AEGIS: ACTION-ENDPOINT GEOMETRY FOR NUISANCE-INVARIANT AND LOCALLY STATE-EQUIVARIANT VLA MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust vision-language-action (VLA) models must remain stable under task-equivalent language and task-irrelevant visual changes while adapting their actions to control-relevant changes in robot state, object configuration, or task progress. Per-sample mean squared error (MSE) supervision anchors individual predictions to demonstrations, but its fixed quadratic metric does not adapt residual weighting or explicitly constrain how predictions should relate across these conditions. We present AEGIS, a demonstration-anchored action-endpoint geometry for VLA models that predict action chunks from single-timestep observations. It combines three complementary objectives in a shared action field: Bounded Heteroscedastic Flow Matching (BHFM) uses bounded heteroscedastic scales to adapt residual weighting; Nuisance Endpoint Invariance (NEI) encourages endpoint agreement under action-preserving language and visual variations; and State Endpoint Equivariance (SEE) encourages local endpoint displacements to match demonstrated action differences across progress-matched states. AEGIS achieves 91.58% success on LIBERO-Plus and 28.24% on RoboCasa365, improving over the $\pi_{0.5}$ baseline by 6.74 and 7.24 percentage points, respectively. Across three real-world manipulation tasks, it achieves 57.33% average success, a gain of 12.33 percentage points over the baseline. These results demonstrate improved task performance in both simulated and physical environments.

1 INTRODUCTION

Vision-language-action (VLA) models map images, instructions, and robot state to temporally extended actions. Action-chunk transformers, diffusion policies, open VLA models, and flow-based action experts have expanded their capabilities (Zhao et al., 2023; Chi et al., 2023; Kim et al., 2025b; Black et al., 2025). Reliable deployment requires selective responses: within the same task, meaning-preserving paraphrases and task-irrelevant image changes should preserve actions, whereas changes in gripper configuration, object pose, robot–object geometry, or task progress that alter the required control should change predictions. Otherwise, models may react to irrelevant perturbations or overlook control-relevant state information.

We use $\pi_{0.5}$ (Physical Intelligence et al., 2025) as the baseline and retain its single-observation action-generation interface.

Challenge 1: A fixed residual metric treats heterogeneous action errors uniformly. Continuous arm motion, gripper actuation, and contact transitions can differ in prediction difficulty across action coordinates and chunk positions. Coordinate normalization does not ensure equal conditional residual scales, motivating adaptive residual weighting beyond the fixed quadratic metric of pure MSE.

Challenge 2: Per-sample supervision leaves connected conditions unstructured. Connected conditions require equal predictions for action-preserving changes and structured differences for action-changing states. Existing consistency objectives organize selected views (Luo et al., 2026; Huang et al., 2026), yet a uniform equality constraint is inappropriate for action-changing state differences.

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.

Insight: Action endpoints should remain unchanged under nuisance variations but shift with control-relevant state changes. We anchor endpoints to demonstrations, enforce equality for action-preserving changes, and match demonstrated displacement for action-changing states. The anchor is necessary because equality admits constant solutions and displacement specifies only relative offsets.

AEGIS implements this view with one objective over the clean-action endpoint implied by a shared flow field. BHFM supplies a bounded, demonstration-grounded anchor; NEI enforces endpoint agreement for action-preserving language and visual changes; and SEE preserves demonstrated displacement for progress-matched states requiring different controls. Figure 1 summarizes the evaluated input variations and endpoint relations.

