When the Agent Turns: Correctness-Aware Multimodal Reasoning over Panoramic Yaw Orbits

Anonymous authors
Paper under double-blind review

Abstract

When an embodied agent turns, the scene remains unchanged, but the correct answer to an egocentric spatial question may change. We identify a failure in panoramic vision–language models (VLMs): stable view-averaged accuracy can coexist with *stale-frame predictions* that retain the source answer when the reference frame requires an update. To measure this failure, we introduce a correctness-aware yaw-orbit benchmark that preserves image content through cyclic shifts and distinguishes invariant questions from equivariant questions whose targets are regenerated from 3D geometry. Across five open VLMs, mean transformed accuracy on invariant questions remains within 3.25 points of clean accuracy, yet full-orbit accuracy is 24.79–35.12% and changed-label accuracy is 16.25–30.58%. We propose **OrbitAlign**, a post-training framework that incorporates the answer transformation law into supervision and optimization. Orbit Supervised Fine Tuning (SFT) learns from geometry-regenerated targets; orbit-based Group Relative Policy Optimization (GRPO) jointly optimizes paired headings; and a correctness-gated Frame-Canonical Orbit (FCO) reward couples predictions only when both are correct in their respective frames. Across three backbones, OrbitAlign improves held-out changed-label accuracy over PanoEnv GRPO by 32.05–42.17 percentage points. With 100 Reinforcement Learning (RL) steps following orbit SFT, the complete framework reduces H200 training GPU-hours by a factor of 4.1–5.2. Ablations examine the contributions of target regeneration, orbit grouping, and cross-view correctness coupling. Together, these results establish current-frame correctness as a criterion for both evaluating and training panoramic spatial reasoning models. The benchmark, code, and model checkpoints will be released.

1 Introduction

A 360° panorama represents the environment surrounding an embodied agent (Dongfang et al., 2026; Zhang et al., 2025; Lin & Zheng, 2026). Its equirectangular projection (ERP) introduces an arbitrary yaw origin: changing this origin cyclically shifts the image while preserving the physical scene. Such transformations arise when an agent changes its heading during navigation and manipulation (Pan et al., 2026), driving (Yuwen et al., 2026), and immersive assistance (Guo et al., 2026). The required answer transformation depends on the question. Object identity and presence remain *invariant*, whereas egocentric relations such as front, left, right, and behind depend on the current reference frame. Repeating the source answer when the target changes constitutes a *stale-frame prediction* (Figure 1).

Existing evaluations do not isolate this failure. Recent benchmarks study omnidirectional understanding (Dongfang et al., 2026; Zhang et al., 2025; Lin & Zheng, 2026), mental rotation and perspective re-anchoring (Zhang et al., 2026; Chen et al., 2026), counterfactual camera rotations (Vellamcheti et al., 2026), and normalized cross-view consistency (Kim et al., 2026). However, camera motion can change visibility, and frame-normalized agreement can remain high when predictions are incorrect in their respective frames. These evaluations therefore leave unresolved whether a model maintains current-frame correctness when a transformation changes the reference heading while preserving scene content. We address this question through *correctness-aware yaw orbits* (Section 3). Integer column rolls of the ERP preserve every source pixel. Invariant questions retain

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

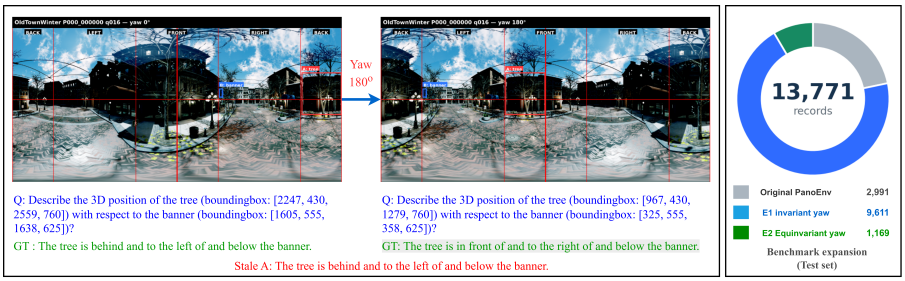

Figure 1: **Correctness-aware yaw transformation.** A 180° yaw shift preserves scene content but changes the agent's reference frame. Invariant answers remain unchanged, whereas egocentric answers must transform with the new frame.

their targets, whereas equivariant targets are regenerated from annotated 3D geometry in the rotated frame. Each prediction is evaluated against its current-frame target, and full-orbit accuracy requires correctness at every heading. Across five open VLMs, including a panorama-specific model, mean transformed accuracy on invariant questions remains within 3.25 points of clean accuracy. Nevertheless, full-orbit accuracy is 24.79–35.12%, and changed-label accuracy is 16.25–30.58% (Section 6.1). View-averaged accuracy therefore does not establish reliability across reference frames.

This distinction also determines the supervision required for post-training. Training on a single heading does not explicitly supervise answer transformations, while copying source labels under yaw augmentation can contradict the current-frame target. We propose **OrbitAlign** to supervise these transformations through two training stages (Section 4). Orbit SFT learns from invariant targets and geometry-regenerated equivariant targets. Orbit-grouped GRPO subsequently optimizes paired headings, with the correctness-gated FCO reward coupling their current-frame correctness. Its advantage estimator standardizes the base reward and adds the centered orbit term, retaining the cross-view learning signal. Held-out experiments evaluate OrbitAlign across three model backbones (Section 6.2). An ablation then examines the contributions of supervision, orbit grouping, and FCO (Section 6.3). Under identical yaw augmentation, regenerating targets instead of copying source labels improves changed-label accuracy by 27.08 points. Within the 100-step RL comparison, adding FCO to orbit-grouped GRPO improves changed-label accuracy by 1.93 points and reduces stale-answer rate by 12.13 points. Supplementary analyses examine unseen angles, coordinate-free prompts, and a second dataset using OrbitAlign checkpoints. The evaluation also distinguishes geometric reasoning from label transformation alone. A label-determinism audit shows that 67.2% of changed-label test targets cannot be determined from the source label and yaw without additional scene geometry. Thus, supervision and evaluation must account for the target in each reference frame.

Our contributions are threefold:

- *We introduce a **correctness-aware panoramic yaw benchmark** that separates invariant from equivariant questions, preserves scene pixels under transformation, regenerates current-frame targets from 3D geometry, and evaluates correctness over complete yaw orbits.*
- *We identify the **stale-frame failure** in panoramic VLMs: stable view-averaged accuracy can coexist with errors in full-orbit prediction and egocentric answer updating, including on targets that cannot generally be inferred from the source label and yaw alone.*
- *We propose **OrbitAlign**, a post-training framework that integrates geometry-regenerated orbit SFT, orbit-grouped GRPO, and a correctness-gated FCO reward. Held-out experiments evaluate the framework across model backbones, and an ablation examines the contribution of each component.*

2 Related Work

Panoramic and viewpoint-conditioned spatial reasoning. OSR-Bench evaluates counting, distance, and direction in omnidirectional scenes (Dongfang et al., 2026); 360-R1 stud-

ies panoramic question answering with rule-based reinforcement learning (Zhang et al., 2025); PanoEnv provides geometry-grounded questions with a two-stage GRPO curriculum (Lin & Zheng, 2026); PanoWorld adds panorama-native localization and frame transformation (Wang et al., 2026); and CoSpace and H* Bench study multi-view integration and active search (Zhu et al., 2025b; Yu et al., 2026). Further diagnostics cover observer-centered reasoning, 3D orientation, uncommon viewpoints, and perspective-taking (Shiri et al., 2024; Wang et al., 2025; Ma et al., 2025; Kong et al., 2026; Zhang et al., 2026). The closest transformation-oriented benchmarks differ in how viewpoints and frames are constructed and evaluated (Table 1). PCSR-Bench reasons about hypothetical observer transformations from a supplied panorama (Chen et al., 2026); CVT-Bench evaluates counterfactual camera transformations without exposing the target view (Vellamcheti et al., 2026); and SCOPE evaluates predictions across distinct camera views after mapping them to a common frame (Kim et al., 2026). Our setting isolates reference-frame changes under content-preserving yaw and scores correctness in the current frame over complete orbits.

Table 1: Protocol comparison with transformation-oriented benchmarks. Views: transformed visual inputs; Source perm.: exact source-pixel permutation; Frame GT: frame-specific ground truth; Inv./equiv.: explicit target contracts; Orbit GT: joint correctness over a prescribed orbit; Orbit train.: grouped orbit post-training.

Benchmark	Views	Source perm.	Frame GT	Inv./equiv.	Orbit GT	Orbit train.
SpinBench	✓	✗	✓	✗	✗	✗
PCSR-Bench	✗	✗	✓	✗	✗	✗
CVT-Bench	✗	✗	✓	✗	✗	✗
SCOPE	✓	✗	✓	✗	✗	✗
Ours	✓	✓	✓	✓	✓	✓

Transformation robustness, consistency, and equivariance. Group-equivariant and spherical CNNs encode equivariance into visual representations (Cohen & Welling, 2016; Cohen et al., 2018). We instead study *behavioral equivariance* in language-conditioned reasoning, where the required answer transformation depends on the question. Robustness benchmarks such as ImageNet-P (Hendrycks & Dietterich, 2019), VQA-Rephrasings (Shah et al., 2019), and MM-R^3 (Chou et al., 2025), and studies of VLM sensitivity to geometric transformations (Qiu et al., 2026), equivariant multimodal similarity (Wang et al., 2023), and paired invariant/equivariant edits (Khanbayov & Kurban, 2026), emphasize invariant targets, pairwise consistency, or representation-level behavior. Ours scores ground-truth correctness over complete orbits under explicit contracts, so a consistently wrong orbit receives no credit.

Post-training for spatial reliability. GRPO (Shao et al., 2024) has been adapted to panoramic and spatial reasoning in 360-R1, PanoEnv, and PCSR-Bench (Zhang et al., 2025; Lin & Zheng, 2026; Chen et al., 2026), and Spatial-SSRL derives self-supervised rewards from geometric transformations in perspective images (Liu et al., 2026). These approaches optimize independently scored prompts. OrbitAlign combines supervision on geometry-regenerated targets with joint optimization of transformed views from the same instance. Orbit SFT establishes the target transformation contract, while orbit-grouped GRPO and FCO couple current-frame correctness across headings (Section 4).

3 CORRECTNESS-AWARE YAW ORBITS

Yaw transformation and target contract. Let $x \in \mathbb{R}^{H\times W\times 3}$ be an ERP panorama, $\mathcal{G}$ the set of evaluated yaw rotations, including $e = 0°$, and q a spatial question about x. Each $g \in \mathcal{G}$ is implemented as an integer circular shift of ERP columns, preserving all source pixels without resampling. The yaw orbit is

$$\mathcal{O}(x) = \{gx : g \in \mathcal{G}\}. \tag{1}$$

The transformed input is $T_g(x, q) = (gx, \tau_g(q))$, where τ_g updates image-coordinate references in the question q.

The target depends on the question contract:

$$y_g = \begin{cases} y, & \text{invariant,} \\ H(R_g Z, q), & \text{equivariant,} \end{cases} \tag{2}$$

where Z is the annotated 3D geometry, R_g expresses it in the rotated frame, and H is the fixed PanoEnv relation rule. Thus, identity, category, and supported attribute questions retain their targets,

whereas egocentric relations such as left, right, front, and behind are regenerated in the current frame. Consequently, y_g need not be determined by the source label y and yaw alone.

