance and execution consistency over general-purpose and VLA-specific distillation? **RQ2 (contribution):** what does each component contribute? **RQ3 (mechanism):** through what changes in student learning and interaction behavior do the components act? Sections 4.2–4.4 address them in turn.

4.1 Experimental setup

Domains and training. For **UAV navigation (CoNav-UAV)** (Song et al., 2026) in AirSim (Shah et al., 2018), an FPV image, current pose, and target coordinates condition three action tokens encoding planar displacement and yaw change. For **robotic manipulation (LIBERO-10)** (Liu et al., 2023), an agent-view image and task instruction condition seven tokens: three translations, three rotations, and a categorical gripper (Appendix B). Qwen2.5-VL-7B and Qwen3.5-9B teachers (Bai et al., 2025; Qwen Team, 2026) supervise Qwen3.5-0.8B students in the two domains. All distillation methods include the full-vocabulary cross-entropy base loss, and PGOT uses $(\alpha, \beta_r, \gamma, w_1) = (5, 5, 0.5, 10)$ in both domains. CoNav-UAV training uses 250,475 samples, 9,785 optimizer steps, and learning rate 2×10^{-6}; LIBERO-10 uses 101,469 samples, 10,000 steps, and 10^{-5}. All methods train on four NVIDIA H100 GPUs with effective batch size 128 and seed 42.

Evaluation. Each policy is executed three times per initial configuration, sampling action tokens autoregressively from a softmax over the 256 action tokens at temperature 0.3. CoNav-UAV covers 150 start–goal tasks, 30 per 100 m distance band up to 500 m, with a 200-step horizon and a 0.5 m collision threshold; LIBERO-10 covers 50 official initial states for each of 10 tasks, with a 520-step horizon (Appendices E and F). We report success rate (SR, mean $\pm$ sample standard deviation over three rounds) and Pass^3, the fraction of configurations completed in all three executions. Since repeated executions share an initial state and differ only in sampled tokens, Pass^3 measures whether a policy solves a state *every* time, and a gap between SR and Pass^3 reveals sensitivity to decoding noise. CoNav-UAV additionally reports collision rate (CR), timeout rate (TR), and mean minimum building clearance (MMBC, higher is safer); LIBERO-10 reports mean execution steps (MES) over all episodes (Appendix D).

Baselines. Teacher SFT and Student SFT are supervised references; distillation baselines are forward and reverse KL (FKL, RKL), ABKD (Wang et al., 2025a), ToDi (Jung et al., 2025), and the VLA methods Shallow-π and VLA-AD (Jeon et al., 2026; Shi et al., 2026) adapted to discrete action tokens (Appendix C).

4.2 Overall Effectiveness (RQ1)

Table 1: **Online evaluation performance.** Bold marks the best student value in each column. Green, blue, and purple rows denote KL baselines, LLM distillation, and VLA distillation.

	CoNav-UAV					LIBERO-10		
Policy	**SR** $\uparrow$	**Pass^3** $\uparrow$	**CR** $\downarrow$	**TR** $\downarrow$	**MMBC** $\uparrow$	**SR** $\uparrow$	**Pass^3** $\uparrow$	**MES** $\downarrow$
Teacher SFT	87.56 ± 1.39	78.67	9.78	2.67	5.00	49.13 ± 0.61	37.20	408.02
Student SFT	53.33 ± 1.15	38.67	44.00	**2.67**	2.78	45.47 ± 0.95	31.40	409.95
FKL	85.33 ± 1.33	74.67	11.78	2.89	5.27	44.93 ± 1.33	39.40	413.11
RKL	78.00 ± 2.31	68.00	16.00	6.00	5.00	37.73 ± 0.83	22.20	433.07
ABKD	80.89 ± 1.02	67.33	15.56	3.56	4.98	41.93 ± 1.03	27.40	416.14
ToDi	75.56 ± 3.42	62.00	17.78	6.67	5.00	43.73 ± 0.81	27.40	416.21
Shallow-π	85.33 ± 0.67	74.00	10.89	3.78	5.09	47.07 ± 1.62	32.40	407.13
VLA-AD	74.89 ± 0.38	58.00	20.89	4.22	4.20	39.73 ± 0.90	25.20	424.94
PGOT	$\mathbf{88.00 \pm 2.91}$	**82.00**	**8.00**	4.00	**5.72**	$\mathbf{49.47 \pm 0.64}$	**45.40**	**400.47**

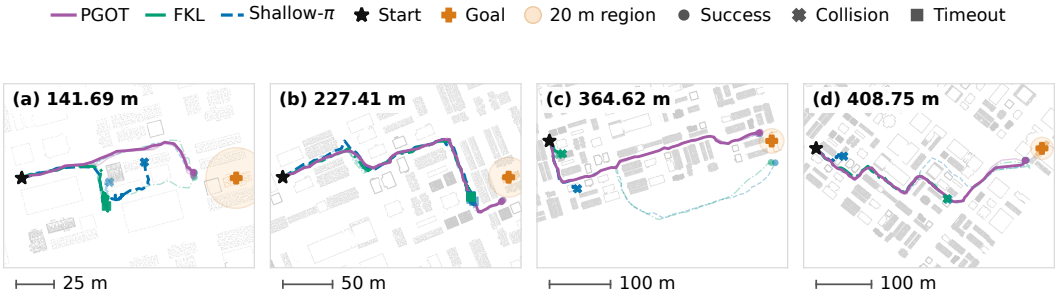


Figure 2: **Matched UAV navigation routes for four tasks, ordered by start–goal distance.** Dark routes mark a trial in which PGOT succeeds and both baselines fail; Appendix E.4 shows eight tasks.

PGOT outperforms all baselines on SR and Pass^3 in both domains (Table 1), and its advantage is concentrated in consistency. Over the strongest baseline on each metric, it adds 2.67 and 2.40 points of SR but 7.33 and 6.00 points of Pass^3 on CoNav-UAV and LIBERO-10. Both students match their teachers' SR and exceed their teachers' Pass^3 (82.00 vs. 78.67 and 45.40 vs. 37.20) with 9–11$\times$ fewer parameters, so distillation with PGOT recovers the teachers' competence and additionally reduces trial-to-trial variation.

The baselines differ in informative ways. The alternative divergences RKL, ABKD, and ToDi all trail plain FKL in both domains: they rebalance how the student covers or concentrates teacher mass

but keep the discrepancy measure categorical, and that reallocation does not improve on FKL here. The VLA-specific methods do not stabilize individual decisions either. Shallow-π reaches 47.07% SR on LIBERO-10, near PGOT, but only 32.40% Pass^3, leaving 30.4% of initial states solved in one or two of three executions against 8.6% for PGOT (Appendix E.5); VLA-AD, which adds offline semantic descriptions, trails FKL by 16.67 and 14.20 points of Pass^3. Enriching what is transferred therefore does not by itself remove per-trial variability, which is where the physical and selective parts of the objective act.

