WHOSE VIEW WINS? INSIDE EGOCENTRIC BIAS IN VISUOSPATIAL PERSPECTIVE TAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visuospatial perspective taking (VSPT) requires reasoning from a viewpoint different from the viewer's own, yet Multimodal Large Language Models (MLLMs) often exhibit egocentric bias. We investigate the representational mechanism underlying this failure using controlled scenes that independently manipulate egocentric and altercentric spatial relations. Linear probing shows that both relations are decodable from hidden representations despite poor VSPT performance under perspective conflict, suggesting that altercentric information is not simply missing. We then decompose hidden representations into egocentric and altercentric components and quantify their contributions along the ego–alter decision direction. Under conflict, the egocentric component contributes systematically more strongly toward its corresponding answer than the altercentric component. This imbalance is accompanied by a large and consistent egocentric advantage in component magnitude, while alignment differences are weaker and model-dependent. Guided by this analysis, we intervene on magnitude and alignment separately; both improve conflict-case performance, with magnitude suppression generally yielding larger gains. The decomposition also transfers to natural-image benchmarks in several settings. Together, our results suggest that egocentric bias stems from an egocentric-favoring imbalance in how concurrently represented perspectives contribute to model predictions.

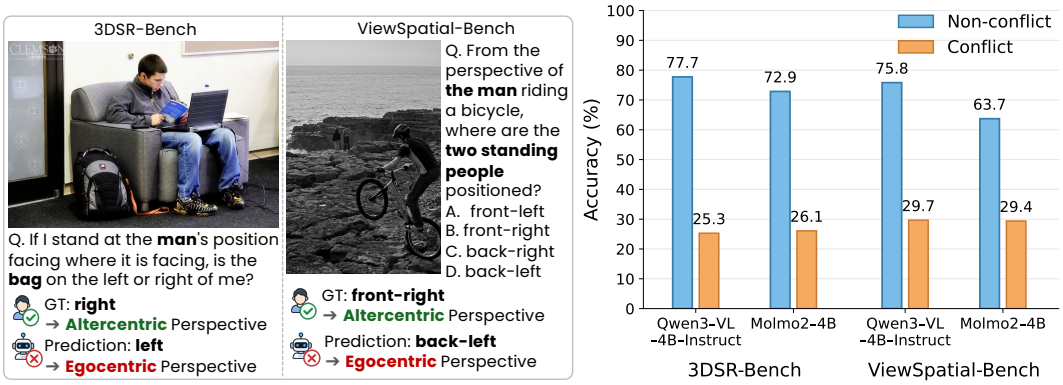
1 INTRODUCTION

Visuospatial Perspective Taking (VSPT) is the ability to understand visual and spatial relations in a scene from a viewpoint that differs from one's own (Quesque et al., 2024; Samuel et al., 2024). In particular, VSPT often requires reasoning from a specified viewpoint (*i.e.*, an *altercentric perspective*) rather than the viewer's own viewpoint (*i.e.*, *the egocentric perspective*). This capability is important for Multimodal Large Language Models (MLLMs), particularly in applications such as navigation, embodied interaction, and human–AI interaction, where models must interpret spatial relations from viewpoints other than their own.

Despite this importance, MLLMs often fail to reason consistently from the viewpoint specified by the task. A prominent failure mode is *egocentric bias*, in which models rely on the viewer's own perspective even when they are instructed to reason from an altercentric perspective (Zhang et al., 2026; Wang et al., 2026; Lee et al., 2025). Figure 1a provides a qualitative example of this behavior, where the model produces an egocentric answer despite being instructed to adopt the target agent's perspective. Figure 1b further shows that this is a systematic failure: performance drops substantially when the egocentric and altercentric perspectives imply conflicting answers, with errors frequently corresponding to the egocentric alternative. These observations motivate us to investigate the source of egocentric bias in MLLMs.