A model is correct over an orbit when

$$f(T_g(x, q)) = y_g, \qquad \forall g \in \mathcal{G}. \tag{3}$$

This distinguishes correctness from consistency: repeating the same wrong answer across headings is consistent but not correct.

Benchmark populations. We build the benchmark from PanoEnv (Lin & Zheng, 2026), using fixed transformation and relabeling rules. *E1* evaluates invariant questions over complete eight-heading orbits $\{0°, 45°, \ldots, 315°\}$: 242 validation orbits and 1,373 frozen-test orbits. *E2* evaluates egocentric relative-position questions on seam-clean clean–transformed pairs: 491 validation pairs, of which 363 change label, and 948 frozen-test pairs, of which 702 change label. Full construction and filtering details are provided in Appendix B.

Correctness-aware metrics. Let $\hat{y}_{ig}$ be the parsed prediction and $a_{ig} = \mathbb{I}[\hat{y}_{ig} = y_{ig}]$. For E1, we report mean transformed accuracy and use *full-orbit accuracy* as the primary metric:

$$A_{\text{orb}} = \frac{1}{N} \sum_{i=1}^{N} \prod_{g \in \mathcal{G}} a_{ig}. \tag{4}$$

It measures the fraction of examples answered correctly at every heading.

For E2, let $\mathcal{P}$ be the eligible pairs and $\mathcal{C} = \{(i, g) \in \mathcal{P} : y_{ig} \neq y_{ie}\}$ the subset whose target changes. Our primary metric is

$$A_{\text{chg}} = \frac{1}{|\mathcal{C}|} \sum_{(i,g) \in \mathcal{C}} a_{ig}, \tag{5}$$

which directly measures whether the model updates its answer when the reference-frame change requires it. We additionally report stale rate, defined as the fraction of parseable pairs in $\mathcal{C}$ for which $\hat{y}_{ig} = \hat{y}_{ie}$.

All results use the official PanoEnv prompt, parser, and strict scoring criterion. Additional metrics, seam sensitivity, construction checks, and uncertainty estimation are reported in the appendix.

4 The OrbitAlign post-training framework

OrbitAlign applies the benchmark's transformation contract to post-training through two stages (Figure 2; Appendix C). Orbit SFT supervises invariant and equivariant answers across headings. Orbit-grouped GRPO then optimizes paired views using current-frame answer rewards and the FCO reward, which couples correctness across views. The framework thus integrates target construction, training-group composition, and reward design.

Stage 1: orbit SFT. The training set contains 7,307 invariant and 984 equivariant orbits over the cardinal headings $0°$, $90°$, $180°$, and $270°$. Equivariant labels are regenerated from rotated geometry rather than copied; camera-view questions without a verified regenerator are excluded. Orbit-balanced SFT samples both behaviors across structured and open-ended questions, rotates the exposed heading across epochs, and alternates changed and stable equivariant targets (Algorithm 2 in Supplementary). This stage supervises both transformation contracts and initializes the policy for subsequent optimization with within-group reward differences.

Stage 2: orbit-grouped GRPO. OrbitAlign samples paired headings to limit the cost of evaluating a complete orbit at each update. For instance i, a group contains the clean view e and one scheduled nonzero cardinal view $g \sim \mu_i$, where μ_i rotates the partner heading and, for equivariant orbits, alternates changed and stable targets. Two responses $p_{v,j}$ are sampled per view v and receive the official current-frame rewards $R^{\text{base}}_{v,j} = \lambda_a R^{\text{answer}}_{v,j} + \lambda_f R^{\text{format}}_{v,j}$ with $(\lambda_a, \lambda_f) = (0.9, 0.1)$. We standardize only the base reward within each view, then add the centered FCO term:

$$\widetilde{R}^{\text{base}}_{v,j} = \frac{R^{\text{base}}_{v,j} - \text{mean}_{j'} R^{\text{base}}_{v,j'}}{\text{std}_{j'} R^{\text{base}}_{v,j'} + \epsilon_A}, \qquad A_{v,j} = 5 \left[\widetilde{R}^{\text{base}}_{v,j} + \lambda_o \left(R^{\text{FCO}}_{v,j} - \text{mean}_{j'} R^{\text{FCO}}_{v,j'} \right) \right], \tag{6}$$

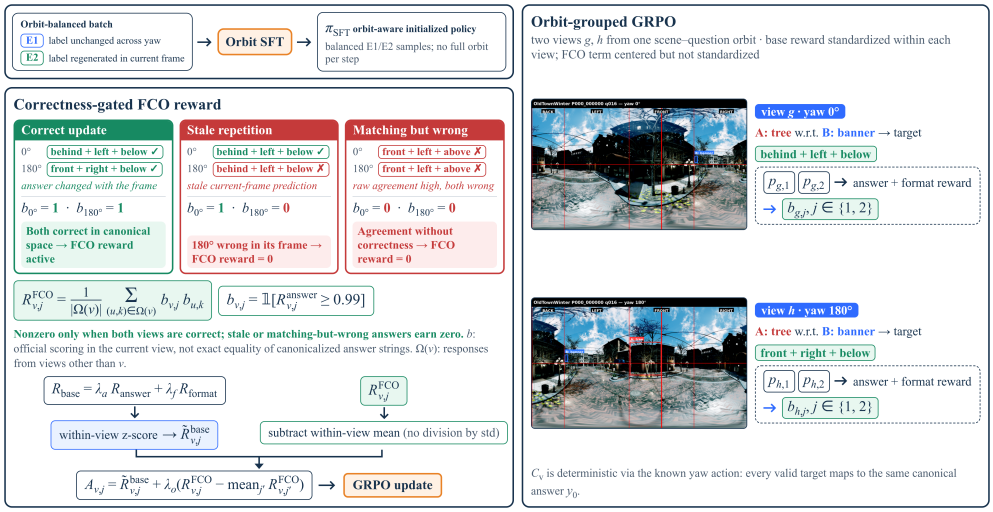

Figure 2: **OrbitAlign.** Orbit SFT trains invariant (E1) and current-frame equivariant (E2) behavior, followed by orbit-grouped GRPO that jointly optimizes paired yaw views. Predictions are mapped to a canonical answer space, where the correctness-gated FCO term rewards pairs only when both answers are correct in their current frames. The FCO advantage estimator retains the cross-view contribution after standardizing the base reward.

with scale 5, sample standard deviation, and $\epsilon_A = 10^{-4}$ for the base reward. Centering the FCO term without standardization preserves its dependence on partner-view correctness. Orbit grouping changes the training distribution: each update samples rotated views scored against geometry-regenerated targets, rather than clean prompts only. It also changes group composition: a group holds two prompts with two responses each, rather than four responses to one prompt. The FCO term additionally weights a response's advantage by the other view's correctness. The ablation separates the orbit-grouped arm without FCO from the arm with the FCO term.

4.1 Correctness coupling through FCO

Orbit grouping is motivated by joint correctness, $J_{\text{pair}} = \mathbb{E}_i \mathbb{E}_{g\sim\mu_i} \mathbb{E}_{p_e,p_g\sim\pi_\theta}[a_{ie} a_{ig}]$ (7), with a_{ie}, a_{ig} the current-frame correctness of the sampled responses. We use a correctness-gated agreement reward. A category-specific map C_v sends each parsed prediction to a canonical label space, $c_{v,j} = C_v(p_{v,j})$, so all valid targets share one canonical answer y_0 (inverse rotation for cardinal directions, record-specific identity or swap for binary answers, a geometry-derived bijection for triplets). The FCO gate uses official current-frame answer scoring, $b_{v,j} = \mathbb{I}[R^{\text{answer}}_{v,j} \geq 0.99]$, rather than exact canonical-string equality, which can reject valid equivalent answers. The Frame-Canonical Orbit (FCO) reward is positive only when both views are correct,

$$R^{\text{FCO}}_{v,j} = \frac{1}{|\Omega(v)|} \sum_{(u,k)\in\Omega(v)} b_{v,j} b_{u,k}, \tag{8}$$

with $\Omega(v)$ the responses of the other view and weight $\lambda_o = 0.5$ in Eq. 6. Unlike raw agreement, this reward gives no credit to a consistently wrong orbit. The partner-view correctness rate $\bar{b}_u = |\Omega(v)|^{-1} \sum_{(u,k)\in\Omega(v)} b_{u,k}$ gives $R^{\text{FCO}}_{v,j} = b_{v,j} \bar{b}_u$. Under Eq. 6, the FCO contribution to the advantage is $5\lambda_o \bar{b}_u (b_{v,j} - \bar{b}_v)$. It is nonzero when responses in the current view differ in correctness and at least one partner response is correct. The policy gradient of Eq. 7 is $\mathbb{E}[(a_{ie} - \bar{a}_{ie})\, \bar{a}_{ig}\, \nabla_\theta \log \pi_\theta(p_e)]$ plus the symmetric term; our estimator retains this partner factor. The 100-step ablation isolates the FCO term at fixed initialization, orbit grouping, and response budget.

5 EXPERIMENTAL SETUP

Models and evaluation. The diagnostic evaluation includes Qwen2.5-VL-3B and 7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), InternVL3-8B (Zhu et al., 2025a), and PanoWorld-9B H* (Wang et al., 2026). All use the same PanoEnv prompts, deterministic decoding, a 256-token completion limit, native image processing, and no correction of parsing failures. Post-training experiments apply OrbitAlign to Qwen2.5-VL-3B and 7B using LoRA on non-vision linear layers and the language head (rank 16, alpha 32, dropout 0.05); InternVL3-8B retains its native visual processor and freezes the vision encoder and projector (Appendix D).

OrbitAlign training and ablation. Orbit SFT uses invariant views and geometry-regenerated current-frame targets for equivariant views. The 3B ablation starts from the orbit-SFT checkpoint at step 1,900. The subsequent orbit-grouped GRPO stage runs for 100 optimizer steps on mixed question types, with two orbit groups per step and four sampled responses per group (two yaw views, two responses each). The ablation compares orbit SFT alone, standard GRPO after the same initializer, orbit-grouped GRPO without FCO, and the complete method with the FCO estimator. The last two arms share the orbit inputs, initialization, and GRPO response budget, isolating the FCO contribution.

Baseline and protocol. PanoEnv GRPO (Lin & Zheng, 2026) starts from the base checkpoint and uses a 1,968-step structured-to-mixed RL curriculum with four responses per group. We retain its published training protocol, which uses unrotated source views without geometry-regenerated targets. For InternVL3-8B, an *orbit-data control* retains the PanoEnv GRPO curriculum, base initialization, and response budget while using OrbitAlign's yaw-augmented orbits and geometry-regenerated targets. This control examines the effect of transformed-data exposure within the PanoEnv training recipe. OrbitAlign's RL stage generates 800 responses versus 15,744 for PanoEnv. The 3B SFT initializer uses 1,900 updates with gradient accumulation eight (15,200 teacher-forced examples). Response counts exclude SFT computation; Table 18 reports total GPU-hours for the full 1,968-step SFT recipe and 100 RL steps. Training uses cardinal yaw angles; the diagonal angles $45°, 135°, 225°, 315°$ are held out. Checkpoint decisions use development data only, and fixed policies are evaluated on the 2,991-question PanoEnv test split and its frozen E1/E2 derivations. Public test labels are not used for test-responsive selection.