The advantage also persists as the navigation horizon grows: PGOT keeps at least 80% Pass^3 in every CoNav-UAV distance band and leads FKL by 10.00 points of Pass^3 in each of the two longest bands (Appendix E.3; Figure 2).

4.3 Component contribution (RQ2)

Table 2: **Component ablations.** Removing Geometry sets $w_1 = 0$, Criticality sets $\gamma = 0$, and Reliability sets $\alpha = \beta_r = 0$. Metrics follow Table 1.

	CoNav-UAV					LIBERO-10		
Objective	**SR ↑**	**Pass^3 ↑**	**CR ↓**	**TR ↓**	**MMBC ↑**	**SR ↑**	**Pass^3 ↑**	**MES ↓**
PGOT	**88.00 ± 2.91**	**82.00**	**8.00**	**4.00**	**5.72**	**49.47 ± 0.64**	**45.40**	**400.47**
w/o Geometry	82.22 ± 2.14	72.67	13.56	4.22	5.25	44.60 ± 0.87	29.80	415.11
w/o Criticality	83.78 ± 1.54	74.67	10.67	5.56	5.52	48.67 ± 0.90	42.40	402.95
w/o Reliability	84.89 ± 3.29	75.33	9.78	5.33	5.37	47.67 ± 1.62	34.20	407.19

Removing any component lowers both metrics in both domains (Table 2), and in all six cases the Pass^3 loss exceeds the SR loss, repeating at the component level the consistency gain of the full method. The components nonetheless differ in where their effect appears. Without Geometry, the student loses the only term that penalizes action errors by physical magnitude; SR falls by 5.78 and 4.87 points, and on CoNav-UAV collisions rise from 8.00% to 13.56% while clearance drops from 5.72 m to 5.25 m, degrading interaction quality near obstacles. Without Reliability, the effect is specific to consistency: LIBERO-10 loses 11.20 points of Pass^3 but only 1.80 of SR, as expected of a mechanism that removes contradicting teacher guidance in a minority of states (Section 4.4.3). Without Criticality, the smallest SR change (0.80 points on LIBERO-10) accompanies a 3.00-point Pass^3 loss, consistent with weights concentrated on sparse transition frames. Configuration-level pass–fail counts appear in Appendix F.3.

4.4 Mechanism analysis (RQ3)

We relate each component to prediction diagnostics and closed-loop measurements (Appendix D.3). *Physical* W_1 and *expected physical error* measure, in normalized action units, how far student probability lies from the teacher and the expert. Behavioral statistics use executions that both compared policies complete, so they measure how a policy succeeds rather than whether it does.

4.4.1 Geometry: a distinct physical learning signal

The transport term is not redundant with FKL. If Geometry merely rescaled the categorical objective, its gradient would be parallel to that of Filtered FKL. Instead, the two stand at mean angles of 73.94° and 72.05° on CoNav-UAV and LIBERO-10, with the transport RMS at 15.13% and 9.94% of the FKL RMS (Figure 3). This largely non-redundant direction changes the student's update despite its smaller magnitude, consistent with Geometry having the largest ablation effect. As training proceeds on CoNav-UAV, the ratio falls from 0.209 to 0.123 while the cosine similarity rises from 0.198 to 0.314 (Appendix G.2): the physical correction decays as the student learns and becomes more aligned with the residual categorical error.

The correction reaches prediction and execution. Against w/o Geometry, CoNav-UAV validation physical W_1 falls by 9.53% and expected physical error by 8.52%, so student mass moves closer to both teacher and expert in physical terms, not only in token index. The change is concentrated in the tail: under teacher forcing on training samples, probability at least 32 bins (one eighth of the

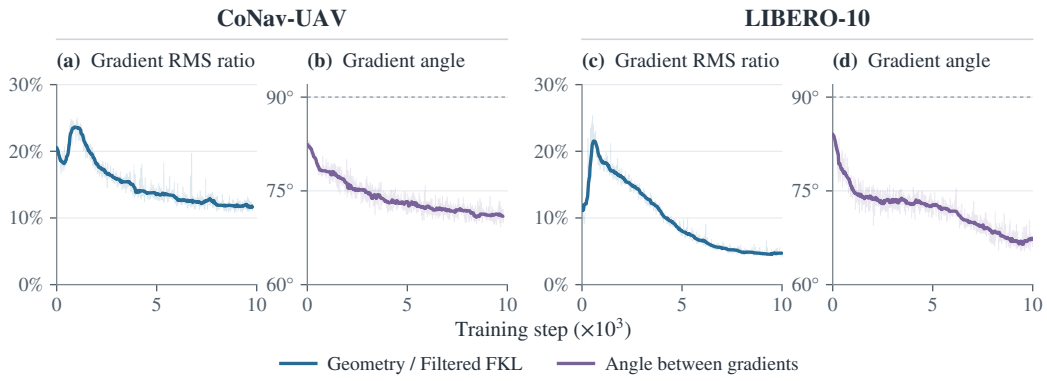


Figure 3: **Geometry versus Filtered FKL gradients during training.** Faint curves show measurements every ten steps; dark curves show 31-measurement rolling medians.

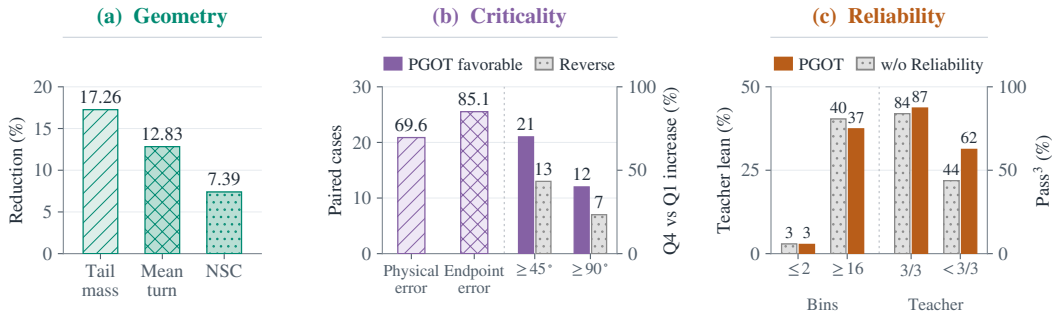


Figure 4: **CoNav-UAV component diagnostics**, each against the named ablation. (a) Reductions in tail mass (probability $\geq$32 bins from the expert under teacher forcing), mean turn, and NSC. (b) Right: Q4-over-Q1 error increases by Criticality weight; left: trials with a late sharp turn won by PGOT or by the ablation. (c) Teacher lean by teacher–expert bin distance, and Pass^3 by the teacher's success count.