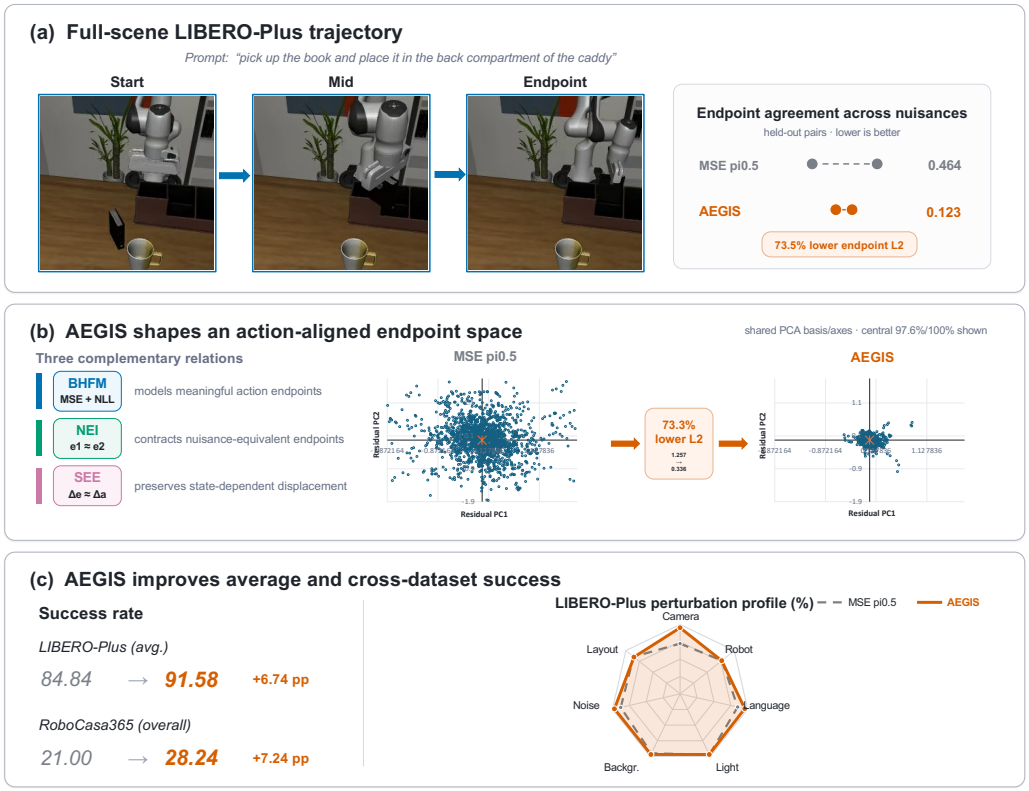


Figure 1: Endpoint geometry of AEGIS. (a) Rollout and nuisance agreement. (b) Training relations and endpoint residuals in a shared PCA basis. (c) Benchmark success and robustness versus $\pi_{0.5}$.

We evaluate all eight combinations of BHFM, NEI, and SEE on the $\pi_{0.5}$ baseline in LIBERO-Plus, and compare complete AEGIS with $\pi_{0.5}$ on RoboCasa365 and three real-world tasks. Held-out diagnostics test nuisance agreement, state-dependent displacement, semantic selectivity under the joint nuisance test, and dependence on meaningful state correspondence.

Our contributions are:

- BHFM provides a bounded heteroscedastic anchor that retains MSE demonstration supervision while adapting element-wise residual weighting.
- NEI enforces endpoint agreement for paired language and visual changes that preserve the demonstrated action.
- SEE matches endpoint displacement to demonstrated action displacement across progress-matched states, yielding a local empirical state relation.

2 Related Work

Vision-language-action models and action fields. Action-chunk transformers and diffusion policies predict temporally extended controls; RT-2, Octo, and OpenVLA transfer heterogeneous robot data or pretrained vision-language representations to control (Zhao et al., 2023; Chi et al., 2023; Zitkovich et al., 2023; Ghosh et al., 2024; Kim et al., 2025b). Recent robot action generators use continuous transport via flow matching (Lipman et al., 2023; Black et al., 2025; Physical Intelligence et al., 2025; Zhang et al., 2025). Heteroscedastic regression learns supervised residual scales (Kendall et al., 2018). AEGIS changes action-field supervision; Section 3 formalizes its endpoint relations.

Robust and relational learning for robot control. Visual randomization improves appearance robustness (Tobin et al., 2017), and VLA consistency relates instructions and views (Luo et al., 2026; Huang et al., 2026). Equivariant manipulation models structure action transformations (Yang et al., 2025; Tie et al., 2025; Huang et al., 2024) using geometric principles (Cohen & Welling, 2016; Bronstein et al., 2017), while behavior-aware representations organize observations by control relevance (Agarwal et al., 2021). Gripper actuation, contact changes, and task progress need not follow a common rigid scene transformation. AEGIS thus uses demonstrated action differences as local endpoint-displacement targets for progress-matched states, alongside endpoint-equality targets for action-preserving pairs (Section 3.4).

3 Method

AEGIS learns a demonstration-anchored action-endpoint geometry. A bounded anchor (BHFM) and two pair-dependent relations (NEI and SEE) specify how endpoints are grounded, shared, or displaced. Figure 2 shows how these terms act on one shared action field.