6 RESULTS

We evaluate reference-frame reliability (§6.1), held-out correctness (§6.2), component contributions (§6.3), and transfer (§6.4). Clean accuracy uses the unshifted views of 242 E1 validation orbits (Table 2), all 1,496 validation questions (Table 4), and all 2,991 test questions (Table 3).

6.1 MARGINAL ACCURACY AND ORBIT CORRECTNESS

Table 2 reports the diagnostic results for five VLMs. For every model, mean transformed E1 accuracy lies within 3.25 points of clean accuracy. Full-orbit accuracy exceeds the heading-independent null, which is below 1% for every model, indicating correlated errors across headings. However, only 52.17–68.64% of clean-correct orbits remain correct at all seven transformed headings. Thus, 31–48% of questions answered correctly at the clean heading receive an incorrect answer at another heading. On E2, changed-label accuracy is 16.25–30.58%, with stale rates of 37.32–66.39%. Scene-level intervals (Appendix A) include zero for the mean-minus-clean difference and exclude zero for the clean-to-orbit drop. The results distinguish stability in marginal accuracy from correctness across reference frames.

6.2 ORBITALIGN IMPROVES HELD-OUT ORBIT CORRECTNESS

Table 3 evaluates OrbitAlign on the frozen test set across three backbones. Relative to base-initialized PanoEnv GRPO, OrbitAlign increases E1 robust accuracy by 9.11, 10.24, and 3.22 points and E2 changed-label accuracy by 38.61, 32.05, and 42.17 points at 3B, 7B, and InternVL3-8B, respectively. Stale-answer rates decrease by 18.09, 4.56, and 23.65 points. These comparisons assess

Table 2: **Panoramic yaw diagnostics on frozen validation sets.** E1 evaluates 242 seam-clean invariant orbits; Clean is accuracy on their unshifted views, Mean is transformed accuracy, and Robust requires all eight headings correct. Two reference lines bracket Robust: Null is the full-orbit accuracy that heading-independent errors would give, $\mathrm{Clean} \times \mathrm{Mean}^7$, and Clean is its ceiling, since a robust orbit must be clean-correct; Kept = Robust/Clean is the share of clean-correct orbits that survive all seven rotations. E2 uses 491 seam-clean records, including 363 changed-label records. Flip measures prediction disagreement; Stale measures repetition when the target changes. All values are percentages. †43/242 unparseable E1 clean outputs.

		E1 — invariant					E2 — equivariant		
Model	Clean	Mean	Null	Robust	Kept	Flip↓	Mean	Changed	Stale↓
Qwen2.5-VL-3B	48.76	52.01	0.50	33.47	68.64	35.60	27.09	20.39	66.39
Qwen2.5-VL-7B	52.48	51.95	0.54	31.82	60.63	33.59	29.94	24.79	49.03
Qwen3-VL-8B	53.31	55.37	0.85	35.12	65.88	34.47	35.03	30.58	37.95
InternVL3-8B	52.07	52.18	0.55	31.82	61.11	33.47	24.64	20.66	37.32
PanoWorld-9B H*†	47.52	45.57	0.19	24.79	52.17	40.85	18.53	16.25	40.87

Table 3: **Frozen held-out test yaw benchmark.** E1 reports mean transformed accuracy, answer flip rate, and full-orbit robust accuracy; E2 reports correct-conditioned equivariance, changed-label accuracy, and stale clean-prediction rate on seam-clean pairs. The **OrbitAlign** rows use 100 optimizer steps after orbit SFT; base-initialized PanoEnv GRPO uses 1,968 RL steps and trains on unrotated views only. The InternVL3-8B orbit-data control uses yaw-augmented orbits and regenerated targets while retaining the PanoEnv GRPO training recipe. The FCO-specific effect is measured by the matched ablation in Table 4. Best value per backbone in bold; all values are percentages.

Model	Clean	E1 — invariant			E2 — equivariant		
	Acc.↑	Mean↑	Flip↓	Robust↑	Cond.↑	Changed↑	Stale↓
Qwen2.5-VL-3B							
Base	44.00	49.91	34.18	33.78	40.41	25.78	80.17
PanoEnv GRPO Lin & Zheng (2026)	48.51	53.75	**22.96**	36.81	46.07	23.36	60.54
PanoEnv GRPO (Orbit data)	45.87	53.48	26.70	40.99	37.97	26.35	55.84
OrbitAlign (Ours)	**52.59**	**58.63**	23.72	**45.92**	**68.41**	**61.97**	**42.45**
Qwen2.5-VL-7B							
Base	47.38	53.62	34.37	33.40	55.91	26.92	44.86
PanoEnv GRPO Lin & Zheng (2026)	51.92	56.90	27.08	40.61	60.98	30.34	44.73
PanoEnv GRPO (Orbit data)	51.19	56.22	24.48	41.94	56.46	28.77	48.01
OrbitAlign (Ours)	**54.76**	**59.77**	**21.12**	**50.85**	**65.93**	**62.39**	**40.17**
InternVL3-8B							
Base	45.30	52.40	36.87	32.83	36.27	19.09	44.36
PanoEnv GRPO Lin & Zheng (2026)	50.32	57.09	27.49	41.18	44.28	26.21	48.29
PanoEnv GRPO (Orbit data)	45.27	54.27	31.85	37.76	39.43	20.66	44.33
OrbitAlign (Ours)	**52.52**	**57.98**	**24.32**	**44.40**	**73.35**	**68.38**	**24.64**

the complete OrbitAlign framework against the published baseline, which trains on unrotated views. Figure 3 shows that OrbitAlign reduces both incorrect repetitions and incorrect answer changes relative to both PanoEnv GRPO controls on InternVL3-8B. For InternVL3-8B, replacing clean training inputs with orbit data in PanoEnv GRPO reduces stale-answer rate from 48.29% to 44.33%, but also reduces changed-label accuracy from 26.21% to 20.66%. Reduced repetition therefore does not imply improved current-frame correctness. Relative to this orbit-data control, OrbitAlign improves clean accuracy by 7.25 points, E1 robust accuracy by 6.64 points, and changed-label accuracy by 47.72 points, while reducing stale-answer rate by 19.69 points. Thus, transformed-data exposure alone does not reproduce OrbitAlign's gains. Evaluating the same 100-step OrbitAlign checkpoints with seam-contaminated records included yields changed-label accuracy gains of 37.18, 32.45, and 42.73 points over PanoEnv GRPO at 3B, 7B, and InternVL3-8B, respectively (Appendix B.6).

Training cost. Including orbit SFT and 100 RL steps, OrbitAlign requires 2.81, 4.39, and 2.48 H200 GPU-hours at 3B, 7B, and InternVL3-8B, respectively. PanoEnv GRPO requires $4.6\times$, $4.1\times$, and $5.2\times$ as many GPU-hours (Appendix D.8, Table 18).

Table 4: **OrbitAlign ablation on Qwen2.5-VL-3B (full validation).** O-SFT: orbit SFT; O-GRPO: orbit-grouped GRPO; S-GRPO: the standard GRPO control. Coupled marks the cross-view FCO term. The three GRPO rows share one initializer and budget (100 steps, two groups per update), and the last two use the same orbit-grouped input and differ only by the FCO term. All values are percentages.

Method	Orbit SFT	GRPO input	Coupled term	E1 Clean↑	E1 Robust↑	E2 Changed↑	E2 Stale↓
Base	–	–	–	42.65	33.47	20.39	66.39
O-SFT	Yes	–	–	51.87	47.11	46.56	50.96
O-SFT + S-GRPO	Yes	Clean	–	52.01	46.69	48.21	47.38
O-SFT + O-GRPO	Yes	Orbit	–	52.01	45.87	51.24	46.01
OrbitAlign	Yes	Orbit	Yes	54.41	45.04	**53.17**	**33.88**
Target contract (development screen)							
Yaw SFT, copied labels				36.00	20.69	36.46	83.33
Yaw SFT, regenerated labels				46.33	24.14	63.54	51.04

Table 5: **E2 relation-family gains on InternVL3-8B.** Differences are OrbitAlign (100 RL steps) minus PanoEnv GRPO (orbit data) on the 702 primary seam-clean changed-label test records, in percentage points. Positive accuracy changes and negative stale-rate changes indicate improvement. Stale rates use parseable changed-label pairs.

Family	Records	ΔChanged ↑	ΔStale ↓
Binary	209	+35.89	−21.12
Cardinal	381	+60.37	−19.25
Triplet	112	+26.79	−18.75

6.3 Contributions of the OrbitAlign components

Table 4 examines the components of OrbitAlign. Orbit SFT increases E1 robust accuracy from 33.47% to 47.11% and E2 changed-label accuracy from 20.39% to 46.56%. The post-SFT comparisons then hold initialization and the 100-step GRPO response budget fixed. Standard GRPO reaches 48.21% changed-label accuracy and a 47.38% stale rate; orbit-grouped GRPO without FCO reaches 51.24% and 46.01%, respectively. Adding FCO raises changed-label accuracy to 53.17% (+1.93 points) and reduces stale rate to 33.88% (−12.13 points); clean accuracy increases by 2.40 points, while E1 robust accuracy decreases by 0.83 points. Fixed orbit inputs, initialization, and response budget isolate FCO's contribution (Section 4.1). Matched Qwen2.5-VL-7B training traces further show that the centered FCO term remains nonzero after view-local centering, with all-zero-advantage groups decreasing from 43.5% to 37.5% (Appendix D.1). The lower panel holds yaw-augmented images and rewritten coordinates fixed, comparing copied labels with regenerated targets satisfying Eq. 3. Regeneration increases changed-label accuracy by 27.08 points and reduces stale rate by 32.29 points; copied labels reduce clean accuracy and yield 83.33% stale answers.

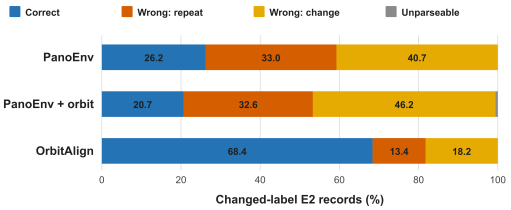

Figure 3: **Prediction transitions on changed-label E2 records.** OrbitAlign substantially increases correct updates while reducing repeated and incorrectly changed answers.

Performance across relation families. OrbitAlign improves changed-label accuracy and reduces stale rates across all three families relative to the InternVL3-8B orbit-data control (Table 5). On 112 geometry-dependent triplet records, whose targets cannot be recovered by source-label rotation, it corrects 45 errors and disrupts 15 correct predictions: 30 net corrections.

6.4 Transfer across angles, prompts, and datasets

Dependence on scene geometry. The label-determinism audit examines whether E2 targets can be inferred by transforming the source answer. In the frozen test manifest (Appendix B.7), 230 of the 702 primary changed-label targets (32.8%) are determined by the source label and yaw: cardinal-relation targets at cardinal yaws; binary targets at diagonal yaws and all triplet targets require geometry not encoded by the source label. On the 177 seam-clean label-determined cardinal records, rotating each model's clean prediction in closed form scores 48.02%, 57.06%, and 65.54% for OrbitAlign on the three backbones, its own transformed predictions score 62.71%, 69.49%, and

71.19%, and stale clean predictions score 14.12–18.08%. A post-hoc label transformation therefore does not account for performance across the benchmark.