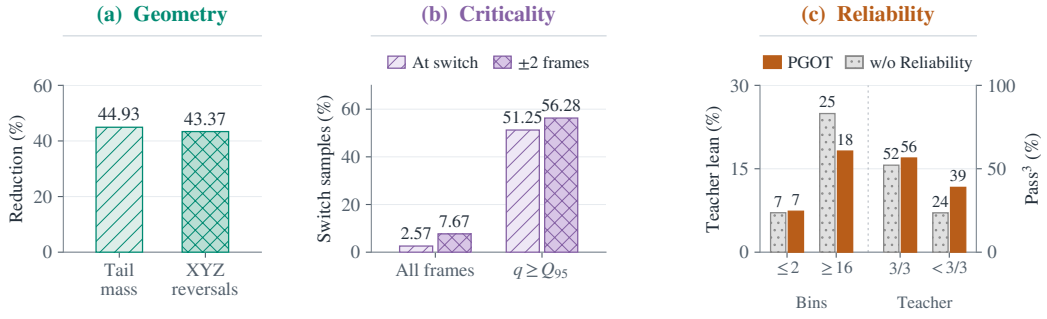


Figure 5: **LIBERO-10 component diagnostics**, each against the named ablation. (a) Reductions in tail mass and in XYZ command reversals beyond 120° on jointly successful executions. (b) Share of expert frames at a gripper switch or within two frames of one, among all frames and among those at the maximum Criticality weight. (c) Teacher lean by teacher–expert bin distance, and Pass^3 on initial states the teacher completes in all or fewer than three executions.

action range) from the expert falls by 17.26% on CoNav-UAV and 44.93% on LIBERO-10, while mass 3–15 bins away rises by 10.5% and 12.0% (Figures 4a and 5a). Geometry thus turns distant errors into near misses, removing the rare far-off commands behind sharp turns and reversals. On

jointly successful attempts, mean turn falls by 12.83% and normalized step change by 7.39%; on LIBERO-10, XYZ command reversals beyond 120° fall from 0.748% to 0.423%.

4.4.2 CRITICALITY: WEIGHTING THE DIFFICULT STATES

High-weight states are those students mispredict. Criticality helps only if the action changes it measures track difficulty. Grouping CoNav-UAV validation samples into quartiles of weight, the highest quartile has 69.6% greater expected physical error and 85.1% greater endpoint error than the lowest, and modal-action accuracy falls monotonically across the four quartiles from 0.268 to 0.118, so the weighting follows a graded difficulty signal rather than a few extreme samples. On LIBERO-10, the same indicator localizes on manipulation decisions: gripper-switch frames make up 51.25% of samples at or above the 95th-percentile action-change scale, against 2.57% overall (Figure 5b).

The emphasis improves learning and maneuver outcomes. Under teacher forcing on CoNav-UAV training samples, PGOT's expected physical error relative to w/o Criticality is unchanged below the median weight but lower by 1.66% between the median and the 95th percentile and by 3.57% above it, so the emphasis acts where it is placed. In w/o Criticality rollouts, a turn of at least 45° occurs in the final five transitions of 73.9% of collisions but only 35.3% of successes, so late sharp maneuvers precede failure. Among matched trials in which the colliding rollout contains such a maneuver, PGOT succeeds in 21 cases where the ablation collides, against 13 in the reverse direction; at 90° the counts are 12 and 7 (Figure 4b).

4.4.3 RELIABILITY: WITHHOLDING DISTANT TEACHER ERRORS

Reliability reweights teacher guidance only where the teacher departs from the expert: with $\alpha = 5$, teacher probability within two bins of the demonstrated action keeps a weight above 0.96 and probability sixteen or more bins away less than 0.73, and the gate r decays likewise. Removing it should thus matter most where teacher and expert are far apart.

Distant teacher errors are inherited less. Under teacher forcing on all training samples, we take the action positions where the teacher's mode $\hat{a}_T$ differs from the expert bin a^* and measure the student's *teacher lean*, $[p_S(\hat{a}_T) - p_S(a^*)]/[p_S(\hat{a}_T) + p_S(a^*)]$, which is zero when the student weighs the two bins equally (Appendix D.3). With the teacher one or two bins from the expert, the two students coincide (2.8% versus 2.9% on CoNav-UAV and 7.4% versus 7.1% on LIBERO-10); with the teacher sixteen or more bins away, PGOT leans less toward it (37.4% versus 40.4% and 18.2% versus 24.9%; Figures 4c and 5c). The students differ only where the filter is active.

Gains concentrate where the teacher is unreliable. Such deviations matter most in states the teacher itself does not solve consistently. On configurations the Teacher SFT policy fails at least once in three executions, PGOT raises Pass^3 over w/o Reliability from 43.75% to 62.50% on CoNav-UAV (32 tasks) and from 23.57% to 38.85% on LIBERO-10 (314 initial states), against 3.39 and 4.30 points where the teacher always succeeds. These states supply 6 of the 10 additional CoNav-UAV tasks and 48 of the 56 additional LIBERO-10 states completed in all three executions. There PGOT mainly turns occasional success into consistent success, which explains why the ablation loss in Table 2 falls mostly on Pass^3.

5 CONCLUSION

PGOT distills discrete-action VLA policies with physically grounded, selective supervision: Geometry aligns action distributions by physical transport, Criticality emphasizes pivotal decisions, and Reliability retains teacher guidance consistent with the expert. On CoNav-UAV and LIBERO-10, 0.8B students match their 7B/9B teachers' success, exceed their three-out-of-three completion, and surpass all distillation baselines, most clearly in repeated completion. Our analyses trace these gains to the components: transport turns distant action errors into near misses, weighting improves learning and outcomes at pivotal maneuvers, and filtering keeps the student from inheriting distant teacher errors, with the largest gains where the teacher is unreliable. Requiring only output distributions and offline demonstrations, PGOT leaves the student's architecture and inference unchanged.

AI USE STATEMENT

Generative AI tools were used to polish the language of the manuscript. The authors take responsibility for the final content.

ETHICS STATEMENT

This work studies UAV navigation and robotic manipulation in simulation using existing expert demonstrations. The experiments do not involve human subjects or public deployment.

REPRODUCIBILITY STATEMENT

Our code is open-sourced and available on our anonymous project page: `https://pgot-7425.pages.dev/`.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 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, 2023a. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, et al. RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems*, 2023b. doi: 10.15607/RSS.2023.XIX.025. URL https://roboticsproceedings.org/rss19/p025.html.

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. URL https://roboticsproceedings.org/rss19/p026.html.