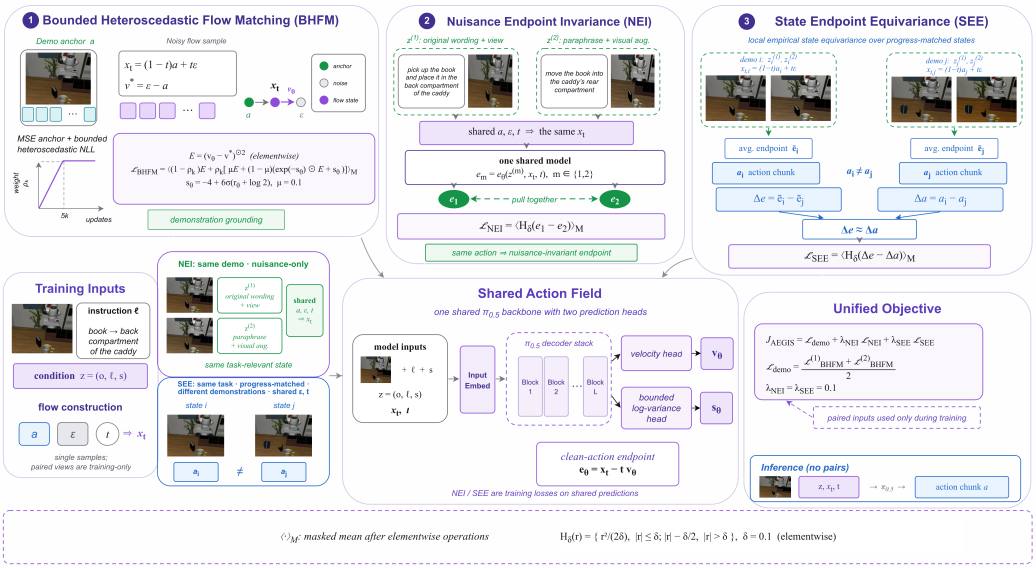


Figure 2: Overview of AEGIS. BHFM anchors actions with bounded residual weighting; NEI aligns endpoints across action-preserving views; SEE matches demonstrated displacements across progress-matched states. The three losses share one action field, with paired inputs used only during training.

3.1 Problem Formulation

Let $z = (o, \ell, s)$ denote the visual observations o, language instruction ℓ, and robot state s available at one observation time step, and let $a \in \mathbb{R}^{H \times D}$ be a demonstrated action chunk. Following the

flow-based action generation used by $\pi_{0.5}$ (Black et al., 2025; Physical Intelligence et al., 2025), we sample $\epsilon \sim \mathcal{N}(0, I)$ and

$$t = 0.001 + 0.999u, \quad u \sim \mathrm{Beta}(1.5, 1), \qquad x_t = (1-t)a + t\epsilon, \tag{1}$$

with target velocity $v^* = \epsilon - a$. The network $v_\theta(z, x_t, t)$ predicts this velocity. Writing the element-wise squared velocity error as $E = (v_\theta - v^*)^2$, the pure-MSE baseline minimizes

$$L_{\mathrm{MSE}} = \langle E \rangle_M, \tag{2}$$

where $\langle \cdot \rangle_M$ averages over valid physical action coordinates. This objective fits each single-timestep condition to its demonstrated action chunk independently.

The same velocity prediction defines a clean-action endpoint in controller action coordinates,

$$e_\theta(z, x_t, t) = x_t - t\, v_\theta(z, x_t, t). \tag{3}$$

When $v_\theta = v^*$, the endpoint equals the demonstration, $e_\theta = a$. Pure MSE therefore supplies an absolute demonstration anchor, but it leaves two aspects of the endpoint geometry unspecified. Its fixed quadratic form prescribes one residual weighting across action elements, and its per-sample supervision does not state whether endpoints from connected conditions should coincide or differ with the required control.

AEGIS addresses these omissions with a bounded anchor and two pair-dependent relations. For paired nuisance views, both predictions retain demonstration-grounded supervision:

$$L_{\mathrm{demo}} = \frac{L^{(1)}_{\mathrm{BHFM}} + L^{(2)}_{\mathrm{BHFM}}}{2}. \tag{4}$$

The complete training objective is

$$\mathcal{J}_{\mathrm{AEGIS}} = L_{\mathrm{demo}} + \lambda_{\mathrm{NEI}} L_{\mathrm{NEI}} + \lambda_{\mathrm{SEE}} L_{\mathrm{SEE}}, \qquad \lambda_{\mathrm{NEI}} = \lambda_{\mathrm{SEE}} = 0.1. \tag{5}$$

Here BHFM anchors the action field to demonstrations, NEI constrains action-preserving pairs, and SEE constrains action-changing pairs. All applicable terms are computed from shared predictions and optimized in one update. These relations exist only during training: alternate views, pair metadata, and shared random variables are discarded at deployment, and the trained checkpoint retains the observation interface and inference-time action-generation paradigm of $\pi_{0.5}$.

3.2 Bounded Heteroscedastic Flow Matching

Reliable endpoint relations require each prediction to remain grounded in its demonstration. Pure MSE provides this anchor but applies the same quadratic residual form to every valid action element. BHFM retains the supervised MSE target while learning bounded element-wise residual weighting. It predicts a raw scale r_θ together with the velocity, maps it smoothly to the bounded log-scale interval $[s_{\min}, s_{\max}] = [-4, 2]$, and defines an element-wise heteroscedastic negative-log-likelihood term (Kendall et al., 2018):

$$s_\theta = s_{\min} + (s_{\max} - s_{\min})\sigma(r_\theta + \log 2), \qquad L_{\mathrm{NLL}} = \exp(-s_\theta)E + s_\theta. \tag{6}$$

Here, $\sigma(u) = (1 + \exp(-u))^{-1}$ is the logistic sigmoid and $\log$ denotes the natural logarithm. The offset maps $r_\theta = 0$ to $s_\theta = 0$. With $\mu = 0.1$, a persisted counter k ramps from MSE to $R = \mu E + (1-\mu)L_{\mathrm{NLL}}$ over $K = 5{,}000$ updates:

$$L_{\mathrm{BHFM}} = \left\langle (1-\rho_k)E + \rho_k \left[\mu E + (1-\mu)L_{\mathrm{NLL}}\right] \right\rangle_M, \qquad \rho_k = \mathrm{clip}(k/K, 0, 1). \tag{7}$$

At $\rho_k = 0$, Equation 7 recovers pure MSE; at $\rho_k = 1$, it combines 0.1 MSE with 0.9 bounded NLL. The effective coefficient on E is $\mu + (1-\mu)\exp(-s_\theta)$, which cannot vanish or diverge because $\mu > 0$ and s_θ is bounded. The ramp avoids abruptly activating the scale head. BHFM thus grounds the velocity field underlying the endpoints without imposing an endpoint relation by itself.