Unseen angles. On the four diagonal angles excluded from training (Table 14 in supplementary), OrbitAlign raises changed-label accuracy over the orbit-SFT-initialized PanoEnv GRPO control from 31.91% to 70.21% at 3B and from 48.94% to 65.96% at 7B; E1 robust accuracy improves at 3B and is unchanged at 7B. These descriptive comparisons use 49 E1 and 24 E2 source questions.

Coordinate-free prompts. Yaw transformations rewrite the bounding-box coordinates in PanoEnv questions, potentially permitting answer updates from textual information. The marker diagnostic replaces these coordinates with visual A/B/C markers (Table 15 in supplementary). OrbitAlign achieves the highest accuracy and lowest stale rate across the reported methods, increasing changed-label accuracy over the orbit-SFT-initialized PanoEnv GRPO control from 39.12% to 47.66%. Its corresponding accuracy with coordinates is 57.85%. Although removing coordinates reduces absolute accuracy, OrbitAlign retains its advantage over the control. The diagnostic evaluates one checkpoint per method on 84 source questions.

Cross-dataset evaluation. PanoEnv-S2D3D applies PanoEnv's unmodified question generator to real indoor Stanford 2D-3D-S scans Armeni et al. (2017) (Appendix G). OrbitAlign achieves the highest clean accuracy across the three backbones, exceeding PanoEnv GRPO by 10.24, 7.70, and 8.06 points. Changed-label accuracy increases by 34.07, 36.30, and 44.17 points, whereas PanoEnv GRPO increases stale-answer rate on two backbones. Differences in invariant E1 metrics are smaller, with PanoEnv GRPO exceeding OrbitAlign on InternVL3-8B. The results support transfer of equivariant answer updating from synthetic to real scenes under the same label convention; improvements in invariant accuracy depend on the dataset.

7 CONCLUSION

Limitations. The benchmark derives from one source dataset and tests horizontal reference-frame changes, not translation, pitch, or physical robot behavior. Seam filtering excludes 221 of 1,169 transformed E2 records from the primary analysis; the all-record analysis retains the gains but cannot remove the ambiguity of seam-enclosing prompts. Training and evaluation targets share the pinned relation rule, so E2 measures its application on held-out scenes rather than transfer to an independently authored convention. The five diagnostic models are open 3B–9B checkpoints, not frontier closed models. The OrbitAlign component ablation uses one 3B training run; training-run variance remains unmeasured. OrbitAlign includes an SFT stage, and its comparison with base-initialized PanoEnv GRPO is not matched in total compute. The GPU-hour comparison includes both SFT and RL (Table 18); total FLOPs are not measured. The matched 100-step ablation isolates the FCO term. The published PanoEnv GRPO baseline trains on unrotated views, so the comparison also changes the training distribution. The InternVL3-8B orbit-data control examines transformed-data exposure within the baseline recipe; component attribution relies on the matched ablation.

Conclusion. Content-preserving yaw orbits reveal discrepancies between view-averaged accuracy, full-orbit correctness, and egocentric answer updating. Geometry-regenerated targets make these discrepancies measurable under invariant and equivariant question contracts. OrbitAlign incorporates the same contracts into post-training through orbit SFT followed by orbit-grouped GRPO with a correctness-gated FCO reward. Held-out comparisons across three backbones show improvements in full-orbit and changed-label accuracy, together with reductions in stale-answer rate. The ablation identifies the contributions of target regeneration, orbit-grouped optimization, and cross-view correctness coupling within this framework. Reliable spatial reasoning across changing reference frames remains an open problem requiring further investigation.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the text, code design and development. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding, 2017. URL https://arxiv.org/abs/1702.01105.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Yuangong Chen, Wai Keung Wong, Jiaxing Li, Ioannis Patras, and Xu Zheng. Beyond localization: A comprehensive diagnosis of perspective-conditioned spatial reasoning in mllms from omnidirectional images, 2026. URL https://arxiv.org/abs/2605.12413.

Shih-Han Chou, Shivam Chandhok, Jim Little, and Leonid Sigal. MM-r^3: On (in-)consistency of vision-language models (VLMs). In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4762–4788, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.246. URL https://aclanthology.org/2025.findings-acl.246/.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/cohenc16.html.

Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hkbd5xZRb.

Zihao Dongfang, Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Danda Pani Paudel, Luc Van Gool, Kailun Yang, and Xuming Hu. Are multimodal large language models ready for omnidirectional spatial reasoning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9759–9769, June 2026.

Xucheng Guo, Bing Li, Yuhan Zhou, Han Zhou, Majed Elwardy, Yan Hu, Yuanfeng Zhou, Xiaoming Chen, Ling Wang, and Yiran Shen. Towards event-guided panoramic hdr video reconstruction for indoor immersive vr: A novel dataset and approach. In *2026 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 66–76, 2026. doi: 10.1109/VR67842.2026.00030.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Rasul Khanbayov and Hasan Kurban. Consistency has a computable blind spot: A commutation theory of label-free reliability for vision-language figure reading, 2026. URL https://arxiv.org/abs/2608.05675.

Yoonji Kim, Jieun Kim, Yujin Jeong, and Sung-Bae Cho. Diagnosing spatial consistency across perspectives and viewpoints in large vision-language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32803–32827, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1514. URL https://aclanthology.org/2026.acl-long.1514/.

Fei Kong, Jinhao Duan, Kaidi Xu, Zhenhua Guo, Xiaofeng Zhu, and Xiaoshuang Shi. Lrr-bench: Left, right or rotate? vision-language models still struggle with spatial understanding tasks, 2026. URL https://arxiv.org/abs/2507.20174.

Zekai Lin and Xu Zheng. Panoenv: Exploring 3d spatial intelligence in panoramic environments with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9647–9657, June 2026.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9570–9581, June 2026.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934, October 2025.

Haotian Pan, Shibo Huang, Jian Yang, Jinpeng Mi, Ke Li, Xiong You, Peidong Liang, Jinbo Yang, Yingjie Liu, Jianfeng Zhang, Muyu Wang, Jie Yang, Xinyu Zhang, Lijun Zhao, Mingsong Chen, Jie Zhou, and Xian Wei. Robot navigation via foundation language models: A review. *ACM Comput. Surv.*, 58(11), April 2026. ISSN 0360-0300. doi: 10.1145/3802539. URL https://doi.org/10.1145/3802539.

Jason Qiu, Zachary Meurer, Xavier Thomas, and Deepti Ghadiyaram. Semantic richness or geometric reasoning? the fragility of vlm's visual invariance, 2026. URL https://arxiv.org/abs/2604.01848.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models, 2024. URL https://arxiv.org/abs/2411.06048.

Shanmukha Vellamcheti, Uday Kiran Kothapalli, Disharee Bhowmick, and Sathyanarayanan N. Aakur. Cvt-bench: Probing spatial-state integrity through counterfactual viewpoint transformations, 2026. URL https://arxiv.org/abs/2603.21114.

Changpeng Wang, Xin Lin, Junhan Liu, Yuheng Liu, Zhen Wang, Donglian Qi, Yunfeng Yan, and Xi Chen. Panoworld: Towards spatial supersensing in 360° panorama world, 2026. URL https://arxiv.org/abs/2605.13169.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11998–12008, October 2023.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large mutimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24669–24679, June 2025.

Heyang Yu, Yinan Han, Xiangyu Zhang, Baiqiao Yin, Bowen Chang, Xiangyu Han, Xinhao Liu, Jing Zhang, Marco Pavone, Chen Feng, Saining Xie, and Yiming Li. Thinking in 360deg: Humanoid visual search in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22445–22455, June 2026.

Xuan Yuwen, Ziwang Lu, Jicheng Chen, Long Chen, and Hui Zhang. Panoramic active visual system for distant traffic sign recognition. *IEEE Transactions on Intelligent Transportation Systems*, 27(4):4047–4057, 2026. doi: 10.1109/TITS.2026.3662047.

Xinshen Zhang, Zhen Ye, and Xu Zheng. Towards omnidirectional reasoning with 360-r1: A dataset, benchmark, and grpo-based method, 2025. URL https://arxiv.org/abs/2505.14197.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. Spinbench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r7rUDgGYC4.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025a. URL https://arxiv.org/abs/2504.10479.

Yiqi Zhu, Ziyue Wang, Can Zhang, Peng Li, and Yang Liu. Cospace: Benchmarking continuous space perception ability for vision-language models, 2025b. URL https://arxiv.org/abs/2503.14161.

Table 6: **Stage I headline for Qwen2.5-VL-7B with 95% panorama-scene bootstrap intervals** (1000 percentile resamples, frozen seed 20260728; the resampling unit is the panorama scene, because questions are clustered within scenes). The mean-minus-clean intervals include zero; the E1 clean-to-orbit gap interval does not. Strict seven-angle correctness on the 22 complete seam-clean E2 orbits is 0/22; its percentile bootstrap is degenerate and is reported as an observed count rather than an interval.

Metric	Estimate (%)	95% CI
E1 — invariant questions (n=242 orbits, 55 scenes)		
Clean accuracy	52.48	[46.77, 58.01]
Mean transformed accuracy	51.95	[46.22, 57.56]
Mean transformed − clean	-0.53	[-3.99, 2.85]
Full-orbit robust accuracy	31.82	[25.11, 38.10]
Clean → full-orbit drop	20.66	[15.45, 26.47]
Answer-flip rate	33.59	[29.04, 37.36]
Correct-conditioned consistency	84.14	[79.61, 88.40]
E2 — equivariant questions (n=84 questions, 491 records, 49 scenes)		
Paired clean accuracy	29.74	[20.33, 39.80]
Mean transformed accuracy	29.94	[25.42, 34.63]
Mean transformed − clean	0.20	[-9.42, 9.15]
Correct-conditioned equivariance	45.21	[35.57, 55.67]
Accuracy when the label changes	24.79	[19.95, 30.34]
Prediction-change rate	50.97	[44.85, 58.42]
Stale-clean-answer rate	49.03	[41.58, 55.15]

Ethics statement

This work uses public or procedurally generated benchmark data and does not involve human subjects. Its intended use is diagnostic: exposing reliability gaps before panoramic VLMs are deployed in embodied systems. The benchmark does not establish safety for robotics or other high-stakes applications. Release will follow the licenses and redistribution terms of PanoEnv and TartanAir.

Reproducibility statement

The supplementary material specifies data provenance, transformation and label contracts, prompts, parsing, metrics, statistical procedures, training hyperparameters, and checkpoint selection. We will release anonymized frozen E1/E2 manifests, construction and evaluation code, training configurations, and scripts that regenerate paper tables from checksummed metric artifacts.

A Statistical intervals for yaw diagnostics

Table 6 shows why mean transformed accuracy is insufficient. For E1, the mean-minus-clean interval includes zero, while the clean-to-orbit drop is 20.66 points with a 95% CI of $[15.45, 26.47]$. Thus average accuracy can remain stable even though joint correctness over all headings is much lower. E2 exhibits the same marginal ambiguity: its mean-minus-clean interval is $[-9.42, 9.15]$, yet changed-label accuracy is only 24.79% and the stale rate is 49.03%. The intervals therefore support the diagnostic claim about orbit-level failure, not a claim that yaw necessarily lowers marginal accuracy.

Table 7: Frozen transformation-law policy. Record-conditional relative-position questions enter E2 only when the answer contains, or the verified geometry can introduce, a horizontal relation.