Shaoqi Dong, Chaoyou Fu, Haihan Gao, Yi-Fan Zhang, Chi Yan, Chu Wu, Xiaoyu Liu, Yunhang Shen, Jing Huo, Deqiang Jiang, Haoyu Cao, Yang Gao, Xing Sun, Ran He, and Caifeng Shan. VITA-VLA: Efficiently teaching vision-language models to act via action expert distillation. *arXiv preprint arXiv:2510.09607*, 2025. URL https://arxiv.org/abs/2510.09607.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 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.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A. Poggio. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/a9eb812238f753132652ae09963a05e9-Abstract.html.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.090. URL https://www.roboticsproceedings.org/rss20/p090.html.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Xiangyu Huang, Zhenlin Hua, Han Zhou, Shounak Sural, and Ragunathan Rajkumar. RT-VLA: Real-time vision-language-action models via knowledge distillation. *arXiv preprint arXiv:2606.14010*, 2026. URL https://arxiv.org/abs/2606.14010.

Boseong Jeon, Yunho Choi, and Taehan Kim. Shallow-π: Knowledge distillation for flow-based VLAs. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2026. URL https://arxiv.org/abs/2601.20262.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: Robot manipulation with multimodal prompts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14975–15022, 2023. URL https://proceedings.mlr.press/v202/jiang23b.html.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.372. URL https://aclanthology.org/2020.findings-emnlp.372/.

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. ToDi: Token-wise distillation via fine-grained divergence control. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8091, 2025. doi: 10.18653/v1/2025.emnlp-main.409.

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*, 2025a. doi: 10.15607/RSS.2025.XXI.017. URL https://www.roboticsproceedings.org/rss21/p017.html.

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, 2025b. URL https://proceedings.mlr.press/v270/kim25c.html.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

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, 2023.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient action tokenization for vision-language-action models. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.012. URL https://www.roboticsproceedings.org/rss21/p012.html.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073. URL https://www.nowpublishers.com/article/Details/MAL-073.

Qwen Team. Qwen3.5: Towards native multimodal agents. Official Qwen model release, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, pp. 621–635. Springer, 2018.

Jin Shi, Brady Zhang, and Yishun Lu. Offline semantic guidance for efficient vision-language-action policy distillation. *arXiv preprint arXiv:2605.16241*, 2026. URL https://arxiv.org/abs/2605.16241.

Junru Song, Wenhao Zhang, Yang Yang, Xuekai Qiu, Feifei Wang, Weien Zhou, Tingsong Jiang, Ying Wen, Yang Li, and Wen Yao. CoNav-UAV: Cooperative dual-altitude aerial navigation via stackelberg learning. *arXiv preprint arXiv:2608.01802*, 2026. URL https://arxiv.org/abs/2608.01802.

Xiaolou Sun, Wufei Si, Wenhui Ni, Yuntian Li, Dongming Wu, Fei Xie, Runwei Guan, He-Yang Xu, Henghui Ding, Yuan Wu, Yutao Yue, Yongming Huang, and Hui Xiong. AutoFly: Vision-language-action model for UAV autonomous navigation in the wild. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=88RKxlFUNY.

S. S. Vallender. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability and Its Applications*, 18(4):784–786, 1974. doi: 10.1137/1118101.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.

Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. ABKD: Pursuing a proper allocation of the probability mass in knowledge distillation via α-β-divergence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65167–65212, 2025a. URL https://proceedings.mlr.press/v267/wang25dz.html.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic UAV vision-language navigation: Platform, benchmark, and methodology. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/15ce8e7afe5ee95bad56e3b9be28d3d1-Abstract-Conference.html.

Wencheng Ye, Tianshi Wang, Lei Zhu, Fengling Li, Guoli Yang, and Hengtao Shen. ActDistill: General action-guided self-derived distillation for efficient vision-language-action models. *arXiv preprint arXiv:2511.18082*, 2025. URL https://arxiv.org/abs/2511.18082.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. Dual-space knowledge distillation for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18164–18181, 2024. doi: 10.18653/v1/2024.emnlp-main.1010. URL https://aclanthology.org/2024.emnlp-main.1010/.

Zhide Zhong, Haodong Yan, Junfeng Li, Junjie He, Tianran Zhang, and Haoang Li. VLA-OPD: Bridging offline SFT and online RL for vision-language-action models via on-policy distillation. *arXiv preprint arXiv:2603.26666*, 2026. URL https://arxiv.org/abs/2603.26666.

A Objective Details

Base loss and action support. The base loss is cross-entropy against the expert demonstrations,

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} \log \frac{\exp(z_{S,t,y_t^*})}{\sum_{v\in\mathcal{V}_S}\exp(z_{S,t,v})}, \tag{8}$$

where $\mathcal{I}$ indexes supervised token positions in the minibatch, y_t^* is the expert target token, $\mathcal{V}_S$ is the student's full vocabulary, and $z_{S,t,v}$ is its logit for token v at position t. The distillation loss instead uses distributions restricted to the action-token indices (Equation 1). Teacher and student action-token logits are aligned by token string, which accommodates backbones that assign different vocabulary IDs to the same action token. The cost in Equation 2 and the transport in Equation 3 are computed from decoded values after sorting, so they do not depend on the vocabulary order of the action tokens; bins that decode to the same center contribute zero-width intervals.

Filtered-teacher interpretation. At one action position, write $p_i = p_T(i)$, $q_i = p_S(i)$, and $w_i = \exp[-\alpha c(i, a^*)]$. Define the retained teacher mass and its normalized distribution as $Z = \sum_i w_i p_i$ and $\widetilde{p}_i = w_i p_i / Z$. The weighted FKL term in Equation 7 decomposes as

$$\sum_i w_i \mathcal{L}_{\mathrm{FKL}}(i) = Z\,\mathrm{KL}(\widetilde{p}\|q) + Z\log Z - \sum_i w_i p_i \log w_i. \tag{9}$$

The last two terms are constant with respect to the student. Thus the filtered objective has the student gradient of a KL toward the expert-conditioned teacher, scaled by its retained mass Z. This scaling also reduces supervision from teacher distributions concentrated far from the expert. The transport and complete-objective definitions are given in Equations 3 and 7.

B Task Goals and Input–Output Templates

Each policy maps a visual observation and task context to one discrete action token per control dimension. Below, $i_d \in \{0, \ldots, 255\}$ denotes the token index for dimension d.

UAV navigation (CoNav-UAV)

Task goal. Navigate to within 20 m of the target coordinates while avoiding buildings, at a fixed flight altitude of 15 m.

Visual input. One first-person-view RGB image.

System message.

```
you are a vision-language-action model
```

User text template.

```
Current pose:  [x, y, ψ]
Target coordinate:  [x_g, y_g]
```

Output template.

```
[action_i1][action_i2][action_i3]
```

The three tokens encode $(\Delta x, \Delta y, \Delta\psi)$: planar displacement and yaw change.

Robotic manipulation (LIBERO-10)

Task goal. Manipulate the specified objects to satisfy the spatial relations and object states described by the instruction. For example, place the white mug on the left plate and the yellow-and-white mug on the right plate.