When NLL is active, increasing the log scale at fixed velocity error reduces the weighted squared error but raises the additive log-scale cost. This trade-off learns residual weighting within the supervised objective. The bounded map restricts weights, while retained MSE keeps every valid action coordinate anchored to its demonstration despite differing weights.

3.3 NUISANCE ENDPOINT INVARIANCE

Demonstration grounding does not determine how predictions should respond when language or visual input changes without changing the required action. For such cases, NEI constructs a nuisance-equivalent pair $z^{(1)} \sim_{\mathcal{N}} z^{(2)}$ from two views of the same demonstration. Robot configuration, task-relevant state, demonstrated action, flow noise, and time are shared, while the instruction form or task-irrelevant visual preprocessing may differ. The desired relation is

$$z^{(1)} \sim_{\mathcal{N}} z^{(2)} \quad \Longrightarrow \quad e_\theta(z^{(1)}, x_t, t) = e_\theta(z^{(2)}, x_t, t). \tag{8}$$

For endpoints e_1 and e_2, NEI realizes this relation with

$$L_{\mathrm{NEI}} = \langle H_\delta(e_1 - e_2) \rangle_M , \qquad \delta = 0.1, \tag{9}$$

where $H_\delta(r) = \frac{r^2}{2\delta}$ for $|r| \le \delta$ and $|r| - \frac{\delta}{2}$ otherwise (Huber, 1964). The primary view preserves the configuration's language marginal. For the LIBERO-Plus experiments, nuisance pairs use distinct prompts drawn from the canonical instruction and task-equivalent variants generated offline by GPT-5.5. For RoboCasa, the paired prompt applies one of three fixed wrappers to the original task text. Visual augmentations are sampled during training with independent randomness for the two views. Sharing robot state, action, flow time, and noise isolates the action-preserving change in the condition.

The endpoint constraint also has a direct residual interpretation. Because both views share x_t, t, and target velocity v^*, letting $\Delta v_k = v_k - v^*$ gives

$$e_1 - e_2 = -t(v_1 - v_2) = -t(\Delta v_1 - \Delta v_2). \tag{10}$$

Thus NEI is a time-scaled robust consistency penalty on velocity residuals expressed in common action-space coordinates. Endpoint equality is appropriate for these action-preserving pairs, but extending it to states that require different controls would erase useful variation.

3.4 STATE ENDPOINT EQUIVARIANCE

SEE defines a local empirical displacement relation between progress-matched samples i and j drawn from distinct trajectories of the same task. Their robot configurations or object poses may differ, and the target geometry is

$$(i, j) \in \mathcal{P}_{\mathcal{S}} \quad \Longrightarrow \quad e_{\theta,i} - e_{\theta,j} = a_i - a_j. \tag{11}$$

Each state retains two nuisance-equivalent views, averaged as $\bar{e}_i = (e_i^{(1)} + e_i^{(2)})/2$ and likewise for j. Pair members share flow noise and time, while each flow state uses its own action. We minimize

$$L_{\mathrm{SEE}} = \langle H_\delta[(\bar{e}_i - \bar{e}_j) - (a_i - a_j)] \rangle_M . \tag{12}$$

Let $\Delta v_i = \bar{v}_i - v_i^*$ be the view-averaged velocity prediction error. Equation 3 gives $\bar{e}_i = a_i - t\Delta v_i$, and hence

$$(\bar{e}_i - \bar{e}_j) - (a_i - a_j) = -t(\Delta v_i - \Delta v_j). \tag{13}$$

SEE penalizes time-scaled differences between view-averaged velocity errors, encouraging endpoint displacements to match demonstrated action differences. Progress matching reduces phase mismatch, and the Huber penalty bounds the influence of imperfect pairs. A state-pair update evaluates two views of each state, allowing the same endpoints to participate in both relational constraints. The resulting relation is local to the constructed pair distribution rather than a global group action.

Equations 10 and 13 expose the common structure of the two relations: both penalize time-scaled differences in velocity prediction errors. Nuisance pairs share an action target, while state pairs retain their own targets and subtract the demonstrated displacement. In the latter case, the errors are averaged over the nuisance views of each state. An error shared by both pair members cancels in either relational residual. The demonstration anchor therefore constrains absolute prediction error, while pair supervision constrains how that error varies across conditions. Sharing flow time and noise holds the flow draw fixed across the pair.

4 EXPERIMENTS

We evaluate AEGIS on LIBERO-Plus, RoboCasa365, and three physical manipulation tasks, and use endpoint diagnostics to study how the training objectives organize predicted actions.