Population	Included question families	Target rule
E1 invariant	Environment; depth and metric distance; shape, size, and volume; metric inter-object distance	Copy the source answer after checking that the subcategory is yaw-invariant. Every source question must have all seven transformed members.
E2 equivariant	Horizontal `relpos_binary`, `relpos_cardinal`, and directional `relpos_triplet_extreme` records	Rotate the official 3D centroids, then re-run the pinned PanoEnv relation rule in the current frame. The source answer is never copied as the transformed target.
Excluded	Camera-view questions without a verified geometry regenerator; unknown subcategories; ambiguous triplet rankings; invalid or incomplete orbits	Exclude with a recorded reason. No heuristic label is introduced.

B SUPPLEMENTARY BENCHMARK PREPARATION

This section specifies the construction protocol for the frozen yaw benchmark, distinguishing the source PanoEnv data, the generated transformations, and the filtering decisions made before inference. No model output is used to select a question, yaw angle, or seam-clean record.

B.1 SOURCE POPULATION AND EXACT YAW CONSTRUCTION

The frozen source is the 2,991-question PanoEnv test split, containing 1,523 multiple-choice, 600 open-ended, and 868 true/false questions from 120 panorama scenes. Each source image is a 2560×1280 ERP. For $g \in \{45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ, 315^\circ\}$, we apply an integer circular horizontal roll of $s_g = 2560g/360$ pixels. All seven shifts are integral, so benchmark image files are exact permutations of the source pixels and require no interpolation. The model's native processor may subsequently resize or tile the file; the pixel-permutation guarantee therefore applies to the benchmark files rather than necessarily to the final model tensor.

Bounding-box literals in the question are shifted by the same s_g and rewritten in the transformed image coordinates. When a box crosses the ERP boundary, the PanoEnv convention represents its minimum enclosing interval as a full-width box. Such records remain in the construction audit, but are excluded from the primary E2 seam-clean pairwise metrics because a full-width box can obscure the queried object. E1 additionally requires a complete seam-clean orbit.

B.2 FROZEN ANSWER-BEHAVIOR POLICY

Before generating any transformed record, every subcategory is assigned a behavior contract. Unknown or unsupported cases fail closed instead of being silently treated as invariant. Table 7 summarizes the policy used for the reported benchmark.

For every E2 source question, the geometry regenerator is first executed at 0°. The record is accepted only when that output reproduces the official PanoEnv answer. At nonzero yaw, the same 3D centroids are rotated consistently with the ERP roll and the answer is re-derived with the pinned upstream label functions. A $0.5\,\mathrm{m}$ significance threshold resolves whether a directional component is present. This zero-yaw reproduction test guards against changing the task definition while constructing transformed labels.

B.3 QUALITATIVE E2 TARGET TRANSFORMATIONS

Figure 4 shows four audited E2 examples from the frozen test construction. Each pair contains the clean ERP and its exact 90° cyclic shift, with the query and reference objects retained and the

Table 8: Frozen test benchmark accounting. Expected-change and expected-stable counts partition all E2 transformed records; seam-clean and seam-contaminated counts form a separate partition.

Population item	Count
Original PanoEnv clean questions / scenes	2,991 / 120
E1 complete invariant source orbits / scenes	1,373 / 120
E1 transformed evaluations ($1{,}373 \times 7$)	9,611
E2 equivariant source orbits / scenes	167 / 97
E2 complete seven-angle seam-clean orbits	46
E2 source questions with primary seam-clean records	165
E2 transformed evaluations (167×7)	1,169
Primary seam-clean records	948
Seam-contaminated records	221
Expected-change records	866
Expected-stable records	303
Total clean plus added yaw evaluations	13,771

ground-truth relation regenerated in the shifted action frame. These are ground-truth construction examples rather than model predictions.

B.4 Construction procedure

Algorithm 1 gives the deterministic construction procedure, including the frozen behavior policy, zero-yaw regeneration check, seam routing, and manifest freeze.

Each transformed row stores both source and transformed identifiers, question text, bounding boxes, answers, yaw angle, shift in pixels, behavior label, regenerator version, and seam flag. These fields make it possible to audit a metric back to the unmodified PanoEnv question.

B.5 Population and attrition accounting

Table 8 distinguishes source questions from added transformed evaluations. In particular, the 9,611 E1 and 1,169 E2 records are new yaw evaluations derived from existing PanoEnv questions; they are not claimed as independently authored source questions.

E1 robust accuracy uses the clean record and all seven transformed members, so one failed heading makes the source orbit incorrect. E2 pairwise metrics use the 948 seam-clean transformed records, of which 702 change the target label. The 167 constructed E2 orbits include 46 complete seam-clean orbits; the remaining 121 contain at least one seam-contaminated view. Complete-orbit E2 statistics are reported separately and never substituted for the pairwise population.

B.6 E2 population cross-tabulation and seam sensitivity

Table 9 cross-tabulates the two E2 eligibility criteria. Seam filtering removes 221/1,169 records (18.91%), including 164 changed-label records. The primary changed-label denominator is therefore 702, compared with 866 when all records are retained. Filtering varies by question family: it excludes 14.76% of binary, 19.05% of cardinal, and 26.34% of triplet-extreme records. Thus, similar overall changed-label proportions do not imply identical question composition across seam subsets.

Table 10 evaluates the same 100-step OrbitAlign checkpoints as Table 3. The seam-clean changed-label accuracy and stale-answer rate therefore match the main table. The analysis compares the primary subset, the excluded seam-contaminated records, and their union using the scored outputs without further inference or selection. OrbitAlign improves changed-label accuracy over PanoEnv GRPO on all three backbones even when seam-contaminated records are retained. For example, the 3B gain is 31.10 percentage points on the excluded records and 37.18 points on all records, compared with 38.61 points on the primary subset. These descriptive comparisons show that the advantage is not confined to seam-clean prompts; they do not establish that filtering is unbiased, since the subsets contain different questions and the full-width box representation remains ambiguous.



(a) AmericanDiner, q004 **(b) AbandonedSchool, q017**

(c) Rome, q006 **(d) CountryHouse, q009**

Figure 4: **Representative E2 current-frame label changes under a** 90° **yaw.** Each panel pairs the clean 0° ERP (left) with its exact cyclic column shift (right). Red and blue boxes identify the query and reference objects; the text below each panorama gives the coordinate-updated question and geometry-regenerated ground truth. The horizontal target changes from (a) front–right to behind–right for the `kitchencounter` relative to the table, (b) behind–left to front–left for the door relative to the tree, (c) front–left to front–right for the `burlapsack` relative to the `ceramicpot`, and (d) right to behind for the plate relative to the fireplace. Vertical relations remain unchanged. The scene pixels are preserved up to cyclic permutation; these panels show benchmark targets, not model outputs.

Table 9: Frozen E2 test population accounting. The upper panel crosses seam status with whether the target changes; the lower panel reports question-family composition. Counts describe transformed records, not independent questions.

Seam status	Changed	Stable	Total
Seam-clean (primary)	702	246	948
Seam-contaminated	164	57	221
All records	866	303	1,169
Question family	**Seam-clean**	**Seam-contaminated**	**Total**
Binary	358	62	420
Cardinal	425	100	525
Triplet extreme	165	59	224

Table 10: E2 seam sensitivity on the frozen test set, using the same 100-step OrbitAlign checkpoints as Table 3. Mean is transformed accuracy; Changed is accuracy on target-changing records; Stale is the repeated clean-prediction rate over parseable changed-label pairs. The three populations contain 948/221/1,169 records and 702/164/866 changed-label records, respectively. All values are percentages. Seam-contaminated results use only the 221 excluded records, not every view of the 121 seam-exposed source orbits.

	Seam-clean (primary)			Seam-contaminated			All records		
Method	Mean↑	Changed↑	Stale↓	Mean↑	Changed↑	Stale↓	Mean↑	Changed↑	Stale↓
Qwen2.5-VL-3B									
Base	26.58	25.78	80.17	24.89	23.17	73.45	26.26	25.29	78.85
PanoEnv GRPO	26.58	23.36	60.54	24.43	23.17	49.39	26.18	23.33	58.43
OrbitAlign (ours)	60.65	61.97	42.45	53.85	54.27	35.37	59.37	60.51	41.11
Qwen2.5-VL-7B									
Base	32.17	26.92	44.86	25.34	19.51	41.46	30.88	25.52	44.21
PanoEnv GRPO	34.49	30.34	44.73	26.70	21.34	45.73	33.02	28.64	44.92
OrbitAlign (Ours)	62.24	62.39	40.17	55.20	55.49	30.49	60.91	61.09	38.34
InternVL3-8B									
Base	24.05	19.09	44.36	25.79	21.34	32.30	24.38	19.52	42.09
PanoEnv GRPO	29.22	26.21	48.29	28.51	23.78	46.95	29.08	25.75	48.04
OrbitAlign (Ours)	**70.04**	**68.38**	24.64	71.49	68.90	18.29	70.32	68.48	23.44

Table 11: Label determinism of primary E2 test targets. A target is determined when it is a function of the source target and yaw alone.

Family	Records	Determined	Changed	Changed determined
Cardinal	425	177	381	177
Binary	358	53	209	53
Triplet extreme	165	0	112	0
All primary	948	230	702	230

Parsing failures remain accuracy errors and are excluded only from the parseable-pair denominator of Stale. Across all records, the three base models have 259, 2, and 21 unparseable transformed outputs at Qwen 3B, Qwen 7B, and InternVL3-8B, respectively. Among the six trained checkpoints, only OrbitAlign-3B has an unparseable output (one seam-clean record); none has a parsing failure on the 221 seam-contaminated records. Thus, the trained-model comparison on the excluded subset uses the same 164-pair stale-answer denominator.

B.7 Label determinism and programmatic-rotation baseline

Table 11 audits, on the frozen E2 test manifest, whether each primary transformed target is a function of the question family, the source target, and the yaw alone. Cardinal-relation targets at cardinal yaws are determined by the quarter-turn rule, which we verify reproduces every such target. Binary targets are checked empirically across the manifest and are determined only when every record sharing the same source answer and yaw has the same transformed answer. Triplet targets are scene-specific object names and have no label-only rule. Overall, 230 of 948 primary records (24.3%) and 230 of 702 changed-label records (32.8%) are label-determined.

On the 177 seam-clean label-determined cardinal records, all of which change the target, a programmatic baseline that parses each arm's clean prediction with the official direction extractor, rotates it by the closed-form quarter-turn rule, and scores it with the official spatial rule reaches 48.02%, 57.06%, and 65.54% for the 1,968-step OrbitAlign checkpoints at 3B, 7B, and InternVL3-8B, against 62.71%, 69.49%, and 71.19% for OrbitAlign's own transformed predictions and 14.12–18.08% for the unrotated clean prediction. Every trained arm's own predictions exceed the rotated baseline; untrained models are almost never strictly correct on these three-axis questions at any

heading. `scripts/build_programmatic_rotation_baseline.py` regenerates the analysis from existing scored outputs.

B.8 PROMPT, PARSING, AND SCORING CONTRACT

All headline evaluations use the frozen `panoenv_official_v1` prompt. The prompt states the native ERP resolution, presents the transformed question, asks for a reasoning span inside `<Reasoning>` tags, and requires the final response inside one `<Answer>` tag. Decoding is deterministic with at most 256 completion tokens. Each model retains its native image processor; the question, target, parser, and metric implementation are otherwise shared.

We score answers with the pinned upstream PanoEnv evaluator and its question-type-specific matching rules. The correctness threshold is 0.99. Unparseable responses are retained as failures and agree with no prediction, including another unparseable prediction. This fail-closed rule prevents formatting failures from appearing artificially consistent across views.