Visual input. One agent-view RGB image.

System message.
`You are a vision-language-action model. Given the current camera observation and task, predict the expert's next robot action.`

User text template.
`Task:` ⟨task instruction⟩
`Predict the next action as exactly 7 discrete tokens in this order: x, y, z, roll, pitch, yaw, gripper.`

Output template.
`[action_`i_1`][action_`i_2`][action_`i_3`][action_`i_4`][action_`i_5`][action_`i_6`][action_`i_7`]`
The seven tokens encode three translations, three rotations, and the categorical gripper state, in the order specified above.

C Baseline Settings

Both domains use the same baseline configurations. ABKD uses divergence coefficients $(\alpha_{\text{ABKD}}, \beta_{\text{ABKD}}) = (0.3, 0.7)$. Shallow-$\pi$ combines expert cross-entropy, forward-KL action distillation, and action-to-vision attention alignment. VLA-AD uses paired inputs with and without offline semantic descriptions, applying equally weighted expert-action cross-entropy to both branches of a shared student; the descriptions are used only during training.

D Metrics and Statistical Units

D.1 Success statistics

Let $y_{nr} \in \{0, 1\}$ indicate success for configuration n in evaluation round r, and let $s_n = \sum_{r=1}^{3} y_{nr}$ count successful executions. There are $N = 150$ start–goal tasks in CoNav-UAV and $N = 500$ task–initial-state pairs in LIBERO-10. For each round, $\text{SR}_r = N^{-1} \sum_n y_{nr}$. Tables report $\overline{\text{SR}} \pm s_{\text{SR}}$, where $s_{\text{SR}} = \sqrt{\frac{1}{2} \sum_{r=1}^{3} (\text{SR}_r - \overline{\text{SR}})^2}$. Mean success and repeated completion are

$$\overline{\text{SR}} = \frac{1}{3N} \sum_n s_n, \qquad \text{Pass}^3 = \frac{1}{N} \sum_n \mathbb{I}[s_n = 3]. \tag{10}$$

Collisions and timeouts count as failures in SR and Pass^3.

D.2 Navigation motion and manipulation execution steps

For a CoNav-UAV trajectory with planar positions $\mathbf{p}_t$, including the initial position $\mathbf{p}_0$, define $\mathbf{d}_t = \mathbf{p}_t - \mathbf{p}_{t-1}$. Retain displacements of at least 0.05 m in temporal order, denoting them by $\mathbf{v}_1, \ldots, \mathbf{v}_K$. For $K \geq 2$, normalized step change is

$$\text{NSC} = \frac{(K-1)^{-1} \sum_{k=2}^{K} \|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2}{K^{-1} \sum_{k=1}^{K} \|\mathbf{v}_k\|_2}. \tag{11}$$

The component diagnostics average episode-level NSC over paired trajectories completed successfully by both policies; lower values indicate smoother motion.

For LIBERO-10, let T_{nr} be the number of execution steps in configuration n and round r. We compute mean execution steps (MES) as $\text{MES} = (3N)^{-1} \sum_{n,r} T_{nr}$, with $N = 500$, including successful and unsuccessful episodes. Episodes end at success or the 520-step horizon. The ten settling steps precede policy execution and are excluded.

D.3 DIAGNOSTIC DEFINITIONS

For the definitions below, $p^S_{n,d,i}$ and $p^T_{n,d,i}$ are the student and teacher probabilities at action bin i at temperature one, $y_{n,d}$ is the expert bin, and c_d is the normalized decoded cost in Equation 2.

Table 3: Training and validation metric definitions. Physical distribution errors are averaged over samples and action dimensions.

Metric	Meaning	Definition
Expected physical error	Student-probability-weighted physical distance from the expert action	$E = \mathrm{mean}_{(n,d)} \sum_i p^S_{n,d,i} c_d(i, y_{n,d})$ over the evaluated samples
Physical W_1 / Geometry	Teacher–student mass displacement on the decoded action axis	$\sum_{j=1}^{B-1} \lvert F^S_j - F^T_j \rvert (\bar{v}_{j+1} - \bar{v}_j)/s_d$ after sorting the support by v
Geometry gradient RMS	Effective transport-gradient strength over M student action-logit coordinates	$\mathrm{RMS}(g_{\mathrm{OT}}) = \sqrt{M^{-1} \sum_{m=1}^{M} g^2_{\mathrm{OT},m}}$
Gradient ratio and angle	Relative strength and directional novelty of Geometry versus Filtered FKL	$\mathrm{RMS}(g_{\mathrm{OT}})/\,\mathrm{RMS}(g_{\mathrm{filt}})$; $\theta = \arccos \frac{g_{\mathrm{OT}}^\top g_{\mathrm{filt}}}{\lVert g_{\mathrm{OT}} \rVert_2 \lVert g_{\mathrm{filt}} \rVert_2}$
Endpoint error	Euclidean error of the argmax planar action endpoint	$\sqrt{(v_{x,\hat{y}_x} - v_{x,y_x})^2 + (v_{y,\hat{y}_y} - v_{y,y_y})^2}$, where $\hat{y}_d = \arg\max_i p^S_{d,i}$
Bin error	Mean absolute token-index distance between modal and expert actions	$(ND)^{-1} \sum_{n,d} \lvert \hat{y}_{n,d} - y_{n,d} \rvert$
Criticality quartile error index	Error in each Criticality quartile relative to the lowest quartile	$I_k = E(Q_k)/E(Q_1)$, where Q_k is the k-th quartile of CoNav-UAV validation samples by h_n. Endpoint error is normalized identically.

For a ground-plane trajectory, let $\mathbf{p}_t$ be position, $\mathbf{d}_t = \mathbf{p}_t - \mathbf{p}_{t-1}$ a valid motion step with $\lVert \mathbf{d}_t \rVert_2 \geq 0.05$, and

$$\theta_t = \arccos \frac{\mathbf{d}_{t-1}^\top \mathbf{d}_t}{\lVert \mathbf{d}_{t-1} \rVert_2 \lVert \mathbf{d}_t \rVert_2}. \tag{12}$$

Let T_θ denote the number of valid direction transitions.

Table 4: Online evaluation, trajectory, and paired-outcome metric definitions.

