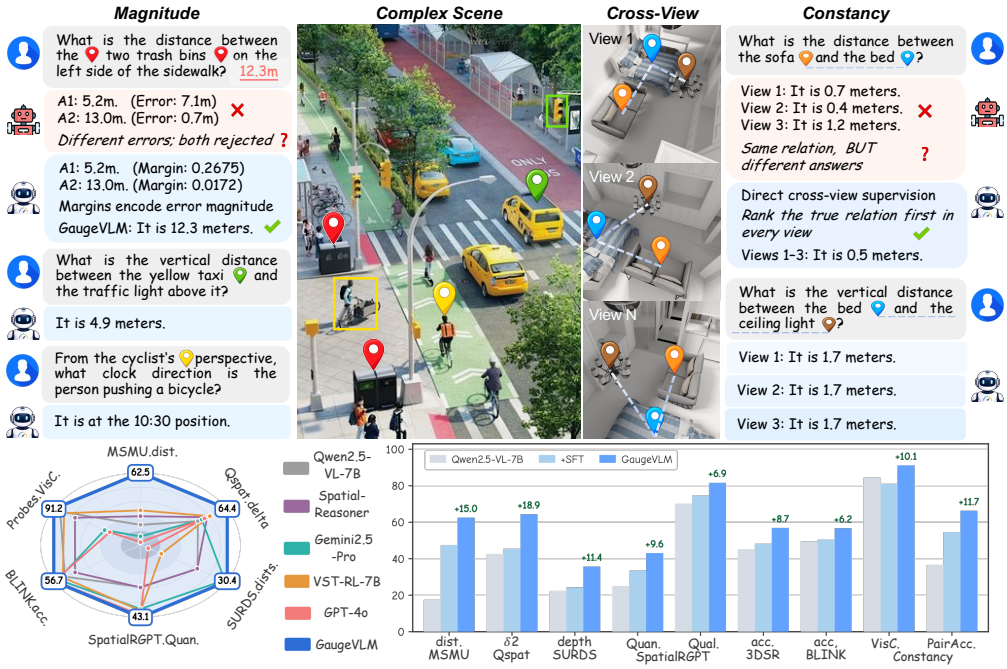
GAUGEVLM: STRUCTURING SPATIAL SUPERVISION WITH MEASURED GEOMETRIC INTERVENTIONS

Hongbo Wang[1,2] **Zihan Lin**[1,3] **Wenkui Yang**[1,3] **Shiran Ge**[1,2,4] **Yuang Ai**[5]
Jie Cao[1,2] **Huaibo Huang**[1,2] **Ran He**[1,2,3*]

[1] NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
[4] National University of Singapore
[5] The Chinese University of Hong Kong

https://github.com/wafer-bob/GaugeVLM



ABSTRACT

Vision-language models (VLMs) can contradict themselves across views of the same spatial relation and fail to respond when that relation changes. Addressing these failures requires supervision that captures error magnitude and geometric dependencies across observations, both of which remain implicit in training on individual answers or ordinal preferences. Therefore, we introduce **GaugeVLM**, which makes this structure explicit through controlled object and camera interventions in explicit 3D scenes, producing linked observations with measured differences between spatial relations and shared truths across views. To translate this structure into learning signals, its core objective, **GaugeDPO**, converts measured errors into preference margins, directly supervises correct canonical rankings across views, and links intervention-induced answer-odds contrasts to measured relation changes with view-specific scales. Our analysis bounds canonical prediction error and establishes that the cross-view and intervention constraints

*Corresponding author.

can be jointly satisfied. Empirically, GaugeVLM improves all 10 established spatial metrics over supervised fine-tuning across three VLM backbones, with the main 7B model gaining 15.0 and 18.9 percentage points on MSMU distance and QSpatial+, respectively. These gains also extend to autonomous driving and embodied reasoning, demonstrating the robust generalization across domains.

1 INTRODUCTION

Despite rapid advances in vision-language models (VLMs) (OpenAI, 2023; Chen et al., 2024c; Bai et al., 2025), spatial reasoning remains brittle. Even after large-scale supervised fine-tuning (SFT) on spatial question answering (Chen et al., 2024a; 2025; Ogezi & Shi, 2025; Helu et al., 2025), VLMs still struggle with distances, directions, and relative relations (Liao et al., 2024; Ma et al., 2025; Fu et al., 2024), especially when relations must hold across viewpoints (Li et al., 2025; Yang et al., 2025; Yeh et al., 2026; Wang et al., 2026). These failures motivate examining the supervision signal alongside data and model capacity. Supervision based only on target answers tells the model what the answer is, but not how far an error is from the truth or whether the same relation should remain consistent across viewpoints.

Preference optimization offers a way to use incorrect answers as supervision, but the ordinal labels used by standard Direct Preference Optimization (DPO) (Rafailov et al., 2023) do not quantify geometric error. For the same true distance, predictions off by ten centimeters and ten meters receive the same rejected label despite their different error magnitudes. This motivates an instance level offset that explicitly reflects how far a rejected answer deviates from the truth. Likewise, since the same allocentric error has the same geometric magnitude across camera views, its assigned offset should also remain unchanged. The central question is therefore how to derive offsets that capture error magnitude while remaining invariant to viewpoint.

Geometry provides the missing scale for spatial preferences. **GaugeDPO** grounds each preference offset in measured error, making supervision sensitive to error magnitude while assigning the same offset to the same allocentric error across views. The measurement comes from the construction itself. **GaugeVLM** perturbs known relations in explicit 3D scenes and verbalizes the resulting alternatives as rejected answers, making their geometric deviations directly available without an intermediate reward estimate. These deviations should depend on geometry regardless of angular origin or distance units. We therefore build a bounded metric from circular direction differences and distance ratios, with explicit saturation. Used as a preference offset, this metric translates the severity absent from ordinal labels into the separation encouraged during training. To understand how it does so, we analyze the offset's local gradient effect and characterize the allocation of preference separation under an equal weight budget, then examine the learned mean gaps empirically.

Beyond measuring individual errors, we develop a 3D scene engine for GaugeVLM to connect observations through controlled changes in the scene. Object interventions produce measured changes in the true relation, while camera interventions preserve it under a fixed object configuration and anchor frame. These linked observations specify when predictions should change and when they should remain constant. Accordingly, direct supervision enforces correct canonical rankings across views, while an intervention profile links answer odds contrasts to measured relation changes without requiring equal confidence across views. The same distinction guides evaluation. For views sharing a fixed relation, we show that half the largest disagreement in a common frame lower bounds the largest prediction error. Object interventions instead require checking predictions against their changed truths. The engine supplies **Gauge-50K** for training and a probe within **Constancy-Bench**.

Across three backbone families, GaugeVLM improves all ten established spatial metrics over SFT. The main 7B model gains 15.0 and 18.9 percentage points on MSMU distance and QSpatial+, respectively, and leads the listed prior open spatial VLMs on nine of ten metrics under the evaluation. Gains extend to driving scenarios, including an 11.4 percentage point improvement on SURDS depth. Matched ablations support the value of measured margins over constant, shuffled, and estimated reward margins. With only 25% of the preference data, GaugeDPO surpasses DPO trained on the full preference dataset across all five evaluated metrics. Further evaluations show improved correctness across views and generalization to unseen intervention magnitudes and combinations.

2 SPATIAL PREFERENCE NEEDS A GEOMETRIC MARGIN

2.1 GEOMETRY-GROUNDED OFFSETS

DPO learns ordinal preferences, SimPO, IPO, AlphaDPO and ODPO regulate gaps or offsets (Rafailov et al., 2023; Meng et al., 2024; Gheshlaghi Azar et al., 2024; Wu et al., 2025; Amini et al., 2024); mDPO adds image comparisons (Wang et al., 2024). Spatial methods use QA, grounding and graded preferences (Chen et al., 2024a; 2025; Ogezi & Shi, 2025; Helu et al., 2025; Cheng et al., 2024; Shen et al., 2025), transformed-input coupling or smooth numerical rewards (Wang & Ling, 2026; Jiao et al., 2026). Ordinal labels do not quantify errors. We map controlled 3D perturbations to offsets with a bounded, unit-invariant metric. Moreover, cross-view and counterfactual studies (Bhat & Yamasaki, 2026; Vellamcheti et al., 2026; Huang et al., 2026; Yu et al., 2026) motivate separate tests of viewpoint constancy and correctness before and after target relocation.

2.2 FORMAL SETUP AND DESIGN REQUIREMENTS

Quantization maps a scene's physical relation in a fixed anchor frame to $r^\star = Q(r^\star_{\text{phys}}) \in \mathcal{R}_Q \subset \mathcal{R}$. Camera v supplies image–prompt input x_v. On $\mathcal{R}_Q$, verbalizer Φ and parser Π satisfy $\Pi(\Phi(r, v)) = r$; invalid answers map to $\perp$. Response-side pairs compare true and distinct rejected labels under one input; camera-relative binary probe labels are separate. For chosen–rejected triples $(x, y_w, y_l) \sim \mathcal{D}$ and $\beta > 0$, define the reference-relative score $\rho_\theta(x, y) = \beta \log[\pi_\theta(y \mid x)/\pi_{\text{ref}}(y \mid x)]$ with positive reference probabilities. With $\sigma(z) = (1 + e^{-z})^{-1}$,

$$\mathcal{L}_{\text{Margin-DPO}} = -\mathbb{E}_{\mathcal{D}} \log \sigma(\rho_\theta(x, y_w) - \rho_\theta(x, y_l) - \gamma), \tag{1}$$

The prescribed offset $\gamma \geq 0$ shifts the gap required at a fixed loss level (Amini et al., 2024; Wu et al., 2025), with $\gamma = 0$ recovering DPO.

For spatial relations, the offset should reflect geometric error, while the policy's predictions should preserve the true relation across cameras. These impose distinct requirements on margins and predictions. Let $\mathcal{E} : \mathcal{R} \times \mathcal{R} \to \mathbb{R}^k_{\geq 0}$ collect deviations, with $\mathcal{E}(r, r^\star) = 0$ iff $r = r^\star$.

(R1) Magnitude. The margin $\tilde{\gamma} : \mathcal{R}^2 \times V \to \mathbb{R}_{\geq 0}$ factors as $\tilde{\gamma}(r^-, r^\star, v) = \gamma(\mathcal{E}(r^-, r^\star))$, where γ is componentwise non-decreasing and strictly increasing before saturation, with $\gamma(e) = 0$ iff $e = 0$. Thus each displacement receives the same assigned margin across cameras.

(R2) Constancy. With the object configuration and anchor frame fixed, an ideal policy satisfies $\Pi(y_v) = r^\star$ for every identifiable view $v \in V$. Correctness implies agreement; agreement alone permits consistent error. Matching training statistics does not establish this property.

Camera invariance alone is insufficient: a constant margin fails (R1). We characterize coordinate-invariant deviations and use a metric to assign offsets, supplying geometric information absent from deterministic ordinal labels.

3 GAUGEDPO: LEARNING FROM GEOMETRIC INTERVENTIONS

Object interventions alter relations and camera changes preserve their truth. GaugeDPO turns this geometry into preference and prediction margins, then structures canonical answer support across interventions.

3.1 FROM GEOMETRY TO MATCHED PREFERENCES

A shared geometric metric. Let $r = (c, d) \in \mathcal{R} := \mathcal{C}_{12} \times \mathbb{R}_{>0}$, with $\mathcal{C}_{12} = \mathbb{Z}/12\mathbb{Z}$. Clock hours advance clockwise about gravity from the anchor's canonical front at 12 o'clock; d is centroid distance in meters. Camera-independent errors should respect circular shifts and changes of units:

$$D_c(c + k, c^\star + k) = D_c(c, c^\star),\ D_d(\mu d, \mu d^\star) = D_d(d, d^\star), \quad k \in \mathcal{C}_{12},\ \mu > 0, \tag{2}$$

We therefore use clock difference and distance ratio, with shortest-arc and capped log-distance:

$$d_{\text{ring}}(c, c^\star) = \tfrac{1}{6}|c - c^\star|_{12}, \quad d_{\text{rel}}(d, d^\star) = \tfrac{1}{\kappa} \min(|\log(d/d^\star)|, \kappa), \tag{3}$$

$$m(r, r^\star) = \alpha\, d_{\text{ring}}(c, c^\star) + (1 - \alpha)\, d_{\text{rel}}(d, d^\star), \tag{4}$$

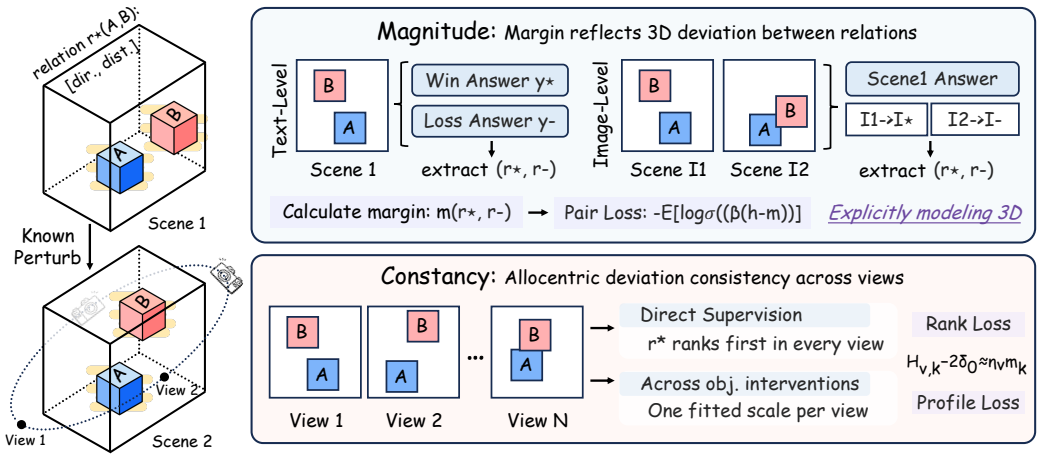


Figure 1: **GaugeDPO turns measured 3D interventions into supervision.** *Left,* controlled perturbations define the margin $m(r^\star, r^-)$. *Top,* it shifts response- and image-side pair losses. *Bottom,* direct supervision and intervention profiles enforce correct, view-consistent predictions.

Here $|\cdot|_{12} \in \{0, \dots, 6\}$, $\kappa > 0$ and $\alpha \in (0, 1)$. The metric $m \in [0, 1]$ satisfies (R1) with $\mathcal{E} = (d_{\text{ring}}, d_{\text{rel}})$, $\gamma(e) = \beta[\alpha e_c + (1-\alpha)e_d]$ and $\widetilde{\gamma} = \beta m$.

Matched response and image changes. Let Π extract the terminal relation from the multi-step response $\Phi(r, v)$, with representable domain $\mathcal{R}_Q \subset \mathcal{R}$. A displacement $\delta \in \mathcal{C}_{12} \times \mathbb{R}_{>0}$ maps $r^\star = (c^\star, d^\star) \in \mathcal{R}_Q$ to $r^- = (c^\star + \delta_c, d^\star \delta_d) \neq r^\star$ in $\mathcal{R}_Q$, fixing geometric error. Response-side pairs compare $\Phi(r^\star, v)$ and $\Phi(r^-, v)$ on one input; image-side pairs change the scene to r^-, fixing the question and response $\Phi(r^\star, v)$. Symmetry gives the same rejected error. Margins measure terminal error; preference scores use full responses. For fields $J \in c, d, c, d$ predeclared, Π_J outputs in $\mathcal{R}_J = \prod_{j \in J} \mathcal{R}_j$, where $\mathcal{R}_c = \mathcal{C}_{12}$ and $\mathcal{R}_d = \mathbb{R}_{>0}$. For $u, u^\star \in \mathcal{R}_J$, use

$$m_J(u, u^\star) = \sum_{j \in J} a_j D_j(u_j, u_j^\star), \quad (a_c, a_d) = (\alpha, 1-\alpha), \quad (D_c, D_d) = (d_{\text{ring}}, d_{\text{rel}}), \tag{5}$$

Each m_J is a metric satisfying (R1) with unchanged weights; $m_{\{c,d\}} = m$ and $\Pi_{\{c,d\}} = \Pi$. Tasks retain their response formats.