4.1 EXPERIMENTAL SETUP

Datasets and metrics. LIBERO-Plus extends LIBERO's language-conditioned manipulation suites with seven perturbation categories: camera viewpoint, robot initial state, language instruction, lighting condition, background texture, sensor noise, and object layout (Liu et al., 2023; Fei et al., 2025). We evaluate 40 tasks from the Spatial, Object, Goal, and LIBERO-10 suites on the fixed 10,030-episode evaluation plan. We report success rate for each category and the benchmark-weighted aggregate, which is not the unweighted mean of the seven category rates. RoboCasa365 extends RoboCasa's randomized household simulation to large-scale multi-task manipulation (Nasiriany et al., 2024; 2026). Models are trained on 300 tasks and evaluated on 50 tasks: 18 Atomic-Seen, 16 Composite-Seen, and 16 Composite-Unseen tasks, with 50 trials per task.

Implementation Details. AEGIS follows the $\pi_{0.5}$ architecture, coupling a SigLIP vision encoder with a Gemma-2B language backbone of hidden size 2048 and a Gemma action expert of hidden size 1024 (Physical Intelligence et al., 2025). For configurations that use BHFM, the scale head defined in Section 3.2 adds 270,624 parameters (approximately 0.01%). Training uses AdamW (Loshchilov & Hutter, 2019) with a global batch size of 256. The peak learning rate is 1×10^{-5} for LIBERO-Plus and 5×10^{-5} for RoboCasa365. Models for both simulation and real-world experiments are trained on eight NVIDIA H200 GPUs. LIBERO-Plus and RoboCasa365 evaluations also use eight H200 GPUs, while real-world evaluation runs on one NVIDIA RTX 4090 GPU. Tables 1 and 3 report benchmark success rates for our $\pi_{0.5}$ baseline and all variants as mean $\pm$ sample standard deviation over three runs (seeds 7, 17, and 27), with benchmark-weighted *Avg.* computed per run before aggregation.

Table 1: Success rates (%) on LIBERO-Plus (upper) and RoboCasa365 (lower). Our pure-MSE $\pi_{0.5}$ and AEGIS results show mean $\pm$ sample SD over seeds 7, 17, and 27 (SD below in the upper panel); published methods retain reported values. *Avg.* is benchmark-weighted; bold marks the higher internal mean.

LIBERO-Plus **Method**	**Cam.**	**Robot**	**Lang.**	**Light**	**Bkg.**	**Noise**	**Layout**	**Avg.**
OpenVLA (Kim et al., 2025b)	0.80	3.50	23.00	8.10	34.80	15.20	28.50	15.60
OpenVLA-OFT (Kim et al., 2025a)	56.40	31.90	79.50	88.70	93.30	75.80	74.20	69.60
WorldVLA (Cen et al., 2025)	0.10	27.90	41.60	43.70	17.10	10.90	38.00	25.00
UniVLA (Bu et al., 2025)	1.80	46.20	69.60	69.00	81.00	21.20	31.90	42.90
StarVLA (StarVLA Community, 2026)	52.50	49.80	88.50	95.70	95.70	73.00	76.90	74.10
π_0 (Black et al., 2025)	13.80	6.00	58.80	85.00	81.40	79.00	68.90	53.60
π_0-FAST (Pertsch et al., 2025)	65.10	21.60	61.00	73.20	73.20	74.40	68.80	61.60
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	72.42 $\pm$0.26	**77.10** $\pm$1.06	85.10 $\pm$1.13	**97.37** $\pm$0.35	95.07 $\pm$0.62	87.76 $\pm$0.35	**85.77** $\pm$1.29	84.84 $\pm$0.51
AEGIS	**95.31** $\pm$1.36	76.77 $\pm$0.26	**95.64** $\pm$0.88	96.94 $\pm$0.47	**96.93** $\pm$0.30	**97.06** $\pm$0.54	85.05 $\pm$0.93	**91.58** $\pm$0.24

RoboCasa365 **Method**	**Atomic Seen**	**Composite Seen**	**Composite Unseen**	**Avg.**
$\pi_{0.5}$	44.11$\pm$1.75	13.38$\pm$0.99	2.63$\pm$0.33	21.00 $\pm$0.42
AEGIS	**52.78**$\pm$0.88	**19.00**$\pm$1.02	**9.88**$\pm$0.45	**28.24** $\pm$0.79

4.2 QUANTITATIVE RESULTS

AEGIS improves robustness on LIBERO-Plus. Complete AEGIS raises mean aggregate success from $84.84 \pm 0.51\%$ for $\pi_{0.5}$ to $91.58 \pm 0.24\%$ (a 6.74-point gain). The strongest improvements occur under camera, language, and sensor-noise perturbations; Table 1 reports the category-level values.

The complete objective improves a second task distribution. On RoboCasa365, AEGIS improves mean aggregate success from $21.00 \pm 0.42\%$ for $\pi_{0.5}$ to $28.24 \pm 0.79\%$, a 7.24-point gain. Unseen compositions remain challenging: success reaches 9.88% on Composite-Unseen, compared with 19.00% on Composite-Seen and 52.78% on Atomic-Seen tasks (Table 1).