Metric	Meaning	Definition
CR / TR	Fractions of executions ending in collision or timeout	$\mathrm{CR} = (3N)^{-1} \sum_{n,r} \mathbb{I}[z_{nr} = C]$, $\mathrm{TR} = (3N)^{-1} \sum_{n,r} \mathbb{I}[z_{nr} = T]$; z_{nr} is the terminal status, with C for collision and T for timeout.
MMBC	Mean episode-minimum distance to non-target buildings	$\lvert\mathcal{E}\rvert^{-1} \sum_{(n,r)\in\mathcal{E}} c_{nr}$, where $c_{nr} = \min_{t,\,\mathbf{b}\in\mathcal{B}_n} \lVert\mathbf{p}_{nrt} - \mathbf{b}\rVert_2$. t indexes recorded waypoints. $\mathcal{E}$ contains complete trajectories; $\mathcal{B}_n$ contains non-target building contour points, excluding noise points.
Mean turn	Average change in motion direction	$T_\theta^{-1} \sum_t \theta_t$
Reversal rate	Fraction of consecutive commands that reverse direction	$R_\delta = \mathrm{mean}_t\, \mathbb{I}[\angle(\mathbf{u}_{t-1}, \mathbf{u}_t) > \delta]$ over span-normalized LIBERO-10 XYZ commands, $\delta = 120^\circ$.
Normalized step change (NSC)	Change between adjacent displacement vectors relative to mean step length	Equation 11
Terminal maneuver rate	Fraction of rollouts with a threshold turn in the final five valid motion transitions	$N^{-1} \sum_j \mathbb{I}[\max_{t\in\mathrm{last5}} \theta_{j,t} \geq \delta]$, $\delta \in \{45^\circ, 90^\circ\}$
Paired outcome transition	Same-task, same-trial recovery or regression between PGOT and an ablation	e.g., $N_{C\to S} = \sum_j \mathbb{I}[A_j = C, F_j = S]$; maneuver-linked counts additionally require the terminal-turn event

CoNav-UAV validation diagnostics evaluate the final checkpoints under teacher forcing on the full validation split, with Criticality quartiles grouping samples by h_n. Online comparisons match task and trial indices. Local motion statistics use jointly successful pairs with valid traces. Paired outcome counts include all evaluated attempts. Relative reductions compare PGOT with the corresponding ablation, weighting eligible episodes equally.

LIBERO-10 diagnostics. For executed translation commands, define $\mathbf{u}_t = (a_{t,x}/R_x, a_{t,y}/R_y, a_{t,z}/R_z)$, where R_d is the tokenizer range of dimension d. The XYZ reversal rate is the fraction of consecutive command pairs with angle greater than 120°; both norms must exceed $0.5/255$. An episode without eligible direction pairs has reversal rate zero. Comparisons match task, initial state, and evaluation round, retain executions completed successfully by both policies, and average episodes equally.

Teacher-forcing diagnostics. The teacher, PGOT, and the three ablations are evaluated under teacher forcing, as in training, on the full training set of each domain. For Geometry, student probability on ordered dimensions is summed by bin distance to the expert and averaged over positions. For Criticality, samples are grouped by their weight h_n (below the median, from the median to the $Q_{.95}$ cap, and at the cap), and the expected physical error of Table 3 is averaged within each group. For Reliability, at an ordered action position (n, d) whose teacher mode $\hat{a}_{n,d} = \arg\max_i p^T_{n,d,i}$ differs from the expert bin $y_{n,d}$, the teacher lean of a student is

$$\ell_{n,d} = \frac{p^S_{n,d,\hat{a}_{n,d}} - p^S_{n,d,y_{n,d}}}{p^S_{n,d,\hat{a}_{n,d}} + p^S_{n,d,y_{n,d}}}, \tag{13}$$

which is zero when the student weighs the two bins equally and positive when it favors the teacher's choice. Positions are grouped by the bin distance $|\hat{a}_{n,d} - y_{n,d}|$, and each group reports the mean of $\ell_{n,d}$ in percent. For the online strata, a configuration counts as teacher-reliable when the Teacher SFT policy succeeds in all three executions, which holds for 118 of the 150 CoNav-UAV tasks and 186 of the 500 LIBERO-10 initial states; Pass^3 is computed within each stratum as in Equation 10.

For expert-frame switch prevalence, both frames straddling a gripper state change are marked as switch frames. The two-frame neighborhood additionally includes two adjacent frames on each side within the same episode, counting each frame once. For each frame set $\mathcal{S}$ (all frames or $q \geq Q_{.95}$), prevalence is $100|\mathcal{S} \cap \mathcal{G}|/|\mathcal{S}|$, with $\mathcal{G}$ denoting switch frames or their neighborhoods. The action-change score q and threshold $Q_{.95}$ follow Equation 4.

E CoNav-UAV Protocol and Supplementary Results

E.1 Training and action interface

The CoNav-UAV dataset contains low-altitude expert trajectories (Song et al., 2026). A PPO expert navigates a geometric representation of the scene built from point clouds. Its waypoint sequences are filtered for spatial diversity and replayed in AirSim to render FPV images; consecutive waypoint differences supply the expert actions. PGOT uses the `airsim16` data at a fixed 15 m altitude.

The training split contains 3,860 trajectories and 250,475 observation–action samples; validation contains 429 trajectories and 27,503 samples. Trajectories are disjoint across the two splits. Each sample pairs one FPV image with the current planar pose (x, y, ψ) and target coordinates (x_g, y_g). These fields are serialized as the model's visual and textual context, with three discretized action tokens as the supervised response.

The teacher is an action-aligned Qwen2.5-VL-7B model and the student is an action-aligned Qwen3.5-0.8B model. The three action tokens represent $(\Delta x, \Delta y, \Delta\psi)$. Their tokenizer ranges are $[-5.9994, 5.9976]$ m, $[-5.9998, 5.9997]$ m, and $[-0.1234, 0.1302]$ rad. All three use ordered 256-bin supports.

All three predicted deltas use linear decoded distances. For Criticality, the positive-change 95th percentile is computed separately for each dataset split: training weights use the training scale, and validation diagnostics use the validation scale.

E.2 Evaluation and action decoding

Single-UAV evaluation in `airsim16` uses a fixed 15 m flight height, a 20 m target radius, and at most 200 control steps. The suite contains 150 start–goal tasks, with 30 targets in each initial-distance band: 0–100, 100–200, 200–300, 300–400, and 400–500 m. Each method is run three times on every target, for 450 attempts per setting. An attempt ends in success, collision, or timeout.

We evaluate collision thresholds of 0.5 and 3.0 m. For each action waypoint, collision is evaluated against the closest non-target obstacle contour; the target building and designated noise points are excluded.

E.3 DISTANCE-STRATIFIED NAVIGATION

Table 5: **Distance-stratified CoNav-UAV performance at the 0.5 m collision threshold.** Each band contains 30 tasks with three executions per task. Values are percentages; SR is mean ± sample standard deviation across three rounds. Bold marks the best student value for each metric and distance band.