Proposition 1 (Constancy certificate)**.** *Let $G = \{x_v\}_{v=1}^{V_G}$, $V_G \geq 2$, contain camera views of a fixed object configuration and anchor frame, sharing truth $r_G^\star$. For responses $\{y_v\}$, let $V_{\text{ok}} = \{v : \Pi(y_v) \neq \perp\}$ and $a_v = \Pi(y_v)$ on valid views. If $V_{\text{ok}} \neq \emptyset$,*

$$\max_{v \in V_{\text{ok}}} m(a_v, r_G^\star) \geq \tfrac{1}{2} \max_{u, v \in V_{\text{ok}}} m(a_u, a_v), \tag{6}$$

The factor $\frac{1}{2}$ is optimal; agreement alone does not certify correctness.

3.2 Learning Magnitude and Correctness Across Views

Magnitude-aware preferences. We initialize π_θ from SFT and freeze a copy as π_{ref}. Records $z = ((x_w, y_w), (x_l, y_l))$, with image–question inputs x, split into measured $\mathcal{D}_{\text{geom}}$ and ordering-only $\mathcal{D}_{\text{bin}}$. Measured pairs must parse correctly; using each branch's own serialized truth, we require

$$m_{J_z}\big(\Pi_{J_z}(y_w), r^\star_{w,J_z}\big) = 0, \qquad m_z := m_{J_z}\big(\Pi_{J_z}(y_l), r^\star_{l,J_z}\big) > 0, \tag{7}$$

Ordering-only pairs, including vertical comparisons, require a valid preference, not an exact preferred answer. Both use $h_\theta(z) = \log\big(\pi_\theta(y_w \mid x_w)/\pi_{\text{ref}}(y_w \mid x_w)\big) - \log\big(\pi_\theta(y_l \mid x_l)/\pi_{\text{ref}}(y_l \mid x_l)\big)$. This is a within-input response contrast or an mDPO-style image contrast (Wang et al., 2024); the latter does not identify a cross-input reward difference. All sequence scores sum answer tokens through termination, excluding prompt and padding. We set

$$\gamma_z = \beta m_z \ \ (z \in \mathcal{D}_{\text{geom}}), \quad \text{or} \quad \gamma_z = 0 \ \ (z \in \mathcal{D}_{\text{bin}}), \qquad \beta > 0, \tag{8}$$

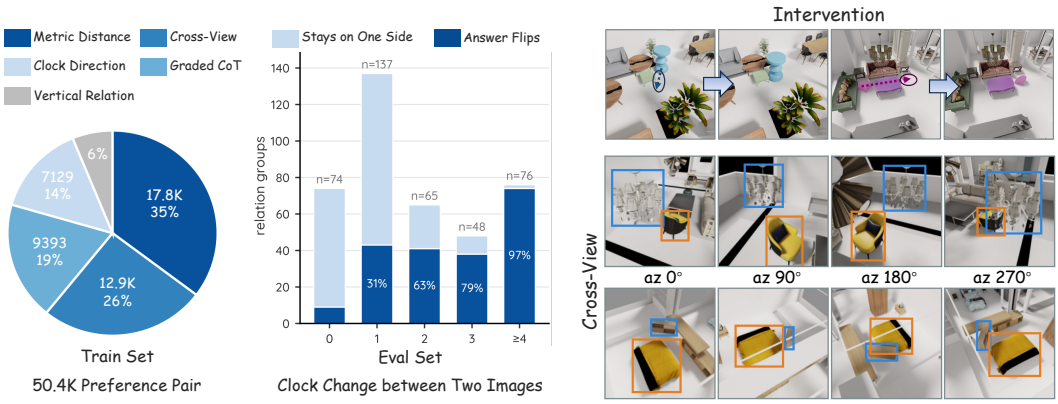


Figure 2: **One construction across training and evaluation.** Left, Gauge-50K by pair family, with geometric displacement annotations for measured pairs. Middle, Constancy-Bench groups held-out scenes by the clock change between two images and reports how often the answer flips. Right, an intervention moves one object and re-renders the scene from up to four views.

Zero offset denotes ordering-only supervision in the pair branch, not zero geometric supervision in the full objective and not zero error. Nonempty strata k index task, pair side and supervision type; fixed weights $w_k > 0$, $\sum_k w_k = 1$, define $\mathcal{D}_{\text{mix}} = \sum_k w_k \,\text{Unif}(\mathcal{D}_k)$ and

$$\mathcal{L}_{\text{pair}} = -\mathbb{E}_{z\sim\mathcal{D}_{\text{mix}}} \log \sigma(\beta h_\theta(z) - \gamma_z)\,, \tag{9}$$

At fixed β, larger measured errors demand larger gaps at a given positive loss level. Zero offsets recover the DPO form in the pair branch; the full objective retains direct and intervention terms.

Canonical prediction. For (R2), fix a finite grid $\mathcal{D}_Q \subset \mathbb{R}_{>0}$ before training independently of test truths, with $\mathcal{C} = \mathcal{C}_{12} \times \mathcal{D}_Q \subseteq \mathcal{R}_Q$ and group truths $r^\star_G = Q(r^\star_{\text{phys},G}) \in \mathcal{C}$. A unique terminated relation sequence $\psi(r)$ satisfies $\Pi(\psi(r)) = r$. Under a fixed prompt q_{rel} and $x^{\text{rel}}_v = (I_v, q_{\text{rel}})$, inference returns $\psi(\hat{r}_v)$ under a fixed tie rule, without test truth or GT rationales, where

$$s_\theta(x^{\text{rel}}_v, r) = \log \pi_\theta(\psi(r) \mid x^{\text{rel}}_v), \qquad \hat{r}_v \in \arg\max_{r\in\mathcal{C}} s_\theta(x^{\text{rel}}_v, r), \tag{10}$$

Direct supervision of the worst view. Set $\Delta(r,r^\star) = \delta_0\, \mathbb{I}[r \neq r^\star] + \xi\, m(r,r^\star)$, where $\delta_0 > 0$ is a classification margin, $\xi > 0$ scales geometry. The most violated view–candidate comparison defines:

$$\ell_{\text{dir}}(G) = \left[\max_{1\le v\le V_G,\, r\in\mathcal{C}\setminus\{r^\star_G\}} \left\{ s_\theta(x^{\text{rel}}_v, r) - s_\theta(x^{\text{rel}}_v, r^\star_G) + \Delta(r, r^\star_G) \right\} \right]_+, \tag{11}$$

$$\mathcal{L}_{\text{dir}} = \mathbb{E}_{G\sim\mathcal{D}_{\text{grp}}} \ell_{\text{dir}}(G). \tag{12}$$

Proposition 2 (Direct control of canonical group predictions)**.** *For a nonempty finite group with finite scores, $r^\star_G \in \mathcal{C}$, and $|\mathcal{C}| \ge 2$, Eq. 10 satisfies*

$$\delta_0\, \mathbb{I}[\exists v : \hat{r}_v \neq r^\star_G] + \xi \max_v m(\hat{r}_v, r^\star_G) \le \ell_{\text{dir}}(G), \tag{13}$$

Thus $\ell_{\text{dir}}(G) < \delta_0$ guarantees correctness, with a unique truth maximizer at zero loss; correct predictions may still incur loss. Also $\frac{\xi}{2} \max_{u,v} m(\hat{r}_u, \hat{r}_v) \le \ell_{\text{dir}}(G)$.

3.3 Linking Intervention Responses to Geometry

Correct rankings do not determine relative gaps across scene changes. A block $B = \{(x^{\text{rel}}_{i,v}, r_i) : i = 0,\dots,K;\ v = 1,\dots,V_B\}$ renders an original state and $K \ge 2$ interventions from $V_B \ge 2$ common cameras. Each state forms a direct group G_i, with $r_i \in \mathcal{C}$ and $m_k = m(r_0, r_k) > 0$ taking at least two distinct values. Suppressing the canonical-input superscript, define

$$H_{v,k} = \underbrace{s_\theta(x_{0,v}, r_0) - s_\theta(x_{0,v}, r_k)}_{\text{original-state answer log-odds}} + \underbrace{s_\theta(x_{k,v}, r_k) - s_\theta(x_{k,v}, r_0)}_{\text{intervened-state answer log-odds}}, \tag{14}$$

Table 1: **Spatial understanding results.** § marks models evaluated without their native additional visual cues for fair comparison. † marks constant outputs.

		MSMU			QSpat	SURDS		SpatialRGPT		3DSR	BLINK	Constancy-Bench		
Model	#P	dist.	width	height	δ_2	dist.	depth	Quan.	Qual.	acc.	acc.	Rank $\rho\uparrow$	VisC. %↑	PairAcc. %↑
Frontier models														
Claude Opus 4.8	–	2.5	1.1	2.2	76.2	11.7	14.7	41.5	55.8	66.4	70.1	.127	52.7	62.0
GPT-4o	–	2.5	0.0	5.5	47.5	2.7	0.0	41.7	74.3	55.3	56.2	.136	32.4	45.4
Gemini-2.5-Pro	–	7.5	6.7	18.7	45.5	29.5	34.3	38.1	80.9	54.2	54.1	.227	37.8	64.4
Kimi-K2.6	1T	15.0	9.0	18.7	64.4	48.5	53.1	42.0	89.3	62.8	72.8	.200	46.0	68.8
Prior open spatial VLMs														
SpaceOm	3B	15.0	5.6	3.3	50.5	3.0	9.5	24.0	47.6	49.9	47.1	−.094	84.6	9.3
SpatialLadder	3B	17.5	1.1	6.6	50.5	0.0	0.0	17.5	43.2	48.2	44.9	−.029	87.2	2.9
VST-RL	7B	30.0	30.3	50.5	51.5	7.3	4.6	39.1	61.2	57.7	50.8	.264	80.4	56.1
SpatialReasoner§	8B	25.0	4.5	9.9	49.5	19.9	23.8	25.3	69.7	56.8	42.8	.277	68.9	59.0
SpatialRGPT§	8B	25.0	2.2	4.4	55.4	2.3	1.5	25.7	60.1	53.4	45.9	.000	85.8	45.9
SD-VLM	7B	47.5	47.2	68.1	27.7	0.01	0.0	22.5	49.2	26.6	16.5	.000†	100.0†	0.0
Ours — Qwen family														
Qwen2.5-VL	7B	17.5	3.4	12.1	42.6	20.0	22.3	25.0	70.0	44.9	49.5	.115	84.5	36.6
GaugeSFT	7B	47.5	48.3	67.0	45.5	22.4	24.3	33.5	74.7	48.2	50.5	.109	81.1	54.6
GaugeVLM	7B	**62.5**	**55.1**	**71.4**	**64.4**	**30.4**	**35.7**	**43.1**	**81.6**	**56.9**	**56.7**	**.279**	**91.2**	**66.3**
$\Delta_{\text{vs SFT}}$		+15.0	+6.7	+4.4	+18.9	+8.0	+11.4	+9.6	+6.9	+8.7	+6.2	+.170	+10.1	+11.7
Ours — GLM family														
GLM-4.1V	9B	30.0	1.1	9.9	43.6	29.5	26.4	31.9	71.2	48.2	51.0	.290	62.8	65.4
GaugeSFT	9B	67.5	57.3	67.0	49.5	33.9	29.1	38.4	68.8	46.7	56.2	.235	81.8	67.3
GaugeVLM	9B	**72.5**	**65.2**	**75.8**	**58.4**	**35.2**	**31.3**	**43.5**	**78.7**	**50.8**	**58.2**	**.311**	**90.5**	**69.8**
$\Delta_{\text{vs SFT}}$		+5.0	+7.9	+8.8	+8.9	+1.3	+2.2	+5.1	+9.9	+4.1	+2.0	+.076	+8.7	+2.5
Ours — Mistral family														
Pixtral	12B	10.0	4.5	4.4	8.9	21.3	15.8	27.6	64.8	44.0	37.2	.096	85.1	34.6
GaugeSFT	12B	50.0	53.9	69.2	7.9	22.0	18.8	18.0	60.0	43.5	37.0	.000†	100.0†	32.2
GaugeVLM	12B	**52.5**	**56.2**	**75.8**	**11.9**	**28.5**	**23.7**	**28.0**	**67.0**	**48.0**	**42.6**	**.138**	99.3	**41.0**
$\Delta_{\text{vs SFT}}$		+2.5	+2.2	+6.6	+4.0	+6.5	+4.9	+10.0	+7.0	+4.5	+5.6	+.138	−0.7	+8.8

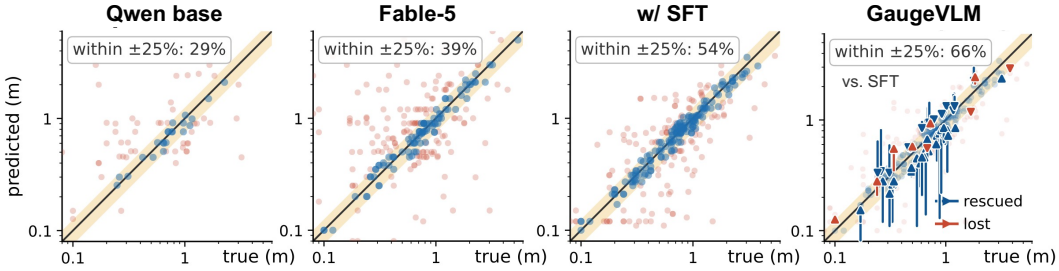


Figure 3: **Predicted versus true values across all MSMU metric tasks.** GaugeVLM achieves the best calibration even outperforming the Claude-Fable-5. Missing or unparseable numerical outputs are not plotted, so point counts vary across models. "Within ±25%" denotes $|\hat{d} - d^\star|/d^\star \leq 0.25$.

This intervention contrast cancels separable score components $a(x) + b(r)$, including input normalizers. Each direct gap includes classification margin δ_0; we fit the excess response:

$$\ell_{\text{int}}(B) = \frac{1}{V_B} \sum_v \min_{\eta \geq 0} \frac{1}{K} \sum_{k=1}^{K} \left(H_{v,k} - 2\delta_0 - \eta m_k\right)^2, \qquad \mathcal{L}_{\text{int}} = \mathbb{E}_{B \sim \mathcal{D}_{\text{int}}} \ell_{\text{int}}(B), \tag{15}$$

Each camera fits one scale across a block's interventions, allowing view-dependent confidence. This prescribed proportionality concerns the sum of two gaps without requiring equality. If all $\ell_{\text{dir}}(G_i)$ and $\ell_{\text{int}}(B)$ vanish, every canonical prediction is correct and $H_{v,k} - 2\delta_0 = \eta_v m_k$ with $\eta_v \geq 2\xi$.

Joint objective. The preference stage combines three separately averaged losses:

$$\mathcal{L}_{\text{GaugeDPO}} = \mathcal{L}_{\text{pair}} + \lambda_{\text{dir}} \mathcal{L}_{\text{dir}} + \lambda_{\text{int}} \mathcal{L}_{\text{int}}, \qquad \lambda_{\text{dir}}, \lambda_{\text{int}} > 0, \tag{16}$$

Table 2: **General Vision-Language Understanding Ability Results.** Average‡ denotes the mean of the seven percentage metrics, excluding MME.