Algorithm 1 Frozen correctness-aware yaw benchmark construction

Require: Frozen PanoEnv records $\mathcal{D}$; nonzero yaw set $\mathcal{G}_+ = \mathcal{G} \setminus \{e\}$; frozen behavior policy π; pinned geometry regenerator H
Ensure: E1 orbit manifest $\mathcal{M}_{\text{inv}}$, E2 record manifest $\mathcal{M}_{\text{eq}}$, exclusion ledger $\mathcal{L}$

Validate the versions of π and H and the uniqueness of $\mathcal{G}_+$
for all $r = (x, q, y, B, Z) \in \mathcal{D}$ **do**
  $b \leftarrow \pi(r)$
  if b is neither invariant nor supported relative-position equivariant **then**
    Append (r, reason) to $\mathcal{L}$; **continue**
  end if
  if b is equivariant **then**
    Require $H(Z, 0^\circ) = y$ ▷ zero-yaw reproduction
  end if
  if r contains a source seam object **then**
    Append $(r, \text{source-seam exclusion})$ to $\mathcal{L}$; **continue**
  end if
  $\mathcal{O}_r \leftarrow \varnothing$
  for all $g \in \mathcal{G}_+$ **do**
    $s_g \leftarrow Wg/360$; require s_g to be integral
    $x_g \leftarrow \textsc{CircularRoll}(x, s_g)$
    $(q_g, B_g, z_g) \leftarrow \textsc{RewriteBoxes}(q, B, s_g)$
    if b is invariant **then**
      $y_g \leftarrow y$
    else
      $Z_g \leftarrow \textsc{RotateCentroids}(Z, g)$
      $y_g \leftarrow H(Z_g)$
    end if
    Add $(x_g, q_g, y_g, g, s_g, z_g)$ and provenance to $\mathcal{O}_r$
  end for
  Require $|\mathcal{O}_r| = |\mathcal{G}_+|$
  if b is invariant and every member is seam clean **then**
    Append the complete orbit $\mathcal{O}_r$ to $\mathcal{M}_{\text{inv}}$
  else if b is equivariant **then**
    Append the complete orbit $\mathcal{O}_r$ to $\mathcal{M}_{\text{eq}}$; mark its seam-clean members as primary
  else
    Append $(r, \text{seam-contaminated orbit})$ to $\mathcal{L}$
  end if
end for
Freeze manifests, transformation provenance, and SHA-256 hashes before inference

C Implementation-aligned pseudocode

The following procedures summarize the frozen implementation contract. They make the data flow and optimization logic explicit while omitting framework- specific loading, checkpointing, and distributed-execution code.

Algorithm 1 separates orbit validity from metric eligibility. The behavior policy first prevents unsupported questions from being assigned an implicit transformation law, and the zero-yaw check requires the equivariant regenerator to reproduce the released source target before it can create any new label. Each accepted yaw transforms the image columns and prompt boxes together; equivariant records additionally rotate the 3D centroids and regenerate the current-frame target. Finally, E1 requires every nonzero-yaw member to be seam clean, whereas E2 retains complete source-orbit accounting but marks only individually seam-clean records as primary. Freezing provenance and hashes before inference makes these eligibility decisions independent of model behavior.

Algorithm 2 reflects the implemented behavior-balanced pipeline. The first loop materializes complete four-heading training orbits and audits that mapping each view target back to the canonical frame recovers the source target. The epoch sampler then selects invariant and equivariant source

Algorithm 2 Transformation-law-consistent orbit SFT

Require: PanoEnv training records $\mathcal{D}_{\mathrm{tr}}$; cardinal headings $\mathcal{G}_{\mathrm{tr}} = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$; policy π; regenerator H; initial policy parameters θ
Ensure: Orbit-SFT parameters θ_{SFT}

$\mathcal{O}_{\mathrm{inv}}, \mathcal{O}_{\mathrm{eq}} \leftarrow \varnothing$
for all $r = (x, q, y, B, Z) \in \mathcal{D}_{\mathrm{tr}}$ **do**
    $b \leftarrow \pi(r)$
    if b is invariant or supported relative-position equivariant **then**
        if b is equivariant **then**
            Require $H(Z, 0^\circ) = y$
        end if
        for all $g \in \mathcal{G}_{\mathrm{tr}}$ **do**
            Construct (x_g, q_g) by an online integer ERP roll and box rewrite
            $y_g \leftarrow y$ if invariant; otherwise rotate Z and regenerate with H
            Attach the view-to-canonical map C_g and require $C_g(y_g) = y$ and a valid forward round trip
            Add (x_g, q_g, y_g, g, C_g) to the corresponding complete orbit
        end for
    end if
end for
for each sampler epoch e **do**
    Select equal numbers of invariant and equivariant source orbits
    Rotate orbit/view exposure deterministically using the frozen seed and e
    For equivariant orbits, alternate changed/stable targets when available
    Interleave the two behaviors and shuffle deterministically
    for all selected orbit i **do**
        Choose one scheduled view (x_{ig}, q_{ig}, y_{ig})
        Format y_{ig} with the frozen reasoning/answer response envelope
        Update θ by cross-entropy on assistant tokens only
    end for
end for
return $\theta_{\mathrm{SFT}} \leftarrow \theta$

Algorithm 3 Short-budget Orbit-GRPO with the FCO estimator

Require: Orbit-SFT policy θ_{SFT}; cardinal training orbits; $(\lambda_a, \lambda_f, \lambda_o) = (0.9, 0.1, 0.5)$; $\beta = 0.01$; clip $\epsilon = 0.2$
Ensure: Final OrbitAlign policy θ

$\theta \leftarrow \theta_{\mathrm{SFT}}$
Use mixed open-ended and structured orbits; balance invariant and equivariant records and question buckets
Initialize optimizer and cosine schedule; $\theta_{\mathrm{ref}} \leftarrow \textsc{FrozenCopy}(\theta)$
for each of 100 optimizer steps **do**
  for each of two scheduled orbit groups **do**
    Select the 0° reference and one rotating nonzero-yaw view
    For equivariant orbits, alternate changed/stable views when available
    for all selected views v **do**
      Sample two responses $p_{v,1}, p_{v,2} \sim \pi_\theta(\cdot \mid x_v, q_v)$
      for $j \in \{1, 2\}$ **do**
        Compute official $R^{\mathrm{answer}}_{v,j}$ and $R^{\mathrm{format}}_{v,j}$ in the current view
        Parse $p_{v,j}$ and set $c_{v,j} \leftarrow C_v(p_{v,j})$
        $b_{v,j} \leftarrow \mathbb{I}[R^{\mathrm{answer}}_{v,j} \geq 0.99]$; malformed outputs set $c_{v,j} = \perp$ and $b_{v,j} = 0$
      end for
    end for
    for all responses (v, j) **do**
      $R^{\mathrm{FCO}}_{v,j} \leftarrow |\Omega(v)|^{-1} \sum_{(u,l)\in\Omega(v)} b_{v,j} b_{u,l}$
      $R^{\mathrm{base}}_{v,j} \leftarrow \lambda_a R^{\mathrm{answer}}_{v,j} + \lambda_f R^{\mathrm{format}}_{v,j}$
      $\widetilde{R}^{\mathrm{base}}_{v,j} \leftarrow (R^{\mathrm{base}}_{v,j} - \mathrm{mean}_l\, R^{\mathrm{base}}_{v,l})/(\mathrm{std}_l\, R^{\mathrm{base}}_{v,l} + \epsilon_A)$
      $A_{v,j} \leftarrow 5[\widetilde{R}^{\mathrm{base}}_{v,j} + \lambda_o(R^{\mathrm{FCO}}_{v,j} - \mathrm{mean}_l\, R^{\mathrm{FCO}}_{v,l})]$
    end for
  end for
  Update θ with clipped GRPO using both groups' advantages, ϵ, and the KL penalty to θ_{ref}
end for
return θ

orbits one-to-one, rotates the exposed view across epochs, and alternates changed and stable equivariant targets when both are available. SFT is applied only to assistant response tokens, so the supervision targets the current-frame answer without treating prompt tokens as prediction targets. In particular, geometry-regenerated labels, rather than copied clean answers, supervise equivariant views.

In Algorithm 3, each logical group contains exactly two yaw views and two responses per view. The 100-step run samples mixed questions from the start. Answer and format rewards are computed in each current view; the FCO term links the views through their correctness indicators. Centering that term without standardizing it retains the partner-view factor in the advantage. Because R^{FCO} is zero whenever the current response or its cross-view partner is wrong, agreement on an incorrect canonical answer receives no positive orbit reward.

Table 12: Matched Qwen2.5-VL-7B training-signal diagnostics. Both runs use the same orbit-SFT initialization, training data, policy budget, and response-group count. Values are percentages except mean absolute advantage; changes are percentage points except its relative change. Nonzero-advantage groups contain at least one nonzero response advantage. An active centered-FCO view contains a nonzero FCO contribution after within-view centering, before objective weighting and advantage scaling.

Diagnostic	PanoEnv GRPO	FCO	Change
No-correct groups	41.00	37.00	−4.00
Zero reward-variance groups	34.50	28.00	−6.50
All-zero-advantage groups	43.50	37.50	−6.00
Nonzero-advantage groups	56.50	62.50	+6.00
Nonzero-advantage responses	38.25	41.50	+3.25
Mean absolute advantage	1.352	1.614	+19.4%
Active centered-FCO views	0.00	17.25	+17.25

Table 13: **E2 performance by yaw angle on InternVL3-8B.** Differences are OrbitAlign (100 RL steps) minus PanoEnv GRPO (orbit data) on primary seam-clean changed-label test records, in percentage points. Positive accuracy changes and negative stale-rate changes indicate improvement. Stale rates use parseable changed-label pairs within each angle.

Yaw	Records	ΔChanged↑	ΔStale↓
45°	56	+53.57	−8.93
90°	106	+53.77	−27.36
135°	129	+56.59	−28.68
180°	137	+35.77	−24.40
225°	115	+44.35	−22.57
270°	97	+41.24	−8.67
315°	62	+56.45	+0.87

D Supplementary results and analyses

D.1 FCO training-signal diagnostics

We compare Qwen2.5-VL-7B training traces from a PanoEnv GRPO control and an FCO run with matched orbit-SFT initialization, training data, and 100-step policy budget. Each trace contains 200 response groups of four responses, comprising 400 views and 800 responses. The analysis uses the same number of groups from each run.

Table 12 reports reward differentiation and advantage coverage. No-correct groups contain no response with full answer reward; reward variance is computed over the four responses in each group. FCO reduces all-zero-advantage groups from 43.50% to 37.50% and increases nonzero-advantage responses from 38.25% to 41.50%. The centered FCO term is nonzero for 138 of 800 responses across 69 of 400 views (17.25%), with a mean absolute raw contribution of 0.05875 over all responses before objective weighting and advantage scaling. Mean absolute advantage increases by 19.4%. These trace-level diagnostics show that the coupled term survives view-local centering and is accompanied by broader policy-gradient signal coverage.

D.2 Performance by yaw angle

Table 13 compares the 100-step OrbitAlign checkpoint with the PanoEnv GRPO orbit-data control on InternVL3-8B, partitioning the 702 primary seam-clean changed-label E2 records by yaw angle. Changed-label accuracy improves at every evaluated angle, with gains ranging from 35.77 percentage points at 180° to 56.59 points at 135°. The improvement is therefore not confined to a single rotation. Stale-answer rates decrease at six of seven angles; at 315°, stale rate increases by 0.87 points despite a 56.45-point accuracy gain, distinguishing answer repetition from current-frame correctness.