		Initial navigation distance (m)				
Policy	**Metric**	**0–100**	**100–200**	**200–300**	**300–400**	**400–500**
Teacher SFT	**SR** ↑	88.89 ± 1.92	83.33 ± 3.33	90.00 ± 6.67	86.67 ± 3.33	88.89 ± 5.09
	Pass^3 ↑	86.67	73.33	73.33	80.00	80.00
Student SFT	**SR** ↑	75.56 ± 7.70	67.78 ± 6.94	42.22 ± 5.09	41.11 ± 1.92	40.00 ± 6.67
	Pass^3 ↑	63.33	53.33	26.67	26.67	23.33
FKL	**SR** ↑	84.44 ± 1.92	**85.56 ± 1.92**	87.78 ± 3.85	86.67 ± 3.33	82.22 ± 1.92
	Pass^3 ↑	76.67	**80.00**	76.67	70.00	70.00
RKL	**SR** ↑	85.56 ± 1.92	81.11 ± 1.92	77.78 ± 5.09	68.89 ± 3.85	76.67 ± 3.33
	Pass^3 ↑	80.00	76.67	66.67	50.00	66.67
ABKD	**SR** ↑	85.56 ± 3.85	**85.56 ± 3.85**	70.00 ± 3.33	84.44 ± 1.92	78.89 ± 1.92
	Pass^3 ↑	76.67	76.67	53.33	70.00	60.00
ToDi	**SR** ↑	87.78 ± 1.92	78.89 ± 5.09	73.33 ± 3.33	65.56 ± 9.62	72.22 ± 5.09
	Pass^3 ↑	83.33	70.00	56.67	43.33	56.67
Shallow-π	**SR** ↑	88.89 ± 1.92	**85.56 ± 1.92**	77.78 ± 8.39	88.89 ± 3.85	85.56 ± 3.85
	Pass^3 ↑	83.33	76.67	60.00	76.67	73.33
VLA-AD	**SR** ↑	90.00 ± 5.77	74.44 ± 1.92	70.00 ± 6.67	77.78 ± 1.92	62.22 ± 7.70
	Pass^3 ↑	83.33	63.33	50.00	60.00	33.33
PGOT	**SR** ↑	**92.22 ± 1.92**	81.11 ± 1.92	**88.89 ± 5.09**	**90.00 ± 5.77**	**87.78 ± 8.39**
	Pass^3 ↑	**86.67**	**80.00**	**83.33**	**80.00**	**80.00**

E.4 ADDITIONAL MATCHED ROUTES

Figure 6 extends Figure 2 to eight tasks, four below 300 m and four above 350 m. Each panel keeps all three trials of PGOT, FKL, and Shallow-π and highlights the first matched trial in which PGOT succeeds and both baselines fail.

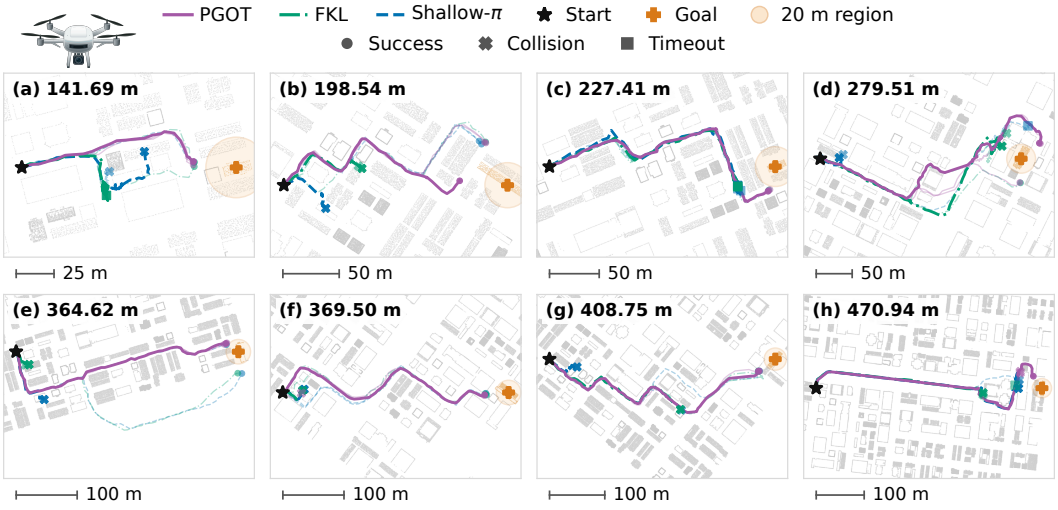


Figure 6: **Matched UAV routes for eight tasks.** Panels give the initial start–goal distance. Dark routes highlight one matched trial where PGOT succeeds and both baselines fail; light routes show the other two trials. Scale bars indicate meters.

E.5 Repeated-execution outcomes

Figure 7 shows, for each initial configuration, how many of the three executions succeed. In both domains PGOT has the smallest share of configurations with one or two successes, 11.33% on CoNav-UAV and 8.60% on LIBERO-10, against 18.00% and 11.20% for FKL, 24.00% and 30.00% for ABKD, and 18.67% and 30.40% for Shallow-π.

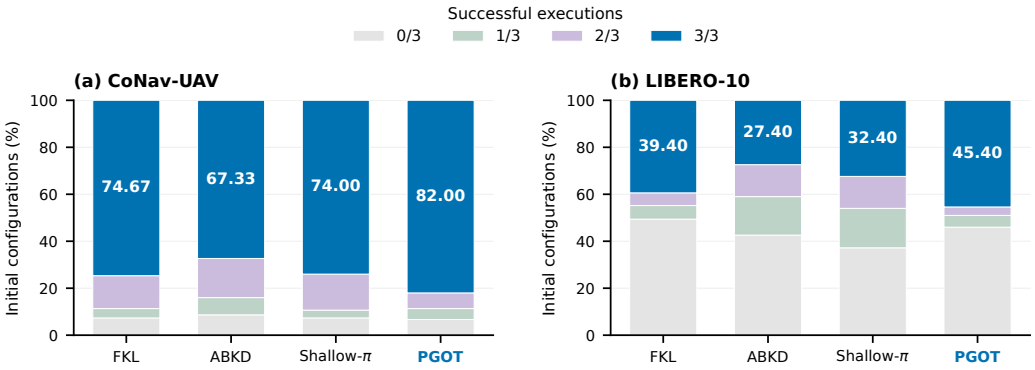


Figure 7: **Distribution of successful executions per initial configuration.** Numbers inside bars give Pass^3 (3/3 successes).

E.6 Collision-threshold sensitivity

Table 6: Online evaluation at the 3.0 m collision threshold. Metrics follow Table 1.