Model	#P	MME^{P}	MME^{R}	POPE	MMStar	AI2D	SEED	SQA	HallB	MMMU	Average‡
Qwen2.5-VL	7B	1606	622	83.7	58.1	78.8	73.0	72.7	63.3	42.8	67.5
GaugeVLM	7B	**1640**	**629**	**85.9**	**60.3**	**80.1**	**75.0**	**78.9**	**64.5**	**44.4**	**69.9**
$\Delta_{\text{vs base}}$		+34	+7	+2.2	+2.2	+1.3	+2.0	+6.2	+1.2	+1.6	+2.4
GLM-4.1V	9B	1593	555	84.6	28.4	57.1	73.1	28.7	49.8	36.6	51.2
GaugeVLM	9B	**1595**	**560**	**88.9**	**32.1**	**64.3**	**74.0**	**42.7**	**52.2**	**40.7**	**56.4**
$\Delta_{\text{vs base}}$		+2	+5	+4.3	+3.7	+7.2	+0.9	+14.0	+2.4	+4.1	+5.2
Pixtral	12B	1598	411	85.4	47.9	78.1	74.5	81.9	60.0	46.0	67.7
GaugeVLM	12B	**1610**	**411**	**85.9**	**50.0**	**79.9**	**75.1**	**84.2**	**60.5**	**49.4**	**69.3**
$\Delta_{\text{vs base}}$		+12	+0.0	+0.5	+2.1	+1.8	+0.6	+2.3	+0.5	+3.4	+1.6

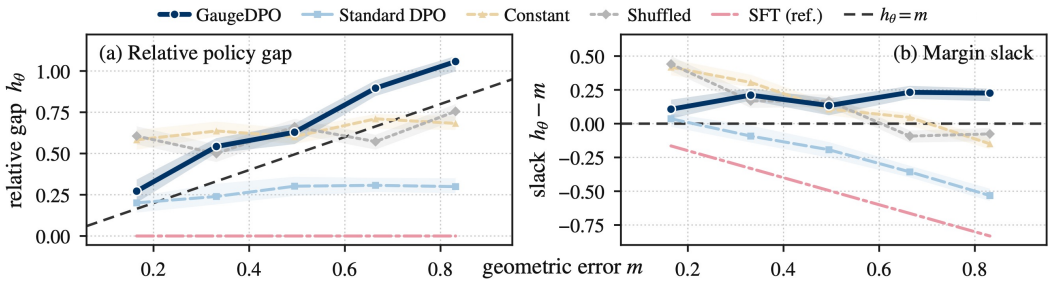


Figure 4: **Measured margins induce graded policy separation.** On held-out scenes, GaugeVLM scales its reference-relative gap with geometric error (a) and maintains positive slack (b). SFT is the reference ($h_{\text{SFT}} = 0$); bands show 95% paired scene-bootstrap CIs.

Only the pair term uses reference-relative scores. Fixed manifests define group and block distributions, with every state directly supervised. Guarantees are pointwise on evaluated groups and blocks for the specified predictor; matched experiments assess optimization and generalization.

3.4 The Gauge Engine

From scene geometry to supervision. The rendering pipeline places Objaverse-LVIS or 3D-FUTURE objects in 3D-FRONT scenes using BlenderProc (Deitke et al., 2023; Fu et al., 2021b;a; Denninger et al., 2019). Object poses determine the anchor-relative clock and centroid distance. Response-side rejections change the serialized relation; image-side pairs move the target and retain the original response. Each measured error is evaluated against its branch's truth. Vertical and reasoning-quality comparisons without a complete geometric target use the ordering-only branch. The supplied Gauge-50K pool reports 50.4K pairs across five task families and 1,174 image-side scene pairs. These are pair-pool counts, not counts of the new intervention blocks.

Matched states and cameras. For the revised objective, each block holds the anchor, object identities, scene context and camera set fixed while changing the target relation at several measured magnitudes. A fixed state supplies a camera group; all states together supply the interaction profile. Blocks require co-visibility, an identifiable anchor frame and sufficient metric-scale evidence. Base scenes and related object assets are separated before generating split-specific states and views. The block manifest records physical and serialized truths, so quantization collisions and out-of-domain cases are explicit. Full derivations are provided in App. A, with data construction details in App. B.

4 Experiments

4.1 Experimental Setup

Models and benchmarks. We instantiate GaugeVLM on Qwen2.5-VL-7B-Instruct, GLM-4.1V-9B, and Pixtral-12B (Bai et al., 2025; GLM-V Team, 2026; Agrawal et al., 2024), using Qwen as

Table 3: **Ablation of preference margin design and SFT initialization.** Zero-, constant-, shuffled- and measured-margin rows retain the same direct/profile supervision; only the pair offset changes.

	Settings	Margin m	MSMU dist.	QSpat δ_2	SRGPT Quan	3DSR acc.	BLINK acc.	Rank $\rho\uparrow$	PairAcc. %$\uparrow$
(a)	SFT Initialization	none	47.5	45.5	33.5	48.2	50.5	.109	54.6
(b)	+ Zero-offset pair loss	$m \equiv 0$	50.0	49.5	35.4	50.3	51.9	.119	56.6
(c)	+ Constant margin	$m \equiv \bar{m}$	52.5	52.5	37.2	51.6	53.1	.146	58.0
(d)	+ Shuffled margin	$m_{\sigma(i)}$	52.5	51.5	36.6	51.8	52.6	.132	57.6
(e)	+ Reward margin	$\hat{m}_\phi$ (estimated)	57.5	56.4	39.7	53.7	52.9	.142	61.0
(f)	+ GaugeDPO	measured (Eq. 4)	**62.5**	**64.4**	**43.1**	**56.9**	**56.7**	**.279**	**66.3**
(g)	GaugeDPO w/o SFT	measured (Eq. 4)	40.0	54.5	35.8	49.4	51.4	.204	52.2

Table 4: **Data efficiency of geometric magnitude supervision.** With only 25% preference data, GaugeDPO surpasses full-data DPO.

Data	Method	MSMU dist.	QSpat δ_2	SRGPT Quan.	3DSR acc.	BLINK acc.
10%	DPO	47.5	46.5	33.7	48.8	50.8
	GaugeDPO	**50.0**	**48.5**	**34.5**	**49.6**	**51.5**
25%	DPO	50.0	47.5	34.2	49.3	51.2
	GaugeDPO'	**52.5**	**51.5**	**36.1**	**51.0**	**52.6**
50%	DPO	50.0	48.5	34.9	49.8	51.6
	GaugeDPO	**57.5**	**57.4**	**39.5**	**53.4**	**54.7**
100%	DPO	50.0	49.5	35.4	50.3	51.9
	GaugeDPO	**62.5**	**64.4**	**43.1**	**56.9**	**56.7**

Table 5: **Sensitivity analysis.** λ_{dir} weights direct supervision; λ_{int} weights intervention fitting. The other weight stays at its shaded default.

Para.	Value	MSMU dist.	QSpat δ_2	SRGPT Quan.	3DSR acc.	BLINK acc.	PairAcc. %$\uparrow$
λ_{dir}	0	57.5	60.4	40.6	54.8	55.3	64.4
	0.03	60.0	63.4	42.4	56.1	56.3	65.9
	0.1	**62.5**	**64.4**	43.1	**56.9**	56.7	**66.3**
	0.3	**62.5**	**64.4**	**43.3**	56.7	**56.8**	65.9
	1.0	60.0	62.4	42.1	55.8	56.0	65.4
λ_{int}	0	60.0	63.4	42.5	56.1	56.2	60.0
	0.25	**62.5**	**64.4**	42.9	56.7	56.5	64.9
	0.5	**62.5**	**64.4**	**43.1**	56.9	**56.7**	**66.3**
	0.75	**62.5**	**64.4**	43.0	**57.0**	**56.7**	**66.3**
	1	60.0	63.4	42.6	56.6	56.3	65.4

the main backbone. We compare with SpaceOm, SpatialLadder, VST-RL, SpatialReasoner, Spatial-RGPT, SD-VLM (Remyx AI, 2025; Li et al., 2026; Yang et al., 2026; Shen et al., 2025; Cheng et al., 2024; Chen et al., 2025), and frontier models Kimi-K2.6, Claude-Opus-4.8, GPT-4o, Gemini-2.5-Pro (Moonshot AI, 2026b; Anthropic, 2026b; OpenAI, 2024; Google DeepMind, 2025). Evaluation spans benchmarks from diverse domains to assess both in-domain reasoning and cross-domain generalization, including general spatial understanding on MSMU, QSpatial$^+$ (QSpat.), SpatialRGPT-Bench, 3DSRBench, and BLINK (Chen et al., 2025; Liao et al., 2024; Cheng et al., 2024; Ma et al., 2025; Fu et al., 2024), with driving-oriented reasoning on SURDS (Guo et al., 2025).

Evaluation metrics. In Tab. 1, *Rank* is Spearman correlation between predicted and true distances; *VisC.* is prediction agreement (%) under photometric changes at fixed geometry and camera. Let $\mathcal{P}, \mathcal{G}, \mathcal{B}$ contain relocation pairs, fixed-truth camera groups and canonical multi-state blocks. Set $C_S = \prod_{j\in S} \mathbb{I}[\hat{a}_j = a_j^\star]$ for $S \in \mathcal{P} \cup \mathcal{B}$. For the tolerance-based view audit, C_G^{tol} and A_G^{tol} indicate valid groups passing correctness and agreement thresholds, respectively; overbars denote manifest means. *ViewAgree*$_{\text{tol}}$ is $100\overline{A^{\text{tol}}}_{\mathcal{G}}$, *WrongAgree*$_{\text{tol}}$ is $100\overline{A^{\text{tol}}(1-C^{\text{tol}})}_{\mathcal{G}}$, and *Mean worst* m is $\overline{\max_{i,v} m(\hat{r}_{i,v}, r_i)}_{\mathcal{B}}$. We average E_{shape} over eligible block–camera profiles, and main 7B PairAcc uses 205 pairs. With $U_v = H_v - 2\delta_0\mathbf{1}$, $m_B = (m_1, \ldots, m_K)$ and the fitted η_v^* from Eq. 15,

$$(\text{PairAcc}, \text{GroupAcc}_{\text{tol}}, \text{BlockAcc}) = 100\big(\overline{C}_{\mathcal{P}}, \overline{C^{\text{tol}}}_{\mathcal{G}}, \overline{C}_{\mathcal{B}}\big), \tag{17}$$

$$E_{\text{shape}}(B, v) = \frac{\|U_v - \eta_v^* m_B\|_2^2}{\|U_v\|_2^2}, \qquad \|U_v\|_2 > \tau_U > 0, \tag{18}$$

Training and scoring units. All backbones use spatial SFT on 25,378 QA turns followed by preference optimization, with a frozen SFT checkpoint as reference. In the SFT stage, we use the OSD dataset following SpatialRGPT together with 10K samples drawn from the MSMU training set, and fine-tune the language-model parameters with LoRA of rank 128 and $\alpha_{\text{LoRA}} = 128$, while the vision encoder is frozen. The SFT adapter is merged before the subsequent GaugeDPO stage, which uses our constructed data and performs full-parameter fine-tuning of the model and vision encoder. The recorded defaults are $\lambda_{\text{dir}} = .1$, $\lambda_{\text{int}} = .5$, $\beta = .3$, geometric $\alpha = .5$, and $\kappa = \log 5$, with $\tau_c = 1/6$ and $\tau_d = 0.25$ in normalized ring/log-distance units. Chunking reduces peak activation

Table 6: **Supervision comparison.** Worst m is the mean worst-case error. Profile metrics use eligible profiles ($|U_v| * 2 > 0.05$, finite U_v); $E * \mathrm{shape}$ and zero-$\eta_v^\star$ use their 1,022-profile intersection.

	Behavior		Profile diagnostics		
Objective	Block acc.↑	Worst $m \downarrow$	$E_{\mathrm{shape}} \downarrow$	Eligible (%)	$\eta_v^\star$=0 (%)↓
Pair + direct	60.0	.186	.241	88.2	20.9
+ Single-gap	63.7	.170	.112	91.5	9.4
+ EqSim-style	65.0	.163	.174	90.4	14.9
GaugeDPO (full)	**70.0**	**.134**	**.047**	**94.6**	**4.2**

Table 7: **Generalization.** Each setting contains 100 held-out scenes and 200 blocks; "unseen" denotes held-out magnitudes or intervention combinations.

	Pair + direct		GaugeDPO	
Setting	Block acc.↑	Worst $m \downarrow$	Block acc.↑	Worst $m \downarrow$
Familiar	68.0	.140	**76.0**	**.098**
Unseen mag.	60.0	.186	**71.0**	**.135**
Unseen combos.	52.0	.232	**63.0**	**.169**
Overall	60.0	.186	**70.0**	**.134**

memory, not candidate-scoring cost. All training is conducted on four NVIDIA A100 80GB GPUs. Full training details and settings are provided in App. C.

4.2 Main Results

Broad Gains and Competitive Performance. Across backbones, we improve all $30/30$ established spatial-task metrics over both base and SFT checkpoints (Tab. 1). The VisC exception (-0.7) accompanies recovery from SFT output collapse. QSpatial gains reach 18.9 points from preference learning versus 3.0 from imitation. Gains also extend to driving and embodied 3D benchmarks. At 7B, GaugeVLM leads prior open spatial VLMs on $9/10$ established metrics under the evaluated input protocol, trailing VST-RL by < 1 point on 3DSRBench. It leads the listed frontier models on all three MSMU tasks and exceeds GPT-4o and Gemini-2.5-Pro on all reported metrics. Fig. 3 additionally compares distance predictions with Claude-Fable-5 (Anthropic, 2026a). Additionally, across three families in Tab. 2, all metrics from eight general benchmarks (Fu et al., 2025; Li et al., 2023; Chen et al., 2024b; Kembhavi et al., 2016; Li et al., 2024; Lu et al., 2022; Guan et al., 2024; Yue et al., 2024) match or exceed their base-model values, including the hallucination metrics.

Geometric separation and cross-view correctness. On held-out scenes (Fig. 4), only the measured-margin variant shows consistently increasing mean reference-relative gaps with geometric error and positive mean slack at all plotted levels. DPO yields near-uniform mean gaps, while constant and shuffled margins weaken this pattern. At the prediction level, Tab. 8 evaluates tolerance-based cross-view agreement under the three-term objective. The pair-only model ($\lambda_{\mathrm{dir}}, \lambda_{\mathrm{int}} = 0$) achieves 50.5% accurate and 17.5% wrong agreement. Adding the direct or intervention term improves these to $60.0/12.5\%$ and $54.5/15.0\%$, respectively. With both enabled, GaugeVLM reaches 63.0% accurate and 11.5% wrong agreement, improving by 12.5 and 6.0 percentage points over the pair-only baseline.

Table 8: **Tolerance-based cross-view audit.** Ablations set the indicated weights to zero and retain the others. Agreement and correctness use separate tolerances. Invalid groups remain in all denominators. Values are percentages.

	Agreement			Accuracy		Invalid
Model	Total	Correct	Wrong	Group	Item	Invalid
SFT Init	60.0	40.0	20.0	50.0	65.0	5.0
$\lambda_{\mathrm{dir}}, \lambda_{\mathrm{int}} = 0$	68.0	50.5	17.5	60.5	71.5	4.0
$\lambda_{\mathrm{dir}} = 0$	69.5	54.5	15.0	62.5	73.5	3.5
$\lambda_{\mathrm{int}} = 0$	72.5	60.0	12.5	67.5	77.0	3.5
GaugeVLM	**74.5**	**63.0**	**11.5**	**69.5**	**79.0**	**3.0**

4.3 Ablations and Controls

Margin source and geometric assignment. In Tab. 3, measured margins lead all five task metrics and Rank ρ. With direct and interaction supervision fixed, they outperform the zero-offset control by 12.5 points on MSMU distance and 14.9 on QSpatial, raising ρ from .119 to .279. Constant, shuffled, and reward-model offsets also improve over this control but trail measured margins by up to 10.0 and 12.9 points on these tasks. Removing SFT degrades task accuracy and rank correlation, supporting spatial SFT before geometry-aware preference optimization.

Data efficiency and sensitivity analysis. Tab. 4 shows that GaugeDPO outperforms DPO on all metrics at each pair-draw budget; its 25%-budget scores already exceed the full-budget DPO baseline. Tab. 5 varies λ_{dir} and λ_{int} individually while fixing the other at its default; setting either coef-

ficient to zero ablates the corresponding loss term. The results show a broad intermediate plateau, with task accuracy more sensitive to the direct-supervision weight.