AEGIS improves aggregate success over the baseline in all three runs on both benchmarks. LIBERO-Plus varies observations, instructions, and scenes within fixed manipulation task families, revealing gains and regressions across perturbation categories. RoboCasa365 evaluates separately trained models on atomic and composite household tasks, including seen and unseen compositions.

4.3 Real-World Experiment

We compare AEGIS and $\pi_{0.5}$ on *Pick the tomato*, *Build blocks*, and *Plug in the Ethernet cable*, covering object picking, block arrangement, and cable insertion (Figure 3). Each model undergoes 100 trials per task, totaling 300; Table 2 reports successful trials and mean success across tasks.

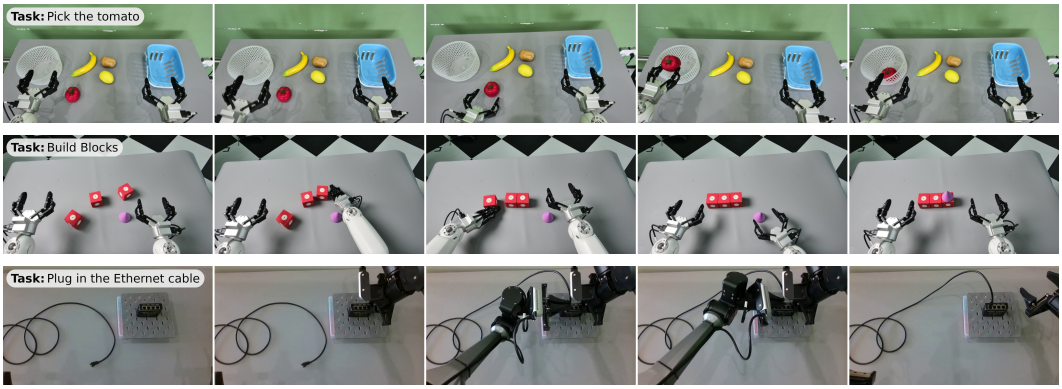


Figure 3: Real-world rollouts. Rows from top to bottom show Pick the tomato, Build blocks, and Plug in the Ethernet cable; frames progress from left to right.

Table 2: Results on three real-world tasks (Figure 3). Entries show successes/100 trials; *Avg.* reports success over all 300 trials per model (%). Bold marks the better result in each column.

Method	**Pick the tomato**	**Build blocks**	**Plug in the Ethernet cable**	**Avg.** (%)
		(successes / trials)		
$\pi_{0.5}$	82/100	33/100	20/100	45.00
AEGIS	**88/100**	**47/100**	**37/100**	**57.33**

AEGIS improves all three tasks, raising average success from 45.00% to 57.33% (+12.33 percentage points). The largest gains are 14 points for block building and 17 for Ethernet insertion, which remain challenging at 47% and 37% success, versus 88% for tomato picking. Equal trial counts give each task equal weight and make the task average equal to pooled success.

4.4 Ablation Study

Table 3 compares all eight combinations of BHFM, NEI, and SEE on LIBERO-Plus across three seeds. Each component alone improves on the baseline's 84.84% mean aggregate success: NEI reaches 87.10%, SEE 86.34%, and BHFM 87.94%. Combining NEI and SEE without BHFM reaches 88.86%.

Complete AEGIS achieves the highest mean aggregate success, $91.58 \pm 0.24\%$, followed by BHFM + NEI at $91.18 \pm 0.28\%$. Removing BHFM, NEI, or SEE lowers this result by 2.72, 1.45, or 0.40 percentage points, respectively. Adding NEI to BHFM improves average success from 87.94% to 91.18% and language success from 90.49% to 94.86%. The category results vary: SEE alone leads

Table 3: LIBERO-Plus component ablation of $\pi_{0.5}$ (Physical Intelligence et al., 2025): success rate (%; mean above, ± sample SD below; seeds 7, 17, 27). AEGIS uses all three components. *Avg.* is benchmark-weighted. Bold and underlining mark the best and second-best means at the displayed precision.