Policy	SR ↑	Pass^3 ↑	CR ↓	TR ↓	MMBC (m) ↑
Teacher SFT	70.22 ± 4.44	51.33	28.44	1.33	5.27
Student SFT	33.78 ± 4.23	22.00	64.67	**1.56**	3.69
FKL	68.67 ± 2.91	55.33	29.11	2.22	5.57
RKL	61.33 ± 1.33	48.67	35.11	3.56	5.27
ABKD	63.11 ± 1.39	44.67	33.33	3.56	5.43
ToDi	63.11 ± 2.04	50.67	31.56	5.33	5.40
Shallow-π	68.22 ± 1.39	54.67	30.00	1.78	5.35
VLA-AD	55.56 ± 2.04	38.67	42.00	2.44	4.74
PGOT	$\mathbf{73.33 \pm 1.33}$	**63.33**	**24.67**	2.00	**5.82**

At the 3.0 m collision threshold, PGOT attains the highest SR and Pass^3 and the lowest CR (Table 6). Compared with FKL, ABKD, and Shallow-π, PGOT improves SR by 4.67, 10.22, and 5.11 percentage points and Pass^3 by 8.00, 18.67, and 8.67 points, respectively.

F LIBERO-10 Experimental Protocol

F.1 Training and action interface

Robotic manipulation uses a Qwen3.5-9B teacher and Qwen3.5-0.8B students. Each sample pairs a single agent-view RGB observation with a natural-language task instruction. Training settings follow Section 4.1.

Each response has seven 256-bin action dimensions: Cartesian translation (x, y, z), orientation $(\mathrm{roll}, \mathrm{pitch}, \mathrm{yaw})$, and gripper state. The first six are ordered physical supports and receive the physical transport loss; the gripper is treated as categorical.

F.2 Evaluation and action decoding

Evaluation uses three rollouts from each of the 50 official initial states for every task, giving 1,500 rollouts per policy. Each rollout begins with ten settling steps and uses at most 520 policy decisions, a 90%-area center crop of the rotated agent-view image, and termination at first success. The simulator seed is zero.

Sampled action tokens are mapped to physical commands; the gripper uses the native simulator convention, -1 for open and $+1$ for close.

F.3 Paired repeated completion

Table 7 compares Pass^3 status configuration by configuration. PGOT solves more configurations in all three executions that its comparator does not than the reverse, for every baseline and ablation.

Table 7: Paired LIBERO-10 three-out-of-three completion over 500 configurations. Counts refer to Pass^3 status. Gains are PGOT minus the comparator in percentage points.

Comparator	Both	Neither	PGOT only	Other only	Gain (pp)
FKL	129	205	98	68	6.00
ABKD	89	225	138	48	18.00
Shallow-π	97	208	130	65	13.00
w/o Geometry	95	219	132	54	15.60
w/o Criticality	134	195	93	78	3.00
w/o Reliability	94	196	133	77	11.20

G Additional Training-Gradient Analysis

G.1 Relative changes during training

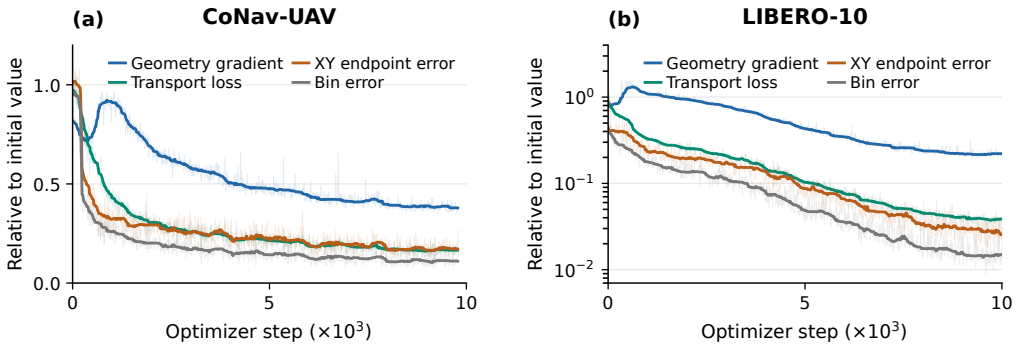


Figure 8: **Training dynamics relative to the initial value.** The LIBERO-10 axis is logarithmic; curves follow Figure 3.

Figure 8 shows the effective Geometry gradient RMS, transport loss, modal XY endpoint error, and mean absolute bin error, each as a fraction of its initial value. In both domains the transport gradient weakens together with the transport loss and prediction errors, consistent with a physical correction that diminishes as the student learns the action distribution.

G.2 CoNav-UAV training-gradient geometry

Gradients are measured on the student's action logits after applying all objective weights. Table 8 summarizes their strength and direction in early, middle, and late epochs of Figure 3(a,b).

Table 8: Gradient geometry by training phase. Ratios use the effective action-logit gradients after Criticality, Reliability, and objective weights.

Phase	$\|g_{\mathrm{OT}}\|/\|g_{\mathrm{filt}}\|$	Cosine	Rotation	Norm gain
Early (E1)	0.209	0.198	11.11°	6.17%
Middle (E2–E3)	0.150	0.277	7.88°	5.15%
Late (E4–E5)	0.123	0.314	6.43°	4.53%

For the phase summaries, let g_f and g_o denote the effective Filtered FKL and transport gradients, and write $g_k = g_f + g_o$. Cosine is $\cos(g_f, g_o)$ and rotation is $\angle(g_f, g_k)$. Norm gain is $100(\|g_k\|/\|g_f\| - 1)$. Entries average these quantities within each epoch group.

G.3 LIBERO-10 TRAINING-GRADIENT MEASUREMENTS

Figure 3(c,d) uses the same weighted Geometry and Filtered FKL terms (Equation 7), evaluated on the action logits of all seven dimensions; the Geometry gradient is zero on the categorical gripper.

H GEOMETRY-WEIGHT SENSITIVITY

We evaluate Geometry weights $w_1 \in \{5, 10, 20\}$ on CoNav-UAV and LIBERO-10 (Table 9), using the primary 0.5 m collision threshold for CoNav-UAV. The default $w_1 = 10$ achieves the highest SR and Pass^3 in both domains. Halving/doubling the weight reduces SR by 5.56/4.89 percentage points on CoNav-UAV and 6.47/7.60 points on LIBERO-10, with corresponding Pass^3 reductions of 10.67/6.67 and 7.80/7.40 points. The default also yields lower collision and timeout rates and greater building clearance on CoNav-UAV, and fewer execution steps on LIBERO-10.

Table 9: **Geometry-weight sensitivity.** Bold marks the best value in each column.

	CoNav-UAV					LIBERO-10		
w_1	SR ↑	Pass^3 ↑	CR ↓	TR ↓	MMBC ↑	SR ↑	Pass^3 ↑	MES ↓
5	82.44 ± 0.77	71.33	11.33	6.22	5.41	43.00 ± 1.04	37.60	416.31
10 (default)	$\mathbf{88.00 \pm 2.91}$	**82.00**	**8.00**	**4.00**	**5.72**	$\mathbf{49.47 \pm 0.64}$	**45.40**	**400.47**
20	83.11 ± 1.39	75.33	11.56	5.33	5.46	41.87 ± 0.92	38.00	420.43