Intervention supervision and generalization. Tabs. 6 and 7 compare matched supervision objectives and held-out test settings. Adding the interaction profile raises BlockAcc from 60.0% to 70.0% and lowers mean worst m from .186 to .134, while improving profile fit over the single-gap and symmetry controls. The BlockAcc gain remains 11.0 points on both unseen magnitudes and unseen magnitude–camera combinations. App. D provides additional transfer evaluations, sensitivity sweeps, and detailed camera outcomes and controls.

5 Conclusion

We presented GaugeVLM, which learns spatial relations from measured geometric errors. GaugeDPO combines geometric preference offsets, direct cross-view supervision, and intervention profiles linking answer-odds changes to geometric magnitude. Across three backbone families, GaugeVLM improves over SFT on established spatial benchmarks. Matched ablations support geometric margin assignment and intervention supervision, with gains extending to unseen magnitudes and magnitude–camera combinations. These results demonstrate the value of geometric supervision for spatial accuracy and responses to object interventions.

AI Use Statement

We used Claude-Fable-5 and GPT-6-Astra to assist with language editing, literature retrieval and discovery, and the generation of synthetic datasets. The synthetic data were produced through the explicit 3D scene engine described in this paper, with scene construction, geometric annotations, filtering, validation, and all experimental procedures reviewed and verified by the authors. The authors take full responsibility for the final manuscript, datasets, analyses, and the accuracy of all statements and claims.

Reproducibility Statement

The supplementary material details the theoretical assumptions and proofs, data construction and filtering procedures, and training and evaluation protocols. It specifies the hyperparameters, canonical candidate-scoring rules, parsing conventions, and aggregation procedures for pair-, group-, and block-level metrics. Matched controls and repeated training runs assess the contributions and robustness of the proposed objectives.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, et al. Pixtral 12B. *arXiv:2410.07073*, 2024.

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. In *ACL Findings*, 2024.

Anthropic. Claude fable 5 and claude mythos 5, 2026a.

Anthropic. Introducing Claude Opus 4.8, May 2026b. Official model announcement, May 28, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL technical report. *arXiv:2502.13923*, 2025.

S. Divakar Bhat and Toshihiko Yamasaki. Consistent yet wrong: Evidence insensitivity in spatial vision-language models. *arXiv:2606.02742*, 2026.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024b.

Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. SD-VLM: Spatial measuring and understanding with depth-encoded vision-language models. In *NeurIPS*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024c.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *CVPR*, 2023.

Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. BlenderProc. *arXiv:1911.01911*, 2019.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS*, 2025.

Huan Fu, Bowen Cai, Lin Gao, Ling-xiao Zhang, Jiaming Wang, Cao Li, Xunqi Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3D furnished rooms with layouts and semantics. In *ICCV*, 2021a.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3D-FUTURE: 3D furniture shape with texture. *IJCV*, 2021b.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *AISTATS*, 2024.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. *arXiv:2507.01006*, 2026.

Google DeepMind. Gemini 2.5 pro model card, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Dujun Nie, Wenke Huang, Chenming Zhang, Shuai Liu, Hao Zhao, and Long Chen. SURDS: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models. In *NeurIPS*, 2025.

Zhi Helu, Jingjing Huang, Xu Wang, Yangbin Xu, Wanyue Zhang, Baoyang Jiang, Shirui Deng, Liang Zhu, Fangfang Li, Tiejun Zhao, Yankai Lin, and Yuan Yao. Scaling spatial reasoning in MLLMs through programmatic data synthesis. *arXiv:2512.16237*, 2025.

Ting Huang, Zhenyu Zhang, Wenyuan Huang, Jian Yang, and Hao Tang. ConsiSpace: Learning geometric consistency matters for video spatial reasoning. In *ECCV*, 2026.

Siwen Jiao, Tianxiong Lv, Kangan Qian, Chenxu Zhao, Xiuyuan Zhu, Tianlun Li, Xiaolong Cheng, Jinyu Li, Zhihao Liao, and Yang Cai. Smooth operator: Smooth verifiable reward activates spatial reasoning ability of vision-language model. *arXiv:2601.07695*, 2026.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *CVPR*, 2024.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv:2505.21500*, 2025.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive training for spatial reasoning in vision-language models. In *ICLR*, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

Yuan-Hong Liao. Q-Spatial Bench: Official evaluation code. `https://github.com/andrewliao11/Q-Spatial-Bench-code`, 2024.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. In *EMNLP*, 2024.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

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 *ICCV*, 2025.

Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic AI for spatial reasoning with a dynamic API. *arXiv:2502.06787*, 2025.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *NeurIPS*, 2024.

Moonshot AI. Kimi K2.6 model card. `https://huggingface.co/moonshotai/Kimi-K2.6`, 2026a.

Moonshot AI. Kimi K2.6: Advancing open-source coding, April 2026b. Official technical blog, released April 20, 2026.

Michael Ogezi and Freda Shi. SpaRE: Enhancing spatial reasoning in vision-language models with synthetic data. In *ACL*, 2025.

OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2023.

OpenAI. GPT-4o system card, August 2024. System card, August 8, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Remyx AI. SpaceOm. `https://huggingface.co/remyxai/SpaceOm`, 2025. Hugging Face model.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in VLMs. In *NeurIPS*, 2025.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In *CVPR*, 2025.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 2005.

Shanmukha Vellamcheti, Uday Kiran Kothapalli, Disharee Bhowmick, and Sathyanarayanan N. Aakur. CVT-Bench: Counterfactual viewpoint transformations reveal unstable spatial representations in multimodal LLMs. *arXiv:2603.21114*, 2026.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *EMNLP*, 2024.

Peiyao Wang and Haibin Ling. SVQA-R1: Reinforcing spatial reasoning in MLLMs via view-consistent reward optimization. In *ICLR Workshop*, 2026.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Fei-Fei Li, and Manling Li. MindCube: Spatial mental modeling from limited views. In *ICLR*, 2026.

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 *ICCV*, 2023.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. AlphaDPO: Adaptive reward margin for direct preference optimization. In *ICML*, 2025.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. In *ECCV*, 2026.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in MLLMs. In *AAAI*, 2026.

Leyuan Yu, Xiao Tang, Minghao Liu, Xinyuan Li, Xiaokai Bai, Sheng Zhou, Qunshu Lin, Weihao Xuan, and Naoto Yokoya. MindEdit-Bench: Benchmarking object-level counterfactual spatial reasoning in VLMs from in-the-wild photos. *arXiv:2607.00491*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3D. *arXiv:2503.22976*, 2025.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. RoboRefer: Towards spatial referring with reasoning in vision-language models for robotics. In *NeurIPS*, 2025.

APPENDIX OVERVIEW

This appendix develops the theoretical foundations of measured geometric supervision and provides details of data construction, training, evaluation, and supplementary experiments.

Contents

A DERIVATIONS

Throughout, $\sigma(t) = (1 + e^{-t})^{-1}$ denotes the logistic function, and expectations are over $(x, y_w, y_l) \sim \mathcal{D}$ unless stated otherwise.

A.1 RELATION METRICS AND CONSISTENCY CERTIFICATES

The same geometric metric defines preference margins and quantifies disagreement across views. We establish its invariance and metric properties, then derive error certificates from the triangle inequality.

Representation and quantization. Clock direction is defined relative to the anchor's canonical front, with the dominant wall normal as a fallback. Both reference frames are independent of the camera. Questions identify the anchor and specify the reference direction through visual evidence or an explicit prompt convention. Metric-distance questions additionally require scale information in the input.

The clock–distance metric is defined on $\mathcal{R} = \mathcal{C}_{12} \times \mathbb{R}_+$, and response templates represent a subset $\mathcal{R}_Q$. The quantizer Q specifies the clock bins and distance precision; on $\mathcal{R}_Q$, the parser exactly inverts the verbalizer. Rejected labels are retained only when they remain distinct from the truth after serialization. Let $e_Q(r) = m(Qr, r)$ for relations in the same clock–distance space. By the triangle inequality,

$$|m(Qr, Qr^\star) - m(r, r^\star)| \le e_Q(r) + e_Q(r^\star).$$

Thus, margins computed from serialized labels approximate the corresponding geometric errors up to the quantization error of the two relations. This bound concerns the discrete clock representation and distance precision. Extending it to continuous azimuth requires a separate angular quantization term.

Invariant discrepancies and metric construction.

Proposition 3 (Invariant discrepancies)**.** *For nonnegative discrepancies D_c, D_d, the clock invariance in Eq. 2 holds iff $D_c(c, c^\star) = \phi(c - c^\star)$ for a unique $\phi : \mathcal{C}_{12} \to \mathbb{R}_{\ge 0}$, and the distance invariance in Eq. 2 holds iff $D_d(d, d^\star) = \psi(d/d^\star)$ for a unique $\psi : \mathbb{R}_{>0} \to \mathbb{R}_{\ge 0}$.*

Proof of Prop. 3. Sufficiency is immediate: $(c + k) - (c^\star + k) = c - c^\star$ in $\mathcal{C}_{12}$ and $(\lambda d)/(\lambda d^\star) = d/d^\star$, so any ϕ, ψ of those arguments give invariant discrepancies. For necessity, take $k = -c^\star$ in $D_c(c + k, c^\star + k) = D_c(c, c^\star)$ to get $D_c(c, c^\star) = D_c\big((c - c^\star) \bmod 12,\, 0\big)$, so $\phi(u) := D_c(u, 0)$ represents it through the cyclic difference alone; take $\lambda = 1/d^\star$ in $D_d(\lambda d, \lambda d^\star) = D_d(d, d^\star)$ to get $D_d(d, d^\star) = D_d(d/d^\star, 1)$, so $\psi(t) := D_d(t, 1)$ represents it through the ratio alone. Since $d/d^\star = \exp(\log d - \log d^\star)$, a function of the ratio is equivalently a function of the log-ratio $s = \log d - \log d^\star$. □

Prop. 3 determines the ratio form of the distance discrepancy; we choose the absolute log-ratio for symmetry and the triangle inequality.

Proposition 4 (Log distance and relative error)**.** *On $\mathbb{R}_+$, $L(d, d') = |\log d - \log d'|$ is a metric. The (uncapped) relative error $R(d, d^*) = |d - d^*|/d^*$ satisfies, with $s = \log d - \log d^*$,*

$$|s|\, e^{-|s|} \;\le\; R(d, d^*) \;=\; |e^s - 1| \;\le\; |s|\, e^{|s|}, \tag{19}$$

so $R = |s| + O(s^2)$ agrees with the log metric to first order.

Proof. L is the pullback of the absolute-value metric under the bijection $\log\colon \mathbb{R}_+ \to \mathbb{R}$, hence a metric. For Eq. 19, apply the mean value theorem to $e^s - 1$. For $s > 0$, $e^s - 1 = s\, e^\xi$ with $e^\xi \in (1, e^s)$ gives $s \le e^s - 1 \le s\, e^s$. For $s < 0$, $1 - e^s = |s|\, e^\xi$ with $e^\xi \in (e^s, 1)$ gives $|s|\, e^{-|s|} \le 1 - e^s \le |s|$. □

Lemma 1 (Capping preserves the triangle inequality)**.** *If d is a metric on a set $\mathcal{X}$ and $\kappa > 0$, then $d_\kappa = \min(d, \kappa)$ is a metric on $\mathcal{X}$.*

Proof. Symmetry and positive definiteness follow directly from those of d. For the triangle inequality, first consider $d(a,b) \geq \kappa$ or $d(b,c) \geq \kappa$. Then

$$d_\kappa(a,b) + d_\kappa(b,c) \geq \kappa \geq d_\kappa(a,c).$$

Otherwise both terms are below the cap, and

$$d_\kappa(a,b) + d_\kappa(b,c) = d(a,b) + d(b,c) \geq d(a,c) \geq d_\kappa(a,c).$$

□

Proposition 5 (The margin is a metric on $\mathcal{R}$)**.** *Let $\alpha \in [0,1]$, $\alpha_c = \alpha$, $\alpha_d = 1-\alpha$, and $\kappa > 0$. The function m of Eq. 4 is a pseudometric on $\mathcal{R} = \mathcal{C}_{12} \times \mathbb{R}_+$, and a metric whenever $\alpha_c > 0$ and $\alpha_d > 0$. It is symmetric, satisfies the triangle inequality, is invariant under Eq. 2, and takes values in $[0,1]$ when $\alpha_c + \alpha_d = 1$.*

Proof. On $\mathcal{C}_{12}$, $\delta(a,b) = \min_k |a - b + 12k|$ is a metric, the triangle inequality following from $\delta(a,c) \leq |a - c + 12(k_1+k_2)|$ for k_1, k_2 attaining $\delta(a,b)$ and $\delta(b,c)$. Its maximum is 6, so $d_{\text{ring}} = \delta/6$ takes values in $[0,1]$. On $\mathbb{R}_+$, L is a metric by Prop. 4, so $\min(L,\kappa)$ is one by Lemma 1 and $d_{\text{rel}} = \min(L,\kappa)/\kappa$ is its positive rescaling. A nonnegative combination of metrics on the two factors is a pseudometric on the product, symmetry and the triangle inequality being inherited coordinatewise. It separates points exactly when both weights are positive, $m(r,r') = 0$ then forcing $c = c'$ and $\log d = \log d'$. Invariance holds because d_{ring} depends only on $(c - c^\star) \bmod 12$ and d_{rel} only on $d/d^\star$. Finally d_{rel} reaches 1 once $|\log(d/d^\star)| \geq \kappa$, so $\alpha_c + \alpha_d = 1$ gives $m \in [0,1]$, with $m = 1$, for two positive normalized weights, only when the direction is opposite and the maximum of the distance ratio and its reciprocal is at least e^κ. □

The metric serves two complementary roles: $m(r^-, r^\star)$ assigns a preference margin from a rejected answer's geometric error, while $m(a_u, a_v)$ measures disagreement between predictions from two views. The latter supports the error certificate below through symmetry and the triangle inequality. For $\kappa = \log 5$, a distance scaled by λ contributes $\alpha_d \min(|\log \lambda|, \kappa)/\kappa$. The scale ranges $[1.3, 1.6]$, $[1.8, 2.5]$, and $[3, 5]$ therefore correspond to 0.16–0.30, 0.37–0.57, and 0.68–1.00 of the maximum distance contribution, respectively. The tolerances τ_c and τ_d in Eq. 30 use these normalized direction and log-distance units.

Disagreement as an error certificate. For views sharing one relation, the triangle inequality converts prediction spread into a lower bound on the largest error. The same argument extends to different truths when a known isometry aligns them.

Proof of Prop. 1. Write a^* for the true answer, so that $m_u = m(a_u, a^*)$ and $m_v = m(a_v, a^*)$. By Prop. 5, m is symmetric and satisfies the triangle inequality on $\mathcal{R}$, hence

$$m(a_u, a_v) \leq m(a_u, a^*) + m(a^*, a_v) = m_u + m_v \leq 2\max(m_u, m_v) \leq 2 \max_{v \in V_{\text{ok}}} m_v.$$

Taking the maximum and dividing by two proves Eq. 6. Sharpness follows by taking predictions one clock step on either side of the truth with equal distances when $\alpha > 0$; when $\alpha = 0$, use two distances symmetric about the truth in log coordinates below the cap. In either case the spread is exactly twice the largest error. □

Remark 1 (Certificate under truncation). The distance component saturates when $|\log(d/d^\star)| \geq \kappa$. Equation 6 remains a valid lower bound in the capped metric, with reduced resolution for larger distance errors.

Remark 2 (Direction of the implication). Equation 6 turns prediction disagreement into a lower bound on error. Because identical predictions can share the same error, we complement this certificate with direct supervision in Eq. 11 and evaluate correctness together with agreement.

Corollary 1 (Sharp distance certificates)**.** *Let d_u, d_v be positive predictions with a shared positive truth $d^\star$. Define $\rho = \max(d_u/d_v, d_v/d_u)$. Then*

$$\max_{w \in \{u,v\}} |\log(d_w/d^\star)| \geq \tfrac{1}{2}\log\rho, \qquad \max_{w \in \{u,v\}} \frac{|d_w - d^\star|}{d^\star} \geq \frac{\rho - 1}{\rho + 1}.$$

Both constants are sharp, with equality attained at the geometric and arithmetic means of the predictions, respectively.