Configuration	Cam.	Robot	Lang.	Light	Bkg.	Noise	Layout	Avg.
$\pi_{0.5}$	72.42 ±0.26	77.10 ±1.06	85.10 ±1.13	**97.37** ±0.35	95.07 ±0.62	87.76 ±0.35	85.77 ±1.29	84.84 ±0.51
+ NEI	79.00 ±1.09	77.00 ±1.34	90.20 ±1.20	97.10 ±0.34	95.60 ±0.49	90.80 ±0.40	85.40 ±0.57	87.10 ±0.22
+ SEE	75.50 ±0.80	**80.60** ±0.18	85.80 ±0.78	97.10 ±0.44	95.10 ±0.59	88.80 ±0.71	**87.20** ±0.75	86.34 ±0.41
+ NEI + SEE	82.50 ±1.09	**80.60** ±0.45	91.20 ±1.11	97.10 ±0.48	95.90 ±0.52	92.30 ±1.02	86.80 ±1.01	88.86 ±0.49
+ BHFM	83.87 ±1.22	76.27 ±0.93	90.49 ±0.42	97.06 ±0.44	95.50 ±0.47	91.91 ±0.61	85.13 ±0.53	87.94 ±0.16
+ BHFM + NEI	<u>95.00</u> ±0.47	76.65 ±0.28	<u>94.86</u> ±0.86	<u>97.29</u> ±0.49	<u>96.19</u> ±0.58	<u>96.31</u> ±0.49	84.72 ±1.25	<u>91.18</u> ±0.28
+ BHFM + SEE	89.50 ±0.76	<u>79.00</u> ±0.98	91.80 ±0.39	96.95 ±0.45	96.10 ±0.56	94.10 ±0.56	<u>86.90</u> ±0.51	90.13 ±0.10
AEGIS (Full)	**95.31** ±1.36	76.77 ±0.26	**95.64** ±0.88	96.94 ±0.47	**96.93** ±0.30	**97.06** ±0.54	85.05 ±0.93	**91.58** ±0.24

on Layout and ties with NEI + SEE on Robot at the reported precision, while the baseline leads on Light. Full AEGIS remains below the baseline on Robot, Light, and Layout by 0.33, 0.43, and 0.72 points. Rollout success summarizes episode-level behavior; the following held-out diagnostics separately probe the action-endpoint relations targeted by the objective.

4.5 Endpoint Geometry Analysis

We examine $\pi_{0.5}$, BHFM, BHFM + NEI, and complete AEGIS on one held-out pair bank, then compare the frozen AEGIS checkpoint's predictions against intended and plausible alternative endpoint relations.

Evaluation protocol. We select 14 cross-episode progress-matched pairs from held-out data for each of 40 LIBERO-Plus tasks and evaluate four shared flow draws per pair, yielding 560 state pairs and 2,240 pair–draw residuals per objective. Across objectives, we fix pair identities, normalization, camera views, flow time and noise, augmentation seeds, action horizon, and active action dimensions. We measure

$$e_\theta = x_t - t\, v_\theta(z, x_t, t)$$

in normalized 10×7 active action-chunk coordinates. Nuisance endpoint L2 measures agreement for an unchanged demonstrated action. State diagnostics compare $e_i - e_j$ with $a_i - a_j$ using cosine similarity for direction, and L2 error and RMSE for direction and magnitude; aligned displacements can still be too large or small.

Action-preserving agreement. Table 4 reports a nuisance endpoint distance of 0.123 for AEGIS versus 0.464 for $\pi_{0.5}$, a 73.5% reduction consistent with the intended contraction of action-preserving language and visual conditions.

State-dependent displacement. On held-out state pairs, AEGIS achieves a cosine similarity of 0.98 and displacement L2 error of 0.34, versus 0.80 and 1.26 for $\pi_{0.5}$.

Even without paired losses, BHFM halves nuisance and state-displacement L2 errors. By Equations 10 and 13, residual weighting can improve endpoint geometry through changes in velocity-error differences. Relative to BHFM, the full objective lowers nuisance L2 from 0.23 to 0.12 and state

Table 4: Held-out LIBERO-Plus endpoint diagnostics. Nuisance L2 measures agreement across action-preserving views. Cosine, State L2, and RMSE compare predicted and demonstrated state-pair displacements in normalized action coordinates. Arrows indicate the preferred direction.

Method	Nuisance L2 ↓	Cosine ↑	State L2 ↓	RMSE ↓
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	0.46	0.80	1.26	0.15
BHFM	0.23	0.86	0.63	0.08
BHFM + NEI	0.17	0.92	0.47	0.06
AEGIS	0.12	0.98	0.34	0.04

L2 from 0.63 to 0.34, while mean rollout success across seeds rises from 87.94% to 91.58% (+3.64 points; Table 3).

Frozen-checkpoint target contrasts. Table 5 reports state-target and instruction contrasts for the frozen AEGIS checkpoint.

Table 5: Frozen AEGIS diagnostics on held-out LIBERO-Plus pairs. Row groups compare SEE/zero-displacement targets, matched/mismatched targets, and equivalent/semantically different instructions. Percentages report pair–draws with lower intended-target L2 error or positive semantic-minus-equivalent endpoint-distance margins.

Diagnostic metric	**Result**
SEE / zero-displacement target L2 ↓	0.34 vs. 2.82
SEE-preferred pair–draws ↑	95.54%
Matched / mismatched target L2 ↓	0.34 vs. 6.47
Matched / mismatched target cosine ↑	0.98 vs. 0.24
Matched-preferred pair–draws ↑	100.00%
Equivalent / semantic-different endpoint L2	0.12 vs. 0.86
Semantic-minus-equivalent margin ↑	0.74
Positive-margin pair–draws ↑	99.15%

Using fixed predictions from one frozen checkpoint, intended SEE targets are preferred to zero displacement on 95.54% of progress-matched pair–draws, and matched targets to mismatched targets on 100.00%. These contrasts test demonstrated displacement versus no change and dependence on the corresponding action difference. The instruction contrast tests semantic selectivity: equivalent and semantically different instructions yield endpoint L2 distances of 0.12 and 0.86, respectively.