Table 14: **Unseen-angle generalization.** Clean accuracy uses the matched 300-question diagonal cohort; orbit metrics use 49 E1 and 24 E2 source questions at $45°$, $135°$, $225°$, and $315°$, none of which is used in training. The PanoEnv GRPO control is initialized from the same orbit-SFT checkpoint as OrbitAlign, so the two trained rows are compute-matched within this supplementary comparison. These 1,968-step checkpoints do not isolate the FCO reward. All values are percentages.

Method	Clean↑	E1 Transf.↑	E1 Robust↑	E2 Transf.↑	E2 Exp.-Change↑	E2 Stale↓
Qwen2.5-VL-3B						
Qwen2.5-VL-3B	40.67	47.45	36.73	41.25	31.91	69.77
Orbit SFT + PanoEnv GRPO	47.67	50.51	28.57	40.00	31.91	59.57
OrbitAlign-3B (ours)	**51.00**	**55.61**	**44.90**	**68.75**	**70.21**	**57.45**
Qwen2.5-VL-7B						
Qwen2.5-VL-7B	48.67	46.94	24.49	47.50	44.68	**48.94**
Orbit SFT + PanoEnv GRPO	51.00	**52.55**	**32.65**	48.75	48.94	57.45
OrbitAlign-7B (ours)	**56.00**	50.00	**32.65**	**66.25**	**65.96**	51.06

Table 15: **Coordinate-free E2 diagnostic on Qwen2.5-VL-3B.** Bounding-box coordinates are replaced by visual A/B/C markers. Marker Clean uses 84 source questions; transformed metrics use seam-clean yaw records. The PanoEnv GRPO control is initialized from the same orbit-SFT checkpoint as OrbitAlign. These 1,968-step checkpoints do not isolate the FCO reward. All values are percentages.

	Source	E2 — Marker-grounded			
Method	Clean↑	Transf.↑	Cond.↑	Changed↑	Stale↓
Base	32.14	32.79	49.69	27.55	52.21
Orbit SFT	40.48	37.68	47.18	32.51	54.55
Orbit SFT + PanoEnv GRPO	42.86	43.38	49.28	39.12	51.24
OrbitAlign	**48.81**	**50.10**	**55.04**	**47.66**	**49.31**

D.3 Checkpoint selection

Development data are used for training choices; frozen-test results are not used to select checkpoints or tune the reward. The 100-step ablation compares arms at the same optimizer-step budget.

D.4 Unseen angles and coordinate-free prompts

Tables 14 and 15 report supplementary transfer diagnostics from the 1,968-step OrbitAlign checkpoints, separate from the 100-step FCO ablation. The unseen-angle cohort evaluates the four diagonal headings excluded from cardinal-angle training; the marker-grounded diagnostic replaces textual bounding-box coordinates with visual A/B/C markers while preserving visual instance references.

D.5 Additional frozen diagnostic models

Table 16 reports frozen-test diagnostic models that are not central to the Qwen2.5-VL training comparison. These rows use the same 2,991 clean records, 1,373 E1 complete orbits, and 948 primary E2 records as the main held-out table.

The diagnostic conclusion transfers across families: mean transformed E1 accuracy remains much higher than strict full-orbit correctness. Thus the clean-to-orbit gap is not specific to one Qwen checkpoint or one image processor.

D.6 InternVL3-8B implementation contract

InternVL3-8B uses 448×448 dynamic tiles, between one and twelve tiles, and an additional thumbnail whenever more than one tile is produced. Tiles are ordered row-major and normalized with the model's ImageNet statistics. This processor is intentionally not replaced by the Qwen processor: the cross-family experiment tests whether orbit training transfers while retaining each architecture's native visual interface.

Table 16: Additional frozen-test diagnostic models. PanoWorld is retained as a panorama-specific reference, but its results should be interpreted cautiously because its model-native output produces substantially more parser failures. All metric entries are percentages.

Model	Clean	E1 — invariant			E2 — equivariant		
	Acc.↑	Mean↑	Flip↓	Robust↑	Cond.↑	Changed↑	Stale↓
Qwen3-VL-8B	48.85	51.67	38.00	32.07	58.86	28.63	39.43
InternVL3-8B	45.30	52.40	36.87	32.83	36.27	19.09	44.36
PanoWorld-9B H*	40.32	48.22	41.34	29.79	43.50	23.08	47.49

Table 17: Post-training settings for the short-budget Qwen ablation.

Setting	Value
LoRA rank / alpha / dropout	16 / 32 / 0.05
Orbit-SFT learning rate / 3B checkpoint step	2×10^{-5} / 1,900
GRPO optimizer steps	100
GRPO learning rate	5×10^{-6}
Generations per logical group	4
Yaw views / responses per view	2 / 2
KL coefficient β / clip ϵ	0.01 / 0.2
Maximum completion length	256
Answer / format / FCO weights	0.9 / 0.1 / 0.5
Advantage scale	5.0
Learning-rate schedule / warmup	cosine / 0.05

For InternVL post-training, the vision encoder and multimodal projector are frozen. LoRA is applied to language-side linear layers with rank 16, alpha 32, and dropout 0.05. The held-out OrbitAlign arm starts from an orbit-SFT adapter and runs 100 GRPO steps with four generations per group, two yaw views and two responses per view, $\beta = 0.01$, clipping $\epsilon = 0.2$, and maximum completion length 256. The answer, format, and FCO reward weights are 0.9, 0.1, and 0.5, respectively.

D.7 Shared post-training hyperparameters

Table 17 records the central Qwen and InternVL settings. Architecture-specific memory adaptations change only the backward microbatch, not the logical four-response group or the response budget.

The 100-step OrbitAlign RL stage adds the centered FCO term after standardizing the base reward. Each logical group contains four responses, organized as two yaw views with two responses per view.

D.8 Training cost

Table 18 reports training cost for the frozen recipes. OrbitAlign includes 1,968 orbit-SFT steps followed by 100 RL steps; PanoEnv GRPO uses two 984-step RL phases. GPU-hours are measured as wall time multiplied by allocated GPUs. Checkpoint selection and validation are excluded for both methods.

E Qualitative orbit examples

Table 19 gives three real validation examples from the geometry-regenerated E2 manifest. The first is visualized with its actual bounding boxes in Figure 5. The examples cover binary, multi-term cardinal, and triplet-extreme answers. In every case, copying the clean answer would be wrong even though the panorama contains the same source pixels.

For the first example, the source prompt contains stairs box $[1737, 463, 1794, 520]$ and chimney box $[2142, 690, 2269, 829]$. After the exact 90° shift, the prompt contains $[2377, 463, 2434, 520]$ and $[222, 690, 349, 829]$, respectively. The official source target is "Yes," whereas rotating the verified

Table 18: Training cost on the frozen recipes, in H200 GPU-hours. PanoEnv GRPO is the base-initialized `standard_panoenv` two-phase run (2×984 steps); OrbitAlign is Orbit-SFT followed by 100 FCO GRPO steps. GPU-hours are wall time multiplied by allocated GPUs; all runs are single-GPU except the 7B RL stage (2 GPUs). Speed-up is the ratio of baseline GPU-hours to total OrbitAlign GPU-hours. Checkpoint selection and validation are excluded for both methods.

Backbone	Recipe	Steps	GPU-h	Speed-up
Qwen2.5-VL-3B	PanoEnv GRPO	1968	12.87	—
	Orbit-SFT	1968	2.33	
	+ FCO GRPO	100	0.48	
	OrbitAlign (total)	2068	**2.81**	4.6×
Qwen2.5-VL-7B	PanoEnv GRPO	1968	18.05	—
	Orbit-SFT	1968	3.80	
	+ FCO GRPO	100	0.59	
	OrbitAlign (total)	2068	**4.39**	4.1×
InternVL3-8B	PanoEnv GRPO	1968	12.98	—
	Orbit-SFT	1968	2.17	
	+ FCO GRPO	100	0.30	
	OrbitAlign (total)	2068	**2.47**	5.2×

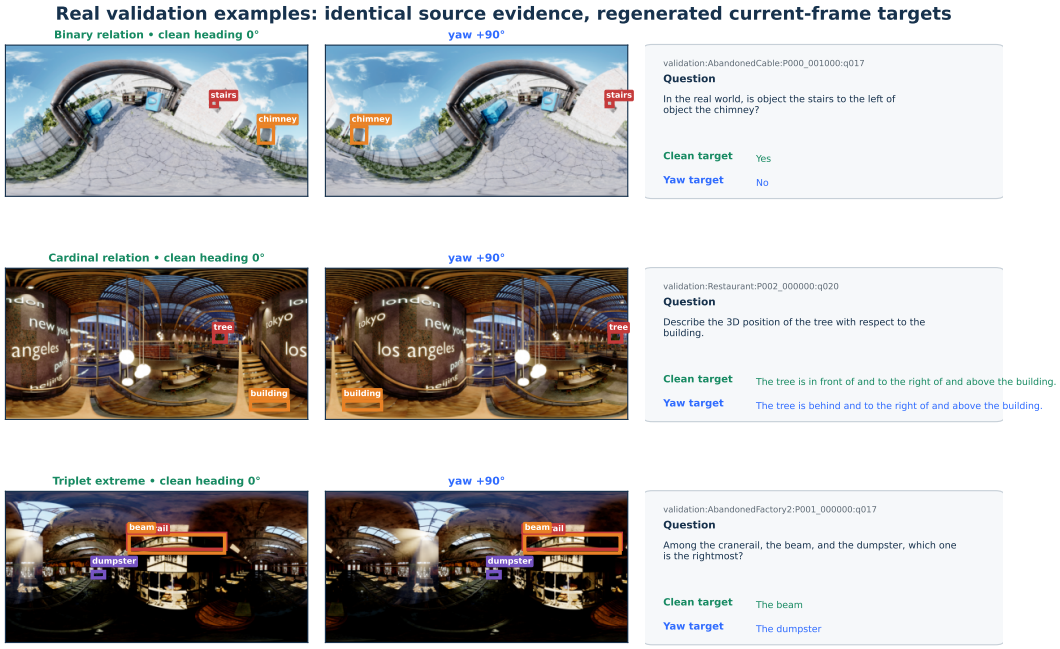

Figure 5: **Qualitative E2 validation orbits.** Each row shows a clean ERP and its exact 90° cyclic shift with the benchmark's actual related-object boxes. The right column gives the corresponding source and geometry-regenerated current-frame targets. Pixel content is preserved, but the correct egocentric answer changes.

centroids and regenerating the relation gives "No." This example demonstrates why transformed coordinates and transformed labels must be produced together.

Table 19: Real geometry-regenerated validation examples. Questions are shortened only by omitting the displayed bounding-box coordinates; object names and answers are unchanged.