Proof. The first inequality follows from $\log \rho \le |\log(d_u/d^\star)| + |\log(d_v/d^\star)|$. For the second, write $a = \min(d_u, d_v)$ and $b = \max(d_u, d_v)$. If both relative errors are at most $r < 1$, then $b \le (1+r)d^\star$ and $a \ge (1-r)d^\star$, whence $\rho \le (1+r)/(1-r)$ and $r \ge (\rho-1)/(\rho+1)$. The bound is automatic for $r \ge 1$. At $d^\star = (a+b)/2$, both errors equal $(b-a)/(b+a)$. The geometric mean similarly attains the log bound. □

For relation-changing interventions, the same argument applies when a known isometry aligns the two ground-truth relations, as formalized below.

Proposition 6 (Transformation-aligned certificate)**.** *Let $(\mathcal{A}, d_\mathcal{A})$ be a metric space, and let $T : \mathcal{A} \to \mathcal{A}$ be a known isometry satisfying $a_B^* = T a_A^*$. For predictions $\hat{a}_A, \hat{a}_B \in \mathcal{A}$, set $e_A = d_\mathcal{A}(\hat{a}_A, a_A^*)$ and $e_B = d_\mathcal{A}(\hat{a}_B, a_B^*)$. Then*

$$\max(e_A, e_B) \ge \tfrac{1}{2} d_\mathcal{A}(T\hat{a}_A, \hat{a}_B). \tag{20}$$

Proof. By the triangle inequality and the isometry property,

$$d_\mathcal{A}(T\hat{a}_A, \hat{a}_B) \le d_\mathcal{A}(T\hat{a}_A, T a_A^*) + d_\mathcal{A}(a_B^*, \hat{a}_B) = e_A + e_B \le 2\max(e_A, e_B).$$

□

For a fixed allocentric relation across cameras, T is the identity. For binary left/right pairs whose ground truth reverses, T swaps the labels. Common cyclic shifts and positive distance rescalings are also isometries of the clock–distance metric. Applying the certificate to an object intervention requires a known transformation of this kind, established from the scene geometry. We evaluate intervention correctness with PairAcc., which requires both answers to match their respective truths.

A.2 Measured Preference Objective and Mixed Supervision

The measured preference term combines a graded comparison model with reference-relative policy scores. We derive its response-side likelihood and gradient, then specify how measured and ordinal-only comparisons enter the mixed objective.

Graded preferences as a Gumbel comparison. Give each response a Gumbel-perturbed reward $\tilde{r}(x, y) = r(x, y) + \epsilon_y$, with $\epsilon_{y_w}, \epsilon_{y_l}$ i.i.d. standard Gumbel. Their difference is standard logistic, $\Pr[\epsilon_l - \epsilon_w \le t] = \sigma(t)$. Conditioning on $\epsilon_w = u$ gives $\int f(u)F(u+t)\,du$, which $z = e^{-u}$ turns into $\int_0^\infty \exp(-z(1+e^{-t}))\,dz = \sigma(t)$.

Lemma 2 (Probability of clearing a gap)**.** *For any $M \in \mathbb{R}$, $\Pr\big[\tilde{r}(x, y_w) > \tilde{r}(x, y_l) + M\big] = \sigma\big(r(x, y_w) - r(x, y_l) - M\big)$. At $M = 0$ this is the Bradley-Terry model.*

Proof. $\Pr[\tilde{r}(x, y_w) > \tilde{r}(x, y_l) + M] = \Pr[\epsilon_l - \epsilon_w < r(x, y_w) - r(x, y_l) - M]$, and the logistic law applies at $t = r(x, y_w) - r(x, y_l) - M$. □

Setting $M = \beta\, m(y_l)$, with β converting the unit-free margin into reward units, gives the chosen graded-comparison model of Sec. 3.2, whose negative log-likelihood gives the measured response-side branch of Eq. 9; Amini et al. (2024) give the same interpretation of an offset preference model. Using $M = \beta m$ expresses the margin in reward units and avoids an additional offset-scale hyperparameter once α and κ are fixed.

From regularized rewards to policy scores. For a reward r and positive reference π_{ref}, assume $0 < Z(x) < \infty$ and consider

$$\max_\pi \; \mathbb{E}_{y\sim\pi(\cdot|x)}\big[r(x, y)\big] - \beta\,\mathrm{KL}\big(\pi(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\big). \tag{21}$$

Define $\pi_r(y \mid x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y \mid x)\exp\big(r(x, y)/\beta\big)$ with $Z(x) = \sum_y \pi_{\text{ref}}(y \mid x)\exp(r(x, y)/\beta)$. Eq. 21 rewrites as

$$\begin{aligned}\mathbb{E}_{y\sim\pi}\Big[r(x, y) - \beta \log \tfrac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\Big] &= -\beta\,\mathbb{E}_{y\sim\pi}\Big[\log \tfrac{\pi(y|x)}{\pi_r(y|x)}\Big] + \beta \log Z(x) \\ &= -\beta\,\mathrm{KL}\big(\pi \,\|\, \pi_r\big) + \beta \log Z(x),\end{aligned}$$

which is maximized at $\pi = \pi_r$. Inverting the definition of π_r,

$$r(x, y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x). \tag{22}$$

For two responses to the same prompt, the $\beta \log Z(x)$ terms cancel:

$$r(x, y_w) - r(x, y_l) = \beta \Big[\log \tfrac{\pi_r(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \tfrac{\pi_r(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \Big] = \beta\, h(x, y_w, y_l). \tag{23}$$

Substituting Eq. 23 into Lemma 2 with $M = \beta m(y_l)$, and parameterizing the optimal policy as π_θ,

$$p_m(y_w \succ y_l \mid x) = \sigma\big(\beta\, h_\theta(x, y_w, y_l) - \beta\, m(y_l)\big),$$

whose negative log-likelihood over measured response pairs is the corresponding branch of Eq. 9:

$$\mathcal{L}_{\text{mag}} = -\,\mathbb{E}\Big[\log \sigma\Big(\beta\big(h_\theta(x, y_w, y_l) - m(y_l)\big)\Big)\Big],$$

Corollary 2 (Degenerate margins)**.** *(i) If $m \equiv 0$, then $\mathcal{L}_{\text{mag}} = \mathcal{L}_{\text{DPO}}$ identically. (ii) If $m \equiv c$ for a constant $c > 0$, then $\mathcal{L}_{\text{mag}}$ is the DPO loss with a fixed target margin $\gamma = \beta c$ inside the Bradley-Terry model, a fixed-offset comparison. SimPO (Meng et al., 2024) also uses a constant offset; the present loss retains the reference model and summed sequence scores.*

Both follow by substitution into this response-side branch, with $\sigma(\beta(h_\theta - c)) = \sigma(\beta h_\theta - \gamma)$ for (ii).

Remark 3 (Definition of a group-preserving shuffled-margin control). Within each measured task–side–supervision stratum, consider groups of equal size, each with a common margin across its views. A group-preserving control permutes margins between these groups and broadcasts each assigned margin to all member pairs. This preserves the margin multiset and its pair-weighted distribution within each stratum. Only the pair offsets in Eq. 9 change; the geometric quantities in the direct and interaction terms remain fixed.

Gradient of the magnitude term. Write the per-sample loss as $\ell(\theta) = -\log \sigma(z_\theta)$ with $z_\theta = \beta\big(h_\theta(x, y_w, y_l) - m(y_l)\big)$. Using $\frac{d}{dz} \log \sigma(z) = 1 - \sigma(z) = \sigma(-z)$,

$$\nabla_\theta \ell = -\,\sigma(-z_\theta)\, \nabla_\theta z_\theta = -\,\beta\, \sigma\big(\beta\,(m(y_l) - h_\theta)\big)\, \nabla_\theta h_\theta.$$

Since π_{ref} is constant in θ,

$$\nabla_\theta h_\theta = \nabla_\theta \log \pi_\theta(y_w \mid x) - \nabla_\theta \log \pi_\theta(y_l \mid x),$$

which together give the response-side gradient of Eq. 9. At $m = 0$ the weight $\sigma(\beta(m - h_\theta))$ equals $\sigma(\rho_\theta(x, y_l) - \rho_\theta(x, y_w))$, recovering DPO's gradient weight exactly. For $m > 0$, the gradient weight equals $1/2$ at $h_\theta = m$. At a fixed gap, larger geometric errors produce stronger scalar gradients; the full parameter gradient also depends on $\nabla_\theta h_\theta$.

At the reference initialization, $h = 0$, $\beta = 0.3$, and $m \in [0, 1]$, the sigmoid weight ranges from 0.5 to $\sigma(0.3) \simeq 0.5744$. The local sensitivity of the scalar gradient magnitude to m is at most $\beta^2/4 = 0.0225$.

Mixed supervision and conditional comparisons. Each task specifies its supervised coordinates J_z before parsing. Measured pairs contain an exactly correct chosen label and a valid rejected label with positive m_{J_z}, evaluated against the ground truth for each input. Clock-only and distance-only pairs retain the coordinate weights in Eq. 5. Valid ordering-only preferences, including vertical comparisons and partially correct chosen responses, enter $\mathcal{D}_{\text{bin}}$ with zero offset. Invalid or unverifiable pairs are excluded. Width and height are evaluated as transfer tasks.

Sampling from $\mathcal{D}_{\text{mix}}$ implements the fixed task–side–supervision weights w_k. Both response- and image-side comparisons score the full response. For image-side pairs, the same response is evaluated against each image's ground truth. The normalization term $C(x) = \beta \log Z(x)$ in Eq. 22 cancels for comparisons within one input but leaves $C(x_w) - C(x_l)$ across inputs; the image-side loss is therefore interpreted as conditional discrimination.

A.3 Identifiability and Interpretation of Measured Margins

Measured margins encode error magnitude in two ways: they shift the separation required for a given loss and increase the scalar gradient at a fixed separation. We derive these effects and characterize the resulting allocation under a fixed separation budget.

Throughout, (y_w, y_i) denotes a response-side pair for prompt x, with positive reference probabilities for both responses. We analyze the reference-relative separations h_i as free real scalars.

Loss-level thresholds and local gradient pressure. The geometric measurement m_z is supplied by the displacement construction. The linear offset $\gamma_z = \beta m_z$ is an explicit choice within offset preference optimization (Amini et al., 2024). Its scalar loss analysis applies independently of the pair side.

For fixed $\beta > 0$, write the per-pair loss as

$$\ell(h; m) = \log(1 + \exp(\beta(m - h))) .$$

For any $\varepsilon > 0$, monotonicity of the exponential and logarithm gives

$$\ell(h; m) \le \varepsilon \quad \Longleftrightarrow \quad h \ge m - \frac{1}{\beta} \log(e^{\varepsilon} - 1) . \tag{24}$$

Indeed, exponentiating the loss inequality gives $\exp(\beta(m - h)) \le e^{\varepsilon} - 1$; taking logarithms and rearranging proves the equivalence. Thus increasing m by t shifts the gap threshold for any fixed positive loss tolerance by t. Equivalently, $\ell(h + t; m + t) = \ell(h; m)$.

The scalar derivatives satisfy

$$\frac{\partial \ell}{\partial h} = -\beta \sigma(\beta(m - h)), \tag{25}$$

$$\frac{\partial}{\partial m} \left| \frac{\partial \ell}{\partial h} \right| = \beta^2 \sigma(\beta(m - h)) \big(1 - \sigma(\beta(m - h))\big) > 0. \tag{26}$$

At fixed h and β, larger margins increase the magnitude of the scalar gradient. This describes the direct effect of the margin on the loss; parameter gradients additionally depend on the model Jacobian.

At $h = m$, the logit is zero, the loss is $\log 2$, and its derivative is $-\beta/2$. The loss continues to decrease for larger h, so the margin specifies a separation threshold rather than a regression target. The attained gaps depend on shared model parameters and the joint objective.

Identification and constrained allocation. Geometric measurements supply information absent from deterministic preference labels: different errors can share the same preference ordering. The following results contrast magnitude information in stochastic labels with its explicit introduction through measured margins.

Lemma 3 (Identification under a stochastic preference model)**.** *Suppose labels follow a Bradley–Terry model with target reward $\rho_g(x, y) = -\beta m(y) + C(x)$, where $m(y_w) = 0$ and $m(y_i) = m_i$. The probability of preferring the truth is then $p_i = \sigma(\beta m_i)$. The unconstrained population optimum of the unshifted binary preference loss is $h_i = m_i$ for every i.*

Proof. The population loss on pair i is the cross-entropy

$$\ell_i(h_i) = -\, p_i \log \sigma(\beta h_i) - (1 - p_i) \log \sigma(-\beta h_i).$$

Differentiating with $\frac{d}{dz} \log \sigma(z) = \sigma(-z)$ and using $\sigma(-z) = 1 - \sigma(z)$,

$$\frac{d\ell_i}{dh_i} = -\,\beta \left[p_i\, \sigma(-\beta h_i) - (1 - p_i)\, \sigma(\beta h_i) \right] = -\,\beta \left[p_i - \sigma(\beta h_i) \right],$$

which vanishes iff $\sigma(\beta h_i) = p_i = \sigma(\beta m_i)$, that is, $h_i = m_i$. The second derivative $\beta^2 \sigma(\beta h_i)\, \sigma(-\beta h_i)$ is positive, so the minimizer is unique and interior. Since $0 < p_i < 1$, the minimum is attained at a finite separation. □

Lemma 4 (Deterministic labels are uninformative about magnitude)**.** *When every chosen label has probability $p_i = 1$ and the pairs have equal weight, the unshifted preference likelihood is functionally independent of the margins, no finite unconstrained minimizer exists, and the optimum under the budget $\sum_i h_i = B$ is $h_i^\star = B/n$ for every i.*

Proof. With $p_i = 1$ the loss on pair i is $-\log \sigma(\beta h_i)$, in which m_i does not appear, so no estimator based on this likelihood distinguishes two margin assignments. Each term is strictly decreasing in h_i, so its infimum is approached as $h_i \to +\infty$. Under $\sum_i h_i = B$, strict concavity of $\sum_i \log \sigma(\beta h_i)$ (since $\frac{d^2}{dz^2} \log \sigma(z) = -\sigma(z)\sigma(-z) < 0$) and the affine constraint give Lagrangian stationarity $\beta\, \sigma(-\beta h_i) = \lambda$ for all i, hence h_i constant in i, hence $h_i^\star = B/n$. □

Proposition 7 (Allocation under an explicit separation budget)**.** *For fixed m_i, equal pair weights, free real separations and the constraint $\sum_{i=1}^n h_i = B$, the shifted loss has the unique minimizer $h_i = m_i + (B - \sum_j m_j)/n$.*

Proof. Let $z_i = h_i - m_i$. The constraint becomes $\sum_i z_i = B - \sum_i m_i$ and $f(z) = -\log \sigma(\beta z)$ is strictly convex. Jensen's inequality shows that $n^{-1} \sum_i f(z_i)$ is uniquely minimized when all z_i equal their fixed mean. This gives the stated solution directly. □

Remark 4 (Scope of constrained allocation). With positive unequal weights w_i, budget stationarity becomes $w_i \beta \sigma(\beta[m_i - h_i]) = \nu$. Within the equal-weight, free-separation, linear-budget family, offsets m_i' reproduce the same optimal pairwise gap differences iff $m_i' = m_i + c$ for one promptwise constant. This result characterizes allocation under the stated budget. We evaluate the ordering of learned gaps in the trained model in Fig. 4.

Design choice and interpretation of learned gaps. A nonnegative increasing mapping s with $s(0) = 0$ could instead supply $\gamma_z = \beta s(m_z)$. Our linear choice preserves increments in the declared geometric-error scale without adding a nonlinear shape parameter; the metric m remains unchanged in the geometric certificates and direct prediction bounds. Fig. 4 evaluates learned mean reference-relative gaps and slack $h - m$ across geometric-error levels.

A.4 Canonical Prediction and Cross-View Supervision

The direct group objective promotes correct predictions of the same relation across cameras. We prove its prediction bound and extend it to quantization and approximate inference.