Pair quality. In the state-pair audit, 39 of 40 pairs (97.50%) agreed at the broad task-stage level.

Progress matching provides comparable task context while allowing gripper configuration and object pose to differ. SEE uses each state's demonstrated action to define the displacement. The audit checks broad-stage correspondence; the displacement diagnostics test the resulting action relation.

5 Conclusion

AEGIS learns an anchored action-endpoint geometry that encourages endpoint agreement for action-preserving changes and demonstrated displacement for control-relevant state changes. The eight-way LIBERO-Plus ablation shows higher average success with each component alone and the highest average with the full objective, with category-dependent gains. Complete AEGIS also lowers held-out nuisance and state-displacement errors relative to the baseline and improves all three real-world tasks. Evaluation on a $\pi_{0.5}$-derived architecture and constructed local pairs motivates contact- and motion-aware pairing, temporal observations, broader VLA architectures and embodiments, and more physical manipulation tasks.

AI USE STATEMENT

During manuscript language polishing, we used generative AI tools, specifically Large Language Models (LLMs), only for minor linguistic refinement, such as correcting grammatical errors. Separately, GPT-5.5 was used to generate task-equivalent language variants offline for LIBERO-Plus. We reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G. Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qda7-sVg84. ICLR 2021 Spotlight.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy 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 *Robotics: Science and Systems XXI*, 2025. doi: 10.15607/RSS.2025.XXI.010. URL https://doi.org/10.15607/RSS.2025.XXI.010.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. doi: 10.1109/MSP.2017.2693418. URL https://doi.org/10.1109/MSP.2017.2693418.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. UniVLA: Learning to act anywhere with task-centric latent actions. In *Robotics: Science and Systems XXI*, 2025. URL https://arxiv.org/abs/2505.06111.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. WorldVLA: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025. URL https://arxiv.org/abs/2506.21539.

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

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999. PMLR, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

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

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 *Robotics: Science and Systems XX*, 2024. doi: 10.15607/RSS.2024.XX.090. URL https://doi.org/10.15607/RSS.2024.XX.090.

Bingqi Huang, Bingchuan Wei, Xuan Wang, Yingkai Cai, and Zhaokui Wang. Cross-view action consistency for camera-robust vision-language-action policies. *arXiv preprint arXiv:2608.06965*, 2026. URL https://arxiv.org/abs/2608.06965.

Haojie Huang, Owen Howell, Dian Wang, Xupeng Zhu, Robert Platt, and Robin Walters. Fourier transporter: Bi-equivariant robotic manipulation in 3D. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=UulwvAU1W0.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018. doi: 10.1109/CVPR.2018.00781. URL https://doi.org/10.1109/CVPR.2018.00781.

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

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. URL https://proceedings.mlr.press/v270/kim25c.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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. doi: 10.52202/075280-1939. URL https://doi.org/10.52202/075280-1939.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jingzhou Luo, Yifan Wen, Yongjie Bai, Xinshuai Song, Yang Liu, and Liang Lin. RoVLA: Multi-consistency constraints for robust vision-language-action models. *arXiv preprint arXiv:2605.19678*, 2026. URL https://arxiv.org/abs/2605.19678.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-scale simulation of household tasks for generalist robots. In *Robotics: Science and Systems XX*, 2024. doi: 10.15607/RSS.2024.XX.050. URL https://doi.org/10.15607/RSS.2024.XX.050.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. RoboCasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tQJYKwc3n4.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025. URL https://arxiv.org/abs/2501.09747.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. URL https://arxiv.org/abs/2504.16054.

StarVLA Community. StarVLA: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026. doi: 10.48550/arXiv.2604.05014. URL https://arxiv.org/abs/2604.05014.

Chenrui Tie, Yue Chen, Ruihai Wu, Boxuan Dong, Zeyi Li, Chongkai Gao, and Hao Dong. ET-SEED: Efficient trajectory-level SE(3) equivariant diffusion policy. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OheAR2xrtb.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133. URL https://doi.org/10.1109/IROS.2017.8202133.

Jingyun Yang, Ziang Cao, Congyue Deng, Rika Antonova, Shuran Song, and Jeannette Bohg. EquiBot: SIM(3)-equivariant diffusion policy for generalizable and data efficient learning. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1048–1068. PMLR, 2025. URL https://proceedings.mlr.press/v270/yang25a.html.

Qinglun Zhang, Zhen Liu, Haoqiang Fan, Guanghui Liu, Bing Zeng, and Shuaicheng Liu. FlowPolicy: Enabling fast and robust 3D flow-based policy via consistency flow matching for robot manipulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(14):14754–14762, 2025. doi: 10.1609/aaai.v39i14.33617. URL https://doi.org/10.1609/aaai.v39i14.33617.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems XIX*, 2023. doi: 10.15607/RSS.2023.XIX.016. URL https://doi.org/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. URL https://proceedings.mlr.press/v229/zitkovich23a.html.