Source record / yaw	Question	Clean target	Current-frame target
`AbandonedCable: P000_001000: q017`, 90°	Are the stairs to the left of the chimney?	Yes	No
`Restaurant: P002_000000: q020`, 90°	Describe the tree's 3D position with respect to the building.	In front of, right of, and above	Behind, right of, and above
`AbandonedFactory2: P001_000000: q017`, 90°	Among the cranerail, beam, and dumpster, which is rightmost?	The beam	The dumpster

F SUPPLEMENTARY REPRODUCIBILITY CHECKLIST

The release package is organized so every reported number can be regenerated without editing the manuscript by hand. It contains:

- frozen clean, E1, and E2 manifests with SHA-256 hashes;
- the behavior-policy configuration and manual-review ledger;
- deterministic ERP-roll, bounding-box rewrite, and geometry-label regeneration code;
- the population lock, seam/exclusion audit, and zero-yaw reproduction checks;
- model revisions, native processor settings, LoRA configurations, training seeds, and phase manifests;
- raw predictions, parser outputs, per-model metric JSON files, and scene-bootstrap scripts; and
- the common paper-metric extractor used to populate the frozen-test tables and teaser figure.

The PanoEnv test answers are public. Consequently, held-out status is a procedural guarantee: populations, arms, checkpoints, endpoints, and the reporting commitment are frozen before test inference. Any post-registration supplementary model arm is labeled as such and cannot replace a primary arm or change the pre-registered conclusions.

G CROSS-DATASET EVALUATION: PANOENV-S2D3D

This supplementary transfer analysis uses the 1,968-step OrbitAlign checkpoints. It does not measure the FCO effect in the 100-step ablation.

G.1 CONSTRUCTION

PanoEnv-S2D3D applies PanoEnv's released question-generation pipeline to the Stanford 2D-3D-S corpus (Armeni et al., 2017). We construct this evaluation set using new scenes while retaining PanoEnv's task definition; the set is not an upstream release.

Source corpus. The imagery is Stanford 2D-3D-S, real indoor building scans with equirectangular RGB, sensor depth, human-annotated semantic instances, and surface normals across seven building areas. PanoEnv, by contrast, is rendered from synthetic TartanAir environments. Every frame must carry all four modalities; missing modalities trigger an error rather than sample exclusion. Native panoramas are 4096×2048 and are resampled to the 2560×1280 ERP resolution of the PanoEnv benchmark, so that the yaw shifts used later are integer column rolls at the same width.

Scene selection. Panoramas are selected deterministically by stratifying on (area, room type), which prevents a single area or the dominant office/hallway pair from dominating the benchmark.

The dev and benchmark sets are drawn room-disjoint: the benchmark is selected first, and generator debugging used only panoramas from rooms absent from it, so no nearby view of a benchmark room influenced construction. Of 71 candidate panoramas, upstream accepted 60; the remainder had fewer eligible objects than upstream's minimum for question generation. The accepted set covers 47 distinct rooms and 11 room types (office, conference room, hallway, copy room, pantry, storage, lounge, lobby, auditorium, open space, and WC).

Object extraction. Objects come from the released semantic instance masks rather than from any detector. A mask is eligible when its true mask area is at least 900 pixels and it fills at least 10% of its bounding box, which rejects masks that are nominally large but thin or scattered. Depth statistics, volume, and 3D centroids are computed from the sensor depth map, and each object carries the polar and seam flags that the ERP-aware sampler uses. These are the fields upstream expects from a TartanAir frame, recomputed from real sensor data.

Question generation. Questions are produced by upstream's unmodified `qa_generation` module. Upstream resolves `import config` at runtime, so we install a configuration shim in `sys.modules` before import; every metric and gating constant in that shim is copied verbatim from upstream's `config.py`, including the $0.5\,\mathrm{m}$ relative-position significance threshold that later defines which directional components a rotated answer contains. Only the environment-category tables are replaced, because TartanAir environment identity has no S2D3D counterpart. The loader verifies the upstream function definitions by AST and records the SHA-256 of the generator source in the build report to document implementation identity.

Quality filtering. Upstream's comparison sampler retries 20 times to find objects that are not mutually contained and then falls back to an unvalidated random sample. We retain the generator but fail closed on that fallback, discarding any comparison question whose objects reach 0.90 directional bounding-box containment. Scenes yielding fewer than ten questions after filtering are dropped entirely. The result is 1,416 questions over 24 sub-categories: 708 multiple-choice, 414 true/false, and 294 open-ended.

Answer-distribution audit. A structural audit flags sub-categories with fewer than 25 questions or a majority answer above 85%. One sub-category is degenerate on this corpus: `env_attribute` has the single answer "Indoors" for all 60 of its questions, because S2D3D is entirely indoor whereas TartanAir spans outdoor scenes. Four further environment sub-categories are flagged for separate reporting for the same reason. We retain all of them in the clean denominator rather than silently reshaping the population; they are 300 of 1,416 questions and affect every arm identically, so they inflate absolute clean accuracy slightly but not the differences between arms.

Yaw derivation. The transformed records use the same machinery as the main benchmark: the same seven angles, the same frozen behavior policy (`panoenv-yaw-v1`, frozen after manual review), the same geometry regenerator (`panoenv-relative-yaw-equivariance-v1`), the same zero-yaw reproduction check, and the same official scorer. The policy assigns 1,039 questions to the invariant contract and 377 to the equivariant relative-position contract. Of the latter, 237 camera-view questions are excluded because the E2 regenerator verifies centroid-based relations only and camera-view labels would require semantic-mask face occupancy, leaving 140 equivariant source questions. No source question contains a seam-enclosing object before transformation.

Reported populations. Table 20 gives the counts. As on the main benchmark, the reported E1 and E2 metrics use the seam-clean primary populations, and the all-record counts are the sensitivity populations. Seam exposure is higher here than on the PanoEnv test split: 662 of 1,039 invariant orbits contain at least one seam-enclosing transformed prompt, leaving 377 complete seam-clean orbits, and only 40 of 140 equivariant orbits are seam-clean at all seven angles. Retaining every record lowers OrbitAlign-7B's E1 full-orbit accuracy from 46.68% to 43.60% and leaves its E2 transformed accuracy essentially unchanged (66.71% to 67.04%), so the ranking in Table 21 does not depend on the seam filter.

Evaluation protocol. Scoring uses the pinned official PanoEnv evaluator and prompt. Records carry the split identity `s2d3d` so the scorer joins predictions against this benchmark rather than the

Table 20: PanoEnv-S2D3D populations. Reported metrics use the primary seam-clean populations; all-record counts are the sensitivity populations.

Population item	Count
Panoramas / rooms / areas	60 / 47 / 7
Clean questions	1,416
Multiple-choice / true-false / open-ended	708 / 414 / 294
E1 invariant source questions	1,039
Complete seam-clean orbits (primary)	377
Primary transformed records	2,639
All transformed records	7,273
E2 equivariant source questions	140
Primary seam-clean records	799
Changed-label / stable-label	584 / 215
Complete seam-clean seven-angle orbits	40
All transformed records	980

PanoEnv validation population. Decoding is greedy and deterministic, as on the main benchmark, but this evaluation allows 512 new tokens and uses scaled-dot-product attention, whereas the main benchmark uses 256 tokens and flash attention; the setting is identical across arms. All trained arms here are the seed 20260810 checkpoints, one per arm.

Scope of the cross-dataset evaluation. The benchmark changes the visual domain from synthetic renders to real sensor scans, the object vocabulary from TartanAir categories to S2D3D furniture and building elements, and the depth source from rendered ground truth to sensor measurement. It does not change the relation rule, the significance threshold, the answer semantics, or the scorer. Results therefore measure whether the learned transformation behavior transfers across domains and scenes under a fixed label convention, not whether it generalizes to an independently authored convention.

G.2 Results

OrbitAlign achieves the highest clean accuracy on all three backbones, improving over the base models by 12.29 points on Qwen2.5-VL-3B (45.97 to 58.26), 14.69 on Qwen2.5-VL-7B (46.33 to 61.02), and 10.53 on InternVL3-8B (50.35 to 60.88), and exceeding PanoEnv-GRPO by 10.24, 7.70, and 8.06 points respectively. On E1, OrbitAlign exceeds PanoEnv-GRPO on the two Qwen backbones in both transformed mean and full-orbit robust accuracy (58.58/44.03 and 62.75/46.68), while on InternVL3-8B PanoEnv-GRPO achieves higher accuracy (59.83 vs. 59.64 mean, 43.24 vs. 42.44 robust). Answer-flip rates are close across the trained arms, between 26.49 and 29.29, with PanoEnv-GRPO lowest on Qwen2.5-VL-3B and InternVL3-8B and OrbitAlign lowest on Qwen2.5-VL-7B. Differences between the methods are larger on E2. PanoEnv-GRPO produces limited changes in equivariant accuracy relative to its base model and raises stale rate on Qwen2.5-VL-7B and InternVL3-8B (48.19 to 53.08 and 43.13 to 49.66), indicating that it repeats the clean answer more often on records whose correct label changes under the transform. OrbitAlign instead improves correct-conditioned equivariance to 67.23, 69.79, and 76.24, raises accuracy on label-changing records by 34.07, 36.30, and 44.17 points over PanoEnv-GRPO, and reduces stale rate to 35.62, 32.02, and 22.09. Consistent with the primary benchmark, PanoWorld-9B with H^* performs below the untrained Qwen3-VL-8B baseline on every metric, including flip rate (46.65 vs. 37.02) and changed-label accuracy (19.86 vs. 30.31).

Table 21: **Cross-dataset evaluation on PanoEnv-S2D3D.** Official PanoEnv scoring on the frozen set. Clean is accuracy on all 1,416 untransformed questions. E1 uses the 377 complete seam-clean invariant orbits and their 2,639 transformed records: Mean is transformed accuracy, Flip the canonicalized answer-flip rate relative to the clean prediction, and Robust requires the clean view and all seven transformed views to be correct. E2 uses the 799 seam-clean transformed records from 140 source questions, of which 584 change the label: Cond. is correct-conditioned equivariance, Changed the accuracy on records whose correct label changes, and Stale the rate of repeating the clean answer on those records. PanoEnv-GRPO is the base-initialized two-phase run used as the primary baseline in the main experiments; OrbitAlign is the complete pipeline at the 1,968-step budget, separate from the 100-step FCO ablation. All trained arms use seed 20260810. Lower is better for Flip and Stale; bold marks the best of the three arms per backbone.

			E1 invariant			E2 equivariant		
Family	Method	Clean↑	Mean↑	Flip↓	Robust↑	Cond.↑	Changed↑	Stale↓
Qwen2.5-VL-3B	Base	45.97	46.87	31.68	31.03	46.53	24.49	61.39
	PanoEnv-GRPO	48.02	55.06	**26.49**	37.93	47.17	24.66	56.51
	OrbitAlign (ours)	**58.26**	**58.58**	26.83	**44.03**	**67.23**	**58.73**	**35.62**
Qwen2.5-VL-7B	Base	46.33	47.10	33.35	25.46	51.75	29.28	48.19
	PanoEnv-GRPO	53.32	57.60	27.40	39.52	47.77	30.31	53.08
	OrbitAlign (ours)	**61.02**	**62.75**	**26.53**	**46.68**	**69.79**	**66.61**	**32.02**
InternVL3-8B	Base	50.35	57.48	29.82	37.40	40.44	25.51	43.13
	PanoEnv-GRPO	52.82	**59.83**	**26.75**	**43.24**	40.82	27.23	49.66
	OrbitAlign (ours)	**60.88**	59.64	29.29	42.44	**76.24**	**71.40**	**22.09**
Qwen3-VL-8B	Base	50.21	56.23	37.02	34.48	52.76	30.31	41.95
PanoWorld-9B	H*	42.30	51.38	46.65	29.97	50.56	19.86	45.07