Canonical prediction bound. A group shares the scene configuration, anchor frame, and serialized relation while varying the camera. Its construction specifies visibility, orientation, and scale requirements before evaluation. Relation-changing interventions define separate groups. Canonical inference scores the terminated relation sequence $\psi(r)$ under the fixed prompt q_{rel}; pairwise training retains the multi-step response format. Scores sum answer tokens through termination, excluding prompt and padding. The candidate domain, quantizer, verbalizer, and tie rule are fixed before training. The bound below applies to this canonical predictor; unrestricted generation is evaluated using the native benchmark protocol.

Fix a group with a shared truth, a finite candidate space of at least two elements containing that truth, finite scores, and exact score-based prediction. If every prediction is correct, the left side of Eq. 13 is zero. Otherwise choose an erroneous prediction with maximal geometric error. Its score is at least the truth score. Substitution into the maximization gives the prediction-failure and geometric-error terms together. This argument also holds for the endpoint pseudometrics: the positive classification term covers erroneous labels that have zero pseudometric distance. The triangle inequality bounds every pairwise disagreement by twice the largest truth-relative error.

Zero direct loss implies $s_\theta(x_v^{\mathrm{rel}}, r_G^\star) - s_\theta(x_v^{\mathrm{rel}}, r) \geq \delta_0 + \xi m(r, r_G^\star) > 0$ for all wrong candidates in each view. Thus, zero direct loss guarantees a unique correct prediction in every view. The bound remains informative at nonzero loss, while a correct argmax can still incur a margin violation.

Quantization and approximate inference. The guarantee concerns $r^\star_G = Q(r^\star_{\mathrm{phys},G})$. Whenever both relations are in the domain of m,

$$\max_v m(\hat{r}_v, r^\star_{\mathrm{phys},G}) \le \frac{\ell_{\mathrm{dir}}(G)}{\xi} + m(Q(r^\star_{\mathrm{phys},G}), r^\star_{\mathrm{phys},G}).$$

The additional term quantifies discretization error in the clock–distance metric. The fixed grid, quantizer, serialization and out-of-domain accounting are specified in App. C.1. Correctness here is exact equality with the serialized label; the tolerance-based camera audit uses a separate event (App. C.3).

To account for approximate inference, let $\widetilde{\ell}_G$ be the loss computed over the searched candidates. Assume a certified nonnegative search error satisfying $\ell_{\mathrm{dir}}(G) \le \widetilde{\ell}_G + \varepsilon_{\mathrm{search}}$. Let the actual predictions $\widetilde{r}_v \in \mathcal{C}$ satisfy

$$s_\theta(x^{\mathrm{rel}}_v, \widetilde{r}_v) \ge \max_{r\in\mathcal{C}} s_\theta(x^{\mathrm{rel}}_v, r) - \varepsilon_v, \qquad \varepsilon_{\mathrm{dec}} = \max_v \varepsilon_v.$$

Substituting the worst erroneous prediction into the exact objective yields

$$\delta_0 \mathbb{I}[\exists v : \widetilde{r}_v \neq r^\star_G] + \xi \max_v m(\widetilde{r}_v, r^\star_G) \le \widetilde{\ell}_G + \varepsilon_{\mathrm{search}} + \varepsilon_{\mathrm{dec}}. \tag{27}$$

If the right-hand side is below δ_0, all predictions are correct. Exact scoring sets both approximation terms to zero. For approximate search, the guarantee requires an upper bound on the missed violation; negative sampling or beam search alone does not provide this bound.

Exact scoring and implementation. Canonical scoring ranks one terminated string per relation, without summing over paraphrases. The guarantee applies to this deterministic sequence-MAP predictor; greedy decoding and random sampling need not preserve it.

Exact candidate scoring can use limited activation storage: for fixed parameters and deterministic scoring, enumerate candidate–view violations in chunks and retain a maximizing index. If its violation is positive, recomputing only that branch with autograd gives the derivative of the full maximum wherever the active branch is unique. At ties, an active branch yields a valid Clarke subgradient; zero is also active at the hinge boundary. The two passes use identical parameters, token masks, score normalization, and deterministic model settings. Chunking reduces activation memory while retaining the full candidate-scoring cost.

Population bound and view aggregation. For a distribution of groups meeting the assumptions, taking expectations in Eq. 13 gives

$$\Pr_G[\exists v : \hat{r}_v \neq r^\star_G] \le \mathbb{E}_G \ell_{\mathrm{dir}}(G)/\delta_0.$$

This margin-rescaling bound (Tsochantaridis et al., 2005) also applies to the empirical training average, where it bounds the corresponding training error.

For nonnegative per-view losses ℓ_v, $\max_v \ell_v = 0$ iff $V^{-1}\sum_v \ell_v = 0$. During optimization, the maximum focuses on the currently most violated observed views within each group.

A.5 Intervention Profiles and Joint Feasibility

The intervention objective links changes in answer support to measured geometric changes. We derive its closed-form scale fit and show that its constraints are compatible with correct rankings across views.

Cancellation and profile fitting. Replacing $s(x, r)$ by $s(x, r) + a(x) + b(r)$ leaves Eq. 14 unchanged: input contributions cancel within each gap, and answer contributions cancel between the two gaps. For $U_v = (H_{v,k} - 2\delta_0)_{k=1}^K$ and $\mathbf{m}_B = (m_k)_{k=1}^K \neq 0$, the fitted scale is

$$\eta^\star_v = \max\left\{0, \frac{\mathbf{m}_B^\top U_v}{\|\mathbf{m}_B\|_2^2}\right\}. \tag{28}$$

The profile loss vanishes exactly when $U_v = \eta^\star_v \mathbf{m}_B$ for every camera. Before camera averaging, its gradient with respect to U_v is $2(U_v - \eta^\star_v \mathbf{m}_B)/K$; the analytic projection needs no iterative optimization of the scale. With $K = 1$, any nonnegative excess can be fitted, motivating multiple distinct magnitudes.

Joint zero-loss properties. If every direct group loss vanishes, both gaps in Eq. 14 are at least $\delta_0 + \xi m_k$. Hence $U_{v,k} \geq 2\xi m_k$. If the interaction loss also vanishes, $m_k > 0$ implies $\eta_v^\star \geq 2\xi$. The direct objective ensures correct predictions, while the intervention objective structures their combined support according to geometric magnitude. The fitted scale accommodates differences in confidence across cameras, and the two individual gaps may differ.

Proposition 8 (Joint feasibility in conditional score space)**.** *Consider a finite collection of groups and blocks with truths in $\mathcal{C}$, such that identical canonical inputs have identical truths. For any finite $t \geq \xi$, a normalized conditional distribution with finite canonical log scores can make all direct and interaction losses vanish simultaneously.*

Proof. Write $r_x^\star$ for the truth of input x and set $f(x, r) = -\delta_0 \mathbb{I}[r \neq r_x^\star] - t\, m(r, r_x^\star)$. For fixed $a \in (0, 1)$, assign

$$\pi(\psi(r) \mid x) = \frac{a \exp f(x, r)}{\sum_{r' \in \mathcal{C}} \exp f(x, r')},$$

and place the remaining mass on noncanonical terminated strings. Distinct terminated verbalizations are disjoint outcomes, so these are valid finite canonical sequence probabilities. Normalization adds only an input-dependent constant. Every truth–competitor gap is therefore $\delta_0 + t\, m(r, r_x^\star) \geq \Delta(r, r_x^\star)$, and every intervention contrast equals $2\delta_0 + 2tm_k$. Thus all direct losses vanish and $\eta_v = 2t$ makes every profile residual zero. The same t works for overlapping blocks and shared inputs. □

Prop. 8 establishes simultaneous feasibility of the direct and interaction constraints in conditional score space.

Matched profile controls. Let $g_{v,k}^{(0)}$ and $g_{v,k}^{(1)}$ denote the two individual gaps in Eq. 14. The single-gap control replaces ℓ_{int} by

$$\ell_{\text{single}}(B) = \frac{1}{2V_B K} \sum_v \min_{a,b \geq 0} \sum_k \left[(g_{v,k}^{(0)} - \delta_0 - am_k)^2 + (g_{v,k}^{(1)} - \delta_0 - bm_k)^2 \right]. \tag{29}$$

It fits two independent nonnegative scales per camera and block. Its zero loss implies zero interaction loss with $\eta_v = a + b$; the converse need not hold because the two gaps can trade residuals. The EqSim-style control uses $\ell_{\text{sym}}(B) = (V_B K)^{-1} \sum_{v,k} (g_{v,k}^{(0)} - g_{v,k}^{(1)})^2$. This control adapts the gap-symmetry principle of Wang et al. (2023) to intervention pairs. It penalizes differences between the two gaps while leaving their dependence on geometric magnitude unconstrained. Both controls retain the same pair and direct terms; their coefficients are selected with equal validation-search budgets.

B Data Construction Details

The construction pipeline produces measured supervision by filtering object pairs, perturbing their relations, and verifying the resulting response- and image-side comparisons. We describe this pipeline, then report corpus composition, serialized margin statistics, and data-quality checks. Preference pairs, view groups, and intervention blocks are counted separately.

B.1 Pair Selection and Perturbation

Object pairs are retained if both objects have annotated visibility ratio at least 0.4, lie within 2 to 50 meters of the camera, and project to bounding boxes of at least 800 square pixels. The retained corpus covers five families, namely metric distance, cross-view comparisons, graded errors, clock direction, and vertical relations. Tab. 9 gives their release and training-pool counts. Evaluation units are specified separately in App. C.2.

For cross-view supervision, a response-side group with V views contains V preference pairs. The complete view group is retained for direct canonical supervision; view comparisons are not counted as additional preference pairs.

Graded perturbations. The initial candidate pool perturbs the clock and the distance jointly, at a severity fixed by its grade. Minor shifts the clock by one hour and scales the distance by a factor drawn uniformly from $[1.3, 1.6]$; moderate by three hours and $[1.8, 2.5]$; fatal by six hours and $[3.0, 5.0]$. The lateral and depth terms are left correct at the minor grade, flipped independently with probability $\frac{1}{2}$ each at moderate, and both flipped at fatal. For the final measured training pool, retained perturbations are reciprocal/sign balanced and every intermediate step is recomputed from the perturbed geometry; the one-sided and rationale-inconsistent variants are controls in Tab. 18, not the final construction. Grades use nominal sampling weights $2 : 3 : 2$ and the reported realized counts are 2,701, 4,021, and 2,671, totaling 9,393. For the stated schedule, $\alpha = 0.5$, and $\kappa = \log 5$, the analytic joint ranges are

grade	$\lvert\delta_c\rvert_{12},\ \delta_d$	m before serialization
minor	$1,\ [1.3, 1.6]$	$[0.164841, 0.229348]$
moderate	$3,\ [1.8, 2.5]$	$[0.432606, 0.534662]$
fatal	$6,\ [3, 5]$	$[0.841303, 1]$

These analytic ranges precede serialization. Each coordinate contributes at most 0.5 with the stated weights, so $m = 1$ requires both coordinates to saturate. Dataset quantiles and extrema are computed separately from serialized labels.

B.2 Task Spaces, Response Generation and Verification

The geometric metric covers clock direction and positive inter-object distance. Measured pairs use the relevant coordinate subset, and ordering-only comparisons, including vertical relations, use zero offset as defined in App. A.2. For graded responses, the geometric margin is computed from the parsed terminal relation. Width and height assess transfer to spatial quantities beyond the training metric.

Response generation and parsing. Chosen and rejected responses share five steps: *Locate anchor*, *Side (left/right)*, *Depth (front/back)*, *Direction (clock)* and *Distance (meters)*. They verbalize the true and perturbed values, respectively, and close with the clock hour and metric distance. The parser Π reads this terminal statement independently of intermediate steps.

Geometric verification. The renderer checks visibility, image bounds, projected object area, relation validity, and duplicate views. Camera-separation checks apply to groups with a fixed object configuration, while image-side intervention pairs share a camera. Template parseability and chosen-label correctness are audited on the candidate pool before filtering. The measured training branch retains only pairs with an exactly correct chosen label and a valid rejected label with positive geometric margin. Tab. 11 reports the candidate-pool audit, and Tab. 9 gives the retained corpus counts.

B.3 Corpus Composition and Margin Statistics

The preference stage combines response- and image-side comparisons with direct supervision of view groups and intervention supervision of multi-state blocks, following Eq. 16.

Corpus composition. Tab. 9 gives the task-family composition of the 50,436-pair Gauge-50K release and the 30,000-pair preference-training pool. These counts describe available preference pairs. The sampling schedule and optimizer-update budget are specified in App. C.1; view groups and intervention blocks are counted separately in App. C.4.

Serialized geometric margins. Clock labels use integer hours and distances use meters to two decimal places. Margins are computed from the final serialized labels with the coordinate weights in Eq. 4. Tab. 10 reports their distributions for the 2,701 minor, 4,021 moderate, and 2,671 fatal perturbations. The three grades occupy distinct ranges, and all retained graded labels have positive margins. Ordering-only preferences use zero offset as described in App. A.2.

Table 9: Corpus allocation by task family for the release and the preference-training pool.

Family	Release pairs	Training-pool pairs
Metric distance	17800	10600
Cross-view	12900	7600
Graded errors	9393	5600
Clock direction	7219	4300
Vertical relation	3124	1900
Total	50436	30000

Table 10: Serialized margin distribution by error grade. Distances use two decimal places; Zero counts labels with zero margin.

Grade	Pairs	Min	p_{10}	Median	p_{90}	Max	Zero
minor	2701	0.1646	0.1718	0.1992	0.2234	0.2299	0
moderate	4021	0.4324	0.4444	0.4883	0.5259	0.5347	0
fatal	2671	0.8413	0.8615	0.9298	0.9872	1.0000	0

B.4 Data Quality and Split Checks

Tab. 11 reports checks on two distinct sets. A sample of $1{,}000$ candidate preference pairs assesses template parseability and agreement with scene ground truth before filtering. The 200 camera groups used in Tab. 8 are checked for anchor-frame identifiability, metric-scale evidence, and completeness. Their pass rates describe different properties of the construction and evaluation sets and should not be pooled. Scene/configuration identifiers and rendered-image hashes are also checked for overlap between training and testing among the inspected units. The fallback conventions and treatment of incomplete camera groups are specified in App. C.3.

C Training and Evaluation Protocols

We specify optimization and sampling budgets, followed by item and pair scoring, fixed-relation camera audits, and multi-state intervention evaluation. Input conventions, invalid-prediction rules, and comparison settings are defined alongside the evaluations to which they apply.

C.1 Training Configuration and Implementation

Base configuration. SFT applies LoRA to the language-model parameters with rank 128 and $\alpha_{\mathrm{LoRA}} = 128$, while the vision encoder remains frozen. SFT uses learning rate 5×10^{-6} and context length 4096. The SFT adapter is merged before preference training; a frozen copy of this merged SFT checkpoint serves as the reference. Preference learning then uses full-parameter fine-tuning of the model and vision encoder, with learning rate 3×10^{-7} and context length 2048, using a cosine schedule, 0.03 warmup ratio and bf16. The geometric-margin coefficients are $\beta = 0.3$, $\alpha = 0.5$, and $\kappa = \log 5$, with offset $\gamma = \beta m$. The joint objective uses $(\delta_0, \xi) = (0.1, 1)$ and $(\lambda_{\mathrm{dir}}, \lambda_{\mathrm{int}}) = (0.1, 0.5)$. Images contain at most $262{,}144$ pixels. The software stack comprises LLaMA-Factory $0.9.5$, PyTorch $2.7.1$, Transformers $5.2.0$ and DeepSpeed ZeRO stage 3. Reported end-to-end model-training times on four A100 80GB GPUs, including SFT and preference training but excluding rendering and reward-estimator fitting, are 7.7, 8.4 and 12.2 hours for Qwen, GLM and Pixtral.

Preference sampling and budget. The full preference schedule draws $30{,}000$ pairs with replacement from the fixed $30{,}000$-pair pool in Tab. 9. With an effective batch size of 32 and a final partial batch, this requires 938 optimizer updates. The budget experiments use 10%, 25%, 50%, and 100% of this pair-draw budget while sampling from the same pool. The 25% setting therefore uses $7{,}500$ draws; it does not restrict access to a quarter of the distinct examples in the pool. These experiments assess performance under a reduced preference pair-draw budget. View-group and intervention-block exposure are separate quantities, with the full block schedule reported in App. C.4.

Table 11: Data-quality and observability checks. Rows 1–2 use 1,000 candidate preference pairs; rows 3–5 use the 200 camera groups from Tab. 8. The final two rows count train/test duplicates among the inspected units. Wilson 95% intervals refer to the pass rate in rows 1–5 and the duplicate rate in rows 6–7. Tab. 12 describes the handling of camera groups requiring fallbacks or marked incomplete.

Check	Counting unit	Audited	Pass / dup.	Rate (%)	Wilson 95%
Template parseability	preference pair	1000	997	99.7	[99.1, 99.9]
Chosen truth matches scene state	preference pair	1000	995	99.5	[98.8, 99.8]
Anchor frame identifiable	evaluation group	200	187	93.5	[89.2, 96.2]
Scale cue sufficient for metric answer	evaluation group	200	178	89.0	[83.9, 92.6]
Complete fixed-configuration groups	evaluation group	200	194	97.0	[93.6, 98.6]
Train/test duplicate scene IDs	scene–configuration	200	0	0.0	[0.0, 1.9]
Train/test duplicate image hashes	rendered item	400	0	0.0	[0.0, 1.0]

Loss-weight selection. The coefficient sweeps use

$$\lambda_{\text{dir}} \in \{0, 0.03, 0.1, 0.3, 1\}, \qquad \lambda_{\text{int}} \in \{0, 0.25, 0.5, 0.75, 1\}.$$

Each sweep holds the other coefficient at its default and retains the optimizer settings above. The shared default yields nine distinct configurations (Tab. 5); setting a coefficient to zero removes the corresponding loss. Hyperparameters are selected on scenes disjoint from training and testing. All main results and ablations use the final checkpoint at their specified update budget. Matched comparisons share the SFT initialization, reference model, training data, update budget, and random seed.

Canonical candidate space. The candidate distance grid is $\mathcal{D}_Q = \{0.1, 0.2, \ldots, 10.0\}$ meters, giving 1,200 clock–distance candidates. Quantization selects the nearest absolute log-distance, breaking ties toward the smaller distance; clock labels retain their declared discretization. Coverage is checked on training scenes before freezing this domain independently of test truths; out-of-domain relations are counted separately, without clipping. The verbalizer uses the fixed terminal-relation template with two distance decimals and model-specific termination. Scores sum answer tokens through termination, excluding prompt and padding, without length normalization; sequence-MAP ties use clock order $1, \ldots, 12$ then ascending distance. Evaluation metadata retain source, scene, configuration, camera, and object identifiers, together with visibility, scale, and parsing information. Native benchmark results use their respective generation and scoring protocols.

C.2 Task Metrics and Intervention Pair Scoring

Numerical accuracy. For a positive prediction $\hat{d}$ and ground truth d, define $q = \hat{d}/d$ and $r_\times = \max(q, q^{-1})$. Ratio accuracy uses $r_\times < t$: $t = 1.25$ for MSMU (Chen et al., 2025) and $t = 2$ for the QSpatial reference implementation (Liao, 2024). The corresponding intervals are $q \in (0.8, 1.25)$ and $q \in (0.5, 2)$. The $\pm 25\%$ band in Fig. 3 instead uses relative error $|\hat{d} - d|/d \leq 0.25$, equivalent to $q \in [0.75, 1.25]$.

Diagnostic interpretation. Spearman correlation measures the rank agreement between predicted and true distances, using midranks for ties. It is undefined for constant predictions; the marked zeros in the main table indicate these cases. VisC. measures prediction agreement when brightness and contrast vary within $[0.8, 1.2]$ while scene geometry and camera pose remain fixed. We report these diagnostics alongside accuracy because agreement alone can be achieved by constant outputs. Cross-camera correctness is evaluated separately below.

Object-relocation accuracy. The object-relocation test contains 205 paired left/right questions (410 individual questions). Each pair shares the scene, camera, and anchor, while moving the target reverses its left/right relation. PairAcc. is the percentage of pairs for which both answers are correct. GaugeVLM-7B correctly answers both states in 136 of 205 pairs, yielding 66.3%. This metric uses the left/right subset of the 710-question intervention probe; the remaining 300 questions assess other relations.

Label balance and parsing. Each pair contributes one left and one right label, giving 205 examples of each. All models are evaluated on the same pairs, with answer options mapped to semantic labels and missing or invalid answers counted as incorrect. Independent uniform binary predictions yield 25% expected PairAcc.; a random first prediction followed by a forced reversal yields 50%. Tab. 17 evaluates sensitivity to answer-option and state order.

C.3 Fixed-Relation Camera Evaluation

The camera evaluation in Tab. 8 measures accuracy and agreement across views of a fixed object configuration. Each group shares the scene, anchor frame, and allocentric ground truth. Questions request clock direction and positive distance in that common frame. The metric compares answers expressed in this common frame. The data-quality audit appears in App. B.4; the corresponding input conventions are specified below.

Let $\mathcal{G}_{\text{view}}$ be the predeclared set of fixed-relation groups, and let $\mathcal{G}_{\text{valid}}$ contain those for which every required view parses successfully. Let V_g contain the group's required views, with $|V_g| \geq 2$, and write $(c_{g,v}, d_{g,v})$ for its parsed allocentric answers. Define

$$\text{ViewAgree}_{\text{tol}} = \frac{100}{|\mathcal{G}_{\text{view}}|} \sum_{g \in \mathcal{G}_{\text{valid}}} \mathbb{I}\left[\max_{u,v \in V_g} d_{\text{ring}}(c_{g,u}, c_{g,v}) \leq \tau_c \ \wedge \ \max_{u,v \in V_g} d_{\text{rel}}(d_{g,u}, d_{g,v}) \leq \tau_d \right]. \tag{30}$$

Invalid or incomplete groups contribute zero and remain in the denominator. Prop. 1 relates disagreement to geometric error for these groups. We evaluate correctness and agreement jointly, as described below. Missing required views and unparseable predictions follow the same invalid-group rule.

Agreement, correctness, and coverage. For a group with all required views parsed, let A_g^{tol} denote passing the two spread thresholds in Eq. 30, and define

$$C_g^{\text{tol}} = \mathbb{I}\left[\forall v : d_{\text{ring}}(c_v, c_g^*) \leq \epsilon_c \ \wedge \ d_{\text{rel}}(d_v, d_g^*) \leq \epsilon_d\right].$$

We set both indicators to zero for invalid or incomplete groups. Correctness uses $\epsilon_c = 0$ and $\epsilon_d = \log 1.25 / \log 5$, independently of the agreement tolerances. Among valid groups, the four outcomes $(A_g^{\text{tol}}, C_g^{\text{tol}}) \in \{0,1\}^2$ distinguish accurate agreement, incorrect agreement, accurate disagreement, and incorrect disagreement. Invalid or incomplete groups form a separate category; all rates use the full evaluation set.

Accurate agreement and wrong agreement are the averages of $100 A_g^{\text{tol}} C_g^{\text{tol}}$ and $100 A_g^{\text{tol}}(1 - C_g^{\text{tol}})$, respectively. Group accuracy averages $100 C_g^{\text{tol}}$, while item accuracy evaluates individual views. Correct predictions within the truth tolerance can still differ by more than the independently chosen agreement tolerance.

For exact canonical evaluation, define

$$C_g^{\text{can}} = \mathbb{I}[\forall v : \hat{r}_v = r_g^\star],$$
$$A_g^{\text{can}} = \mathbb{I}[\text{all required predictions are valid and identical}].$$

Both indicators are zero for invalid or incomplete groups. The resulting metrics are

$$\text{GroupAcc}_{\text{can}} = 100 \overline{C}^{\text{can}}, \tag{31}$$
$$\text{ViewAgree}_{\text{can}} = 100 \overline{A}^{\text{can}}, \tag{32}$$
$$\text{WrongAgree}_{\text{can}} = 100 \overline{A^{\text{can}}(1 - C^{\text{can}})}. \tag{33}$$

Exact correctness implies agreement, so $\text{ViewAgree}_{\text{can}} = \text{GroupAcc}_{\text{can}} + \text{WrongAgree}_{\text{can}}$. These exact metrics correspond to the canonical prediction theorem; the tolerance-based evaluation uses the criteria specified above.

Tolerance units. For $\kappa = \log 5$ and $\tau_d = 0.25$, passing the distance-spread test means that the largest to smallest stated distance ratio is at most $5^{0.25} \approx 1.495$. This is not a $\pm 25\%$ relative-error test. The direction tolerance $\tau_c = 1/6$ is one clock step. The thresholds apply to the normalized metrics and do not enter binary PairAcc.

Table 12: Handling of reference-frame, scale, and completeness checks for the 200 camera groups. Each row partitions the full set into the four listed outcomes. Fallback groups are scored using the specified convention; incomplete groups remain in the denominator and are scored as invalid. No groups are excluded (Excl.).

		Outcome				
Check	Audited	Pass	Fallback	Invalid	Excl.	Fallback rule
Anchor frame	200	187	13	0	0	dominant wall normal
Scale cue	200	178	22	0	0	stated reference extent
Group completeness	200	194	0	6	0	none; scored invalid

Input observability and fallback conventions. Tab. 12 specifies the conventions used to label camera groups that require a reference-frame or scale fallback. These conventions define the targets but do not by themselves establish that the targets are identifiable from the model inputs. Identifiability depends on the orientation and scale information available in those inputs. Fallback groups remain in the evaluation set. The 6 incomplete groups likewise remain in the denominator and are scored as invalid for every model. Additional failures to parse a required prediction also invalidate its group. The model-specific counts are reported in App. D.2.

Object motion and viewpoint changes. The two evaluations assess complementary behaviors. Fixed-configuration camera groups test whether the model predicts the same true relation from different views. Fixed-camera object-relocation pairs test whether its answers follow a change in that relation. The multi-state blocks below combine both forms of variation, enabling joint evaluation of intervention response and viewpoint consistency.

C.4 Block Generalization and Matched Controls

Source scenes are split before constructing states, cameras, or pairs. All test settings use held-out scenes. Familiar settings reuse the training ranges of magnitudes and camera factors; unseen-magnitude settings hold out declared geometric magnitudes, and unseen-combination settings hold out magnitude–camera combinations while retaining each individual factor in training. The manifest specifies these sets, quantization and boundary rules. Held-out magnitudes and held-out joint combinations do not overlap their respective training sets after serialization. Blocks use $K \geq 2$ interventions with at least two distinct positive m_k and $V_B \geq 2$ common cameras. The evaluation design in Tab. 7 allocates three equally sized settings of 200 blocks each; Overall pools all 600 blocks.

Block construction and held-out factors. Tab. 13 reports the block pools, scene counts, numbers of interventions K, and camera counts V_B for each split. Training samples blocks with replacement; the table gives both pool size and the number of distinct blocks visited. Sampling assigns equal nominal weight to the magnitude–camera strata in Tab. 14, with realized frequencies recorded in the training metadata.

Tab. 14 specifies the distance factors, camera-separation bands, and held-out combinations. Held-out distance factors lie between retained training levels, so this setting evaluates generalization to unseen magnitudes within the training range. The combination setting holds out specific magnitude–camera pairs while retaining each factor individually in training.

All supervision variants share scenes, block/state inputs, pair and direct supervision, training updates and selection data. Pair + direct sets $\lambda_{\text{int}} = 0$; full adds Eq. 15; the two controls replace it by the objectives in App. A.5. Coefficient selection receives the same validation budget, and search/decoding conventions are fixed. All variants use the same evaluation blocks.

Metric aggregation. BlockAcc and mean worst m use all 600 test blocks. A block with any missing or invalid required prediction is scored as incorrect and assigned the maximal geometric error of 1. For all methods, the profile diagnostic E_{shape} uses the same intervention contrast $U_v = H_v - 2\delta_0 \mathbf{1}$ from Eq. 18.

Table 13: Intervention-block statistics. Training draws 3,752 blocks with replacement over 938 updates, visiting 1,447 of the 1,600 training blocks. The five splits use disjoint scene sets. Strata correspond to the magnitude–camera combinations in Tab. 14.

		Blocks		Min/median/max			
Split	Role	Distinct	Covered	K	V_B	Scenes	Strata
Train	GaugeDPO updates	1600	1447	2/3/4	2/2/2	800	16
Val	coefficient selection	200	200	2/3/4	2/2/2	100	16
Test	familiar	200	200	2/3/4	2/2/3	100	16
Test	unseen magnitudes	200	200	2/3/4	2/2/3	100	8
Test	unseen combinations	200	200	2/3/4	2/2/3	100	4

A profile is eligible when U_v is finite and $\|U_v\|_2 > \tau_U$, with $\tau_U = 0.05$ fixed before evaluation. Each block–camera pair contributes one profile containing its K intervention responses. Thus, the 600 test blocks contribute $N_{\text{prof}} = \sum_{B \in \mathcal{B}_{\text{test}}} V_B$ profiles in total. Eligible coverage is computed for each method over this common total, and the shared eligible set contains 1,022 profiles. Tab. 6 evaluates both E_{shape} and the fraction with $\eta_v^\star = 0$ on this shared set, ensuring that the profile comparisons use identical observations. All blocks remain in the accuracy and geometric-error metrics. Scene-level paired resampling retains all states, cameras, and blocks from each sampled scene.

Transfer and comparison controls. Training uses 3D-FRONT, MSMU, and OSD. We check overlap across training stages and evaluation using scene/configuration identifiers and image hashes. Comparisons document the images, regions, geometric cues, prompts, decoding, and parsing used by each model. Tab. 22 reports both restricted-input and native-input results for baselines that use additional visual cues. Kimi-K2.6 is an open-weight mixture-of-experts model with 1T total parameters and 32B activated parameters per token (Moonshot AI, 2026a); the main table reports total parameters.

Reward-estimator control. The reward estimator predicts the serialized geometric margin from image and relation-pair features. It is fitted on training scenes with a disjoint validation split, and its outputs are clipped to $[0, 1]$. The text-only control uses the two serialized relations. Tab. 23 reports margin MAE, rank correlation, and fitting and scoring costs against the serialized geometric target. Direct computation evaluates the target formula from scene-state labels and requires no estimator fitting. The input bundles differ across methods, so these results quantify target approximation and compute under the listed inputs.

D Additional Experimental Results

Supplementary experiments assess transfer to additional spatial and embodied benchmarks, sensitivity to loss weights, robustness to input changes and training seeds, and the effect of baseline inputs and reward estimation. Base, SFT, and GaugeVLM checkpoints use the same evaluation pipeline; comparisons against SFT measure the gains from the complete preference stage. Matched objective controls separate the effects of individual supervision terms.

D.1 Additional evaluation

Additional spatial benchmarks. Tab. 15 extends the evaluation to SPAR-Bench (Zhang et al., 2025), QSpatial$^+$ (Liao et al., 2024), Omni3D-Bench (Marsili et al., 2025), and SPBench-SI (Li et al., 2026), covering spatial perception and reasoning, quantitative distance estimation, and object-size estimation. GaugeVLM improves over SFT on every reported metric, including gains of 8.7 points on SPAR-Bench overall, 18.9 points on QSpatial$^+$, and 5.4 points on SPBench-SI size estimation. On the multiple-choice component of Omni3D-Bench (MC), it reaches 60.5%, a gain of 5.8 percentage points. The QSpatial$^+$ result is repeated from the main table for reference. Since task definitions and scoring rules differ across benchmarks, we report their individual metrics rather than aggregate them into a single score.

Table 14: **Distance factors and camera configurations.** Distance levels are reported as the normalized component $d_{\text{rel}} = \min(|\log \lambda|, \kappa)/\kappa$, where $\kappa = \log 5$ and λ is the centroid-distance ratio. The joint margin m_k uses Eq. 4 with the default $\alpha = 0.5$; for a distance-only change, $m_k = (1-\alpha)d_{\text{rel}} = 0.5d_{\text{rel}}$. NN denotes the distance to the nearest training level in these normalized distance units. Camera bands specify azimuth separation. Held-out combinations retain both constituent factors in training.

Factor	Level	λ	d_{rel}	Split role
Magnitude	M1	1.25	0.1386	train, familiar test
Magnitude	M2	1.50	0.2519	train, familiar test
Magnitude	M3	2.20	0.4899	train, familiar test
Magnitude	M4	3.60	0.7959	train, familiar test
Magnitude	M5	4.50	0.9345	train, familiar test
Magnitude	H1	1.80	0.3652	held out; NN 0.1133
Magnitude	H2	2.80	0.6397	held out; NN 0.1498
Camera	V1	$[30°, 60°)$		train and test
Camera	V2	$[60°, 90°)$		train and test
Camera	V3	$[90°, 120°)$		train and test
Camera	V4	$[120°, 150°]$		train and test
Combination	(M1,V4)	—		held out, both factors seen
Combination	(M2,V3)	—		held out, both factors seen
Combination	(M4,V1)	—		held out, both factors seen
Combination	(M5,V2)	—		held out, both factors seen

Table 15: **Additional metric and spatial benchmark results.** Task definitions differ across benchmarks; driving results appear in the main table.

Benchmark	Domain	GaugeVLM main metric	$\Delta_{\text{vs SFT}}$
SPAR-Bench	indoor room (ScanNet)	overall 34.0	+8.7
SPAR-Bench (mv)	robotic multi-view	dist_oo 38.3	+16.5
QSpatial$^+$	indoor/mixed	64.4	+18.9
Omni3DBench	mixed	MC 60.5	+5.8
SPBench-SI	indoor desktop	size 49.8	+5.4

Spatial capabilities relevant to embodied tasks. Tab. 16 further examines multi-view distance estimation on SPAR-Bench (Zhang et al., 2025), spatial configuration understanding on RoboSpatial (Song et al., 2025), and spatial referring on RefSpatial-Bench (Zhou et al., 2025). Here, `dist_oc` and `dist_oo` denote object–camera and object–object distance estimation, respectively. Relative to SFT, GaugeVLM improves these two SPAR-Bench metrics by 14.1 and 16.5 points, while RoboSpatial configuration accuracy increases from 74.8% to 78.1% and the RefSpatial-Bench overall score rises from 16.0 to 17.0. These results support transfer to spatial capabilities relevant to embodied interaction, with a smaller improvement on spatial referring. The SPAR-Bench `dist_oo` result is shared with Tab. 15 and represents the same evaluation.

D.2 Camera Outcomes and Shortcut Controls

The camera evaluation uses 100 scenes, two object configurations per scene, and two camera views per configuration, yielding 200 groups and 400 required views. Each group shares its anchor frame and true clock–distance relation. Under the tolerances in App. C.3, GaugeVLM achieves 69.5% group accuracy and 63.0% accurate agreement, compared with 50.0% and 40.0% for SFT (Tab. 8). Incorrect agreement decreases from 20.0% to 11.5%. These results show that improved agreement is accompanied by improved correctness. All models use the same evaluation groups and denominator, with the scoring rules in App. C.3. The 6 incomplete groups contribute 3.0% to the invalid rate for every model. Additional parsing failures account for 4 groups in SFT, 2 in the pair-only ablation, and 1 in each single-weight ablation. GaugeVLM produces valid outputs for all complete groups, giving an invalid rate of 3.0%.

Table 16: **Additional embodied benchmark results.**

Task (bench)	Domain	sub-metric	base	SFT	GaugeVLM	$\Delta_{\text{vs SFT}}$
SPAR (mv)	embodied	dist_oc (obj-cam)	15.7	27.1	**41.2**	+14.1
SPAR (mv)	embodied	dist_oo (obj-obj)	19.8	21.8	**38.3**	+16.5
RoboSpatial	embodied	configuration	69.1	74.8	**78.1**	+3.3
RefSpatial	embodied	overall	13.0	16.0	**17.0**	+1.0

Table 17: **Input and presentation controls.** Intervention PairAcc. and accurate agreement use separate evaluation sets.

Condition	Intervention PairAcc	Accurate agreement
Default options and state order	66.3	63.0
Permuted answer options	65.9	62.5
Reversed state presentation	65.4	62.5
Image masked	25.9	18.5
Text only	26.8	19.5

Shortcut controls. The controls below separate image dependence, option ordering and negative-construction cues. They complement the matched objective definitions in App. C.4.

Tab. 17 evaluates the same checkpoint under changes to image availability, answer-option order, and state presentation. PairAcc. uses 205 relocation pairs, and accurate agreement uses the 200 camera groups. Permuting options or reversing state order changes PairAcc. by less than one percentage point. Masking the image or using text alone reduces PairAcc. from 66.3% to 25.9% and 26.8%, respectively, indicating reliance on visual input. Predictions are mapped back to semantic labels before scoring.

Tab. 18 separates the effects of reciprocal/sign balancing and coherent reasoning steps in preference construction. Text-only preference accuracy is evaluated on 200 held-out pairs, with randomized candidate order and scene-disjoint fitting and validation. Combining balancing and coherence reduces text-only preference accuracy from 88.0% to 54.5%, while increasing accurate camera agreement from 54.0% to 63.0%. The combined construction therefore reduces detectable textual shortcuts while improving spatial performance.

D.3 Training-Seed Robustness and Paired Uncertainty

Matched preference-stage runs. We use five random seeds, $\{11, 23, 37, 53, 71\}$, for each of seven Qwen2.5-VL-7B objectives in Tab. 19. All 35 runs start from the same SFT checkpoint and frozen reference. Source scenes, pair and block pools, train/validation/test splits, hyperparameters, and evaluation manifests are fixed across seeds. Each run uses 30,000 pair draws, 938 optimizer updates, and four block draws per update. We evaluate the final scheduled checkpoint on the test set using the same checkpoint-selection rule for all runs.

Across seeds, we vary the training RNG streams governing pair/block sampling, data order, and any enabled stochastic model operations. Within a seed, all objectives share the same serialized pair/block draw schedule and aligned stochastic-operation streams wherever applicable. Separate RNG streams handle objective-specific operations, including the group-preserving shuffled-margin permutation. Removing a loss does not change the data exposure of the remaining branches. Zero-, constant-, and shuffled-offset pair losses change only the pair term; their direct and interaction terms remain active. Pair + direct sets $\lambda_{\text{int}} = 0$, while the two profile controls replace only the interaction objective. All settings are frozen before the five-seed sweep.

Evaluation units and retained predictions. Every run uses identical prompts, preprocessing, deterministic decoding, canonical candidate search, parsing, and invalid-output rules. Native benchmark generation and canonical sequence-MAP evaluation remain distinct protocols. The intervention test contains 600 blocks from 300 disjoint scenes: each of three settings contains 100 scenes and two blocks per scene. The camera audit contains 200 groups from 100 scenes. The relocation test contains 205 pairs; all pairs from the same source scene remain one statistical cluster. Public

Table 18: **Construction controls.** Text-only preference accuracy measures discrimination of chosen/rejected pairs, not spatial answer accuracy.

Construction	MSMU	QSpat	Acc. agree.	Text-only pref. acc.
One-sided, inconsistent rationale	60.0	62.4	54.0	88.0
Reciprocal/sign-balanced only	62.5	63.4	56.5	79.0
Coherent rationale only	60.0	63.4	59.0	66.5
Balanced and coherent	62.5	64.4	63.0	54.5

Table 19: **Preference-stage robustness.** Mean $\pm$ sample SD computed from five accuracy scores per cell. Accuracy values are percentages.

Objective	MSMU distance	QSpatial^{+}	PairAcc	GroupAcc$_{\mathrm{tol}}$	BlockAcc
Zero-offset pair loss	49.5 ± 2.1	49.1 ± 1.7	56.2 ± 1.5	60.2 ± 1.4	55.6 ± 1.5
Constant pair offset	52.0 ± 1.1	52.1 ± 1.5	57.8 ± 1.3	62.0 ± 1.3	57.3 ± 1.4
Shuffled pair offset	52.0 ± 2.1	51.3 ± 1.8	57.4 ± 1.6	61.5 ± 1.5	56.8 ± 1.6
Pair + direct	57.5 ± 1.8	61.0 ± 2.8	62.0 ± 1.2	67.2 ± 1.2	59.7 ± 1.3
+ Single-gap profile	59.0 ± 1.4	60.8 ± 1.1	63.3 ± 1.1	68.0 ± 1.1	63.4 ± 1.2
+ EqSim-style symmetry	59.5 ± 1.1	61.6 ± 1.3	63.9 ± 1.2	68.3 ± 1.2	64.7 ± 1.2
GaugeDPO (full)	62.0 ± 1.1	64.0 ± 1.1	66.0 ± 1.0	69.3 ± 1.0	69.6 ± 1.1

benchmarks are clustered by source scene when scene identifiers are available, and otherwise by source image. We retain every run's checkpoint hash, configuration, sampled-unit IDs, predictions, parse flags, and per-unit scores, including failures.

Training variation and paired differences. For each method, we report the mean and sample SD of its five seed-level scores on the unchanged test manifest. For a matched comparison, let $d_s = M_{\mathrm{full},s} - M_{\mathrm{control},s}$. The seed-only 95% interval is $\bar{d} \pm t_{0.975,4}\, s_d/\sqrt{5}$; it quantifies training variation conditional on this test set and, for Tab. 19, this SFT checkpoint. We also report the number of seeds with $d_s > 0$; ties are not wins.

Uncertainty across seeds and test scenes. We additionally use 10,000 paired bootstrap replicates with analysis seed 2027. Each replicate independently resamples the five seed indices and the test-scene clusters with replacement. The same resampled seed indices and scene multiset are used for both methods; the scene multiset is also shared across all sampled seeds. All questions, pairs, configurations, views, and blocks belonging to a selected scene are retained together. For the block test, scene resampling is stratified by the three test settings, preserving their equal weights. We recompute each metric from its original numerator and denominator, then average the paired differences over sampled seeds. The 2.5th and 97.5th percentiles give the seed-and-scene interval. This separates training SD from uncertainty in the method difference and avoids counting repeated predictions on the same scene as independent test observations. Tab. 20 shows positive gains for all reported comparisons across all five seeds. The joint seed-and-scene intervals also remain above zero, supporting robustness to both training variation and the sampled test scenes.

Independent two-stage replications. To assess sensitivity beyond a fixed SFT initialization, we separately repeat the complete SFT–preference pipeline five times for each backbone. Each repetition independently trains SFT from the released base model, then branches into matched zero-offset pair-loss and full-GaugeDPO preference runs, sharing that repetition's SFT reference and draw schedules. Stage-specific RNG streams are derived from the repetition seed. SFT and preference update budgets and the final-checkpoint rule are fixed across repetitions. This adds five SFT runs and ten preference runs per backbone: 15 SFT and 30 preference runs across three backbones. Tab. 21 shows consistent gains across all three backbones when variation in both training stages is included.

Table 20: **Paired gains over controls.** Full GaugeDPO minus each control, in percentage points. Primary comparisons are specified before the repeated runs. Both confidence intervals describe the paired difference.

Control	Metric	Mean Δ	95% CI seeds only	95% CI seeds + scenes	Positive seeds / 5
Primary comparisons					
Shuffled pair offset	QSpatial$^+$	$+12.7$	$[10.7, 14.7]$	$[8.8, 16.4]$	5 / 5
Pair + direct	BlockAcc	$+9.9$	$[8.3, 11.5]$	$[6.9, 12.8]$	5 / 5
Secondary comparisons					
Zero-offset pair loss	QSpatial$^+$	$+14.9$	$[12.9, 16.8]$	$[11.1, 18.5]$	5 / 5
Constant pair offset	QSpatial$^+$	$+11.9$	$[9.9, 13.8]$	$[8.3, 15.4]$	5 / 5
Single-gap	BlockAcc	$+6.2$	$[4.7, 7.7]$	$[3.5, 8.9]$	5 / 5
EqSim-style	BlockAcc	$+4.9$	$[3.2, 6.6]$	$[2.1, 7.6]$	5 / 5
Pair + direct	Unseen-mag. BlockAcc	$+10.6$	$[8.2, 13.0]$	$[6.2, 15.1]$	5 / 5
Pair + direct	Unseen-comb. BlockAcc	$+10.4$	$[7.7, 13.1]$	$[5.4, 15.3]$	5 / 5

Table 21: **Robustness of the complete training pipeline.** Five independent SFT–preference repetitions per backbone. Entries are mean $\pm$ sample SD in percent. SFT checkpoints are shared by the two preference objectives within each repetition.

Backbone	Training	MSMU distance	QSpatial$^+$	PairAcc	BlockAcc
Qwen2.5-VL-7B	SFT	47.0 ± 2.1	45.3 ± 1.9	54.2 ± 1.8	45.2 ± 2.0
	Zero-offset pair loss	49.5 ± 2.1	49.1 ± 1.8	56.3 ± 1.7	55.4 ± 1.9
	GaugeDPO	62.0 ± 2.1	63.8 ± 1.5	65.8 ± 1.4	69.4 ± 1.5
GLM-4.1V-9B	SFT	67.0 ± 2.1	49.1 ± 2.1	67.0 ± 1.9	47.8 ± 2.1
	Zero-offset pair loss	68.5 ± 2.2	52.5 ± 1.9	67.8 ± 1.7	56.9 ± 1.9
	GaugeDPO	72.0 ± 2.1	58.0 ± 1.7	69.5 ± 1.5	71.2 ± 1.6
Pixtral-12B	SFT	49.5 ± 2.7	7.7 ± 1.5	31.8 ± 2.1	38.4 ± 2.3
	Zero-offset pair loss	51.0 ± 2.2	9.3 ± 1.5	35.6 ± 2.0	45.1 ± 2.1
	GaugeDPO	52.0 ± 2.1	11.7 ± 1.5	40.6 ± 1.8	52.7 ± 1.9

D.4 Baseline Inputs and Reward-Estimator Comparisons

These comparisons assess the effects of additional baseline inputs and the accuracy and cost of estimating the geometric target. Their settings are specified in App. C.4.

Restricted and native inputs. Tab. 22 compares restricted inputs with the additional visual cues supported by each baseline. Native cues improve MSMU distance accuracy from 25.0% to 30.0% for SpatialReasoner and to 32.5% for SpatialRGPT. GaugeVLM achieves 62.5% on MSMU distance and 64.4% on QSpatial$^+$, exceeding both native-input baselines on these two metrics under the listed configurations.

Approximation error and computational cost. The image-conditioned regressor in Tab. 23 achieves 0.087 margin MAE and 0.72 rank correlation against the serialized geometric target. Direct computation requires no estimator fitting and 0.02 GPU-hours for scoring in the reported setting. Its zero MAE and unit rank correlation express algebraic agreement with the target formula; they do not measure error against continuous physical geometry. The table therefore complements the downstream margin ablation by reporting the cost and approximation error of the target-construction path under each method's stated inputs.

Table 22: Restricted-input and native-input comparison. Each native configuration retains the model's specified cue bundle.

Model	Input mode	MSMU dist.	QSpatial^{+}
SpatialReasoner	restricted	25.0	49.5
SpatialReasoner	native cue bundle	30.0	53.5
SpatialRGPT	restricted	25.0	55.4
SpatialRGPT	native cue bundle	32.5	60.4
GaugeVLM	standard input	62.5	64.4

Table 23: Reward-estimator comparison against serialized geometric targets. The direct-margin row gives algebraic agreement with the target formula.

Estimator	Inputs	Margin MAE	Rank ρ	Fit GPU-h	Score GPU-h
Scalar regressor	image + two relations	.087	.72	1.2	.3
Text-only regressor	two serialized relations	.104	.65	.5	.1
Direct geometric margin	scene-state labels	0.000	1.00	0.0	.02