To identify where egocentric bias arises, we conduct a stepwise representation-level analysis using a controlled dataset in which egocentric and altercentric spatial relations can be independently manipulated. Our analysis addresses three sequential questions. **(Q1) Are both egocentric and altercentric spatial relations represented?** Linear probing shows that both perspectives are linearly decodable with high accuracy, despite substantially lower VSPT performance, suggesting that the failure cannot be explained simply by the absence of altercentric information. **(Q2) How does each perspective contribute to the model's prediction?** We estimate perspective subspaces, jointly

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



(a) Examples of VSPT failures across benchmarks (b) Accuracy with and without perspective conflict

Figure 1: **Egocentric bias in visuospatial perspective taking.** (a) Representative VSPT failures from 3DSR-Bench (Ma et al., 2025) and ViewSpatial-Bench (Li et al., 2026), where predictions favor the egocentric perspective. (b) Performance drops substantially under perspective conflict, with errors often favoring the egocentric answer.

decompose hidden states into egocentric and altercentric components, and quantify their contributions along the ego–alter decision direction through the frozen LM head. Under perspective conflict, the egocentric component contributes substantially more strongly toward its corresponding answer than the altercentric component. **(Q3) What gives rise to this contribution asymmetry?** We decompose each component's contribution into its magnitude and alignment with the ego–alter decision direction, finding a large and consistent egocentric asymmetry in component magnitude, while alignment shows a weaker and more model-dependent asymmetry.

Based on these findings, we intervene directly on the identified contribution imbalance by reducing the magnitude of the egocentric component or rotating it away from the ego–alter output direction, thereby reducing its relative influence on the output while preserving the altercentric information already encoded in the representation. Under perspective conflict, these interventions consistently reduce egocentric errors and improve perspective-taking performance. Together, these results suggest that egocentric failures in VSPT cannot be explained solely by a lack of altercentric information; rather, they arise from an imbalance in how concurrently represented egocentric and altercentric information contributes to the model's prediction.

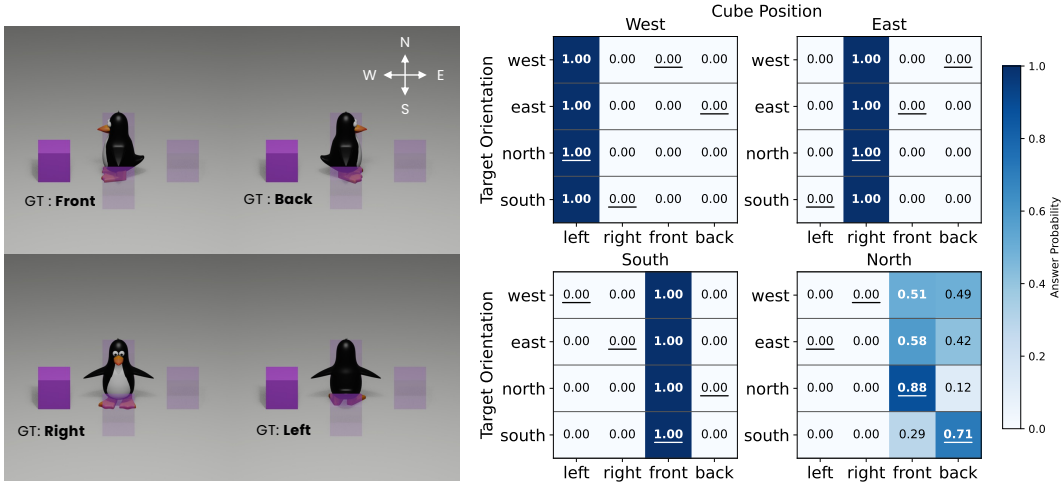
Our main contributions are threefold: (1) We show that both egocentric and altercentric spatial relations are decodable from MLLM representations, ruling out missing altercentric information as the sole explanation for egocentric bias. (2) We decompose representations into egocentric and altercentric components and reveal a systematic egocentric advantage in output contributions under perspective conflict. (3) We show that factor-specific interventions on magnitude and alignment improve VSPT, causally linking contribution asymmetry to egocentric bias.

2 Egocentric Bias in Visuospatial Perspective Taking

Throughout our analysis, we consider two reference frames for spatial relations: the *egocentric perspective*, corresponding to the viewer or camera viewpoint, and the *altercentric perspective*, corresponding to the viewpoint of the target agent. We use *altercentric* to denote this target-centered, decentered perspective, distinguishing it from *allocentric* representations defined relative to an environment- or world-centered reference frame (Mazzarella et al., 2013; Job et al., 2022; Ang et al., 2023). We define a case as *non-conflict* when the egocentric and altercentric perspectives yield the same answer, and as *conflict* when they yield different answers.

2.1 Controlled Diagnostic Dataset

Natural-image VSPT benchmarks entangle perspective with variations in scene layout, appearance, object position, and target orientation (Ma et al., 2025; Li et al., 2026), making perspective-specific effects difficult to isolate. We therefore construct a controlled diagnostic dataset of minimal scenes



(a) Examples from the controlled dataset

(b) Answer probabilities across controlled conditions

Figure 2: **Egocentric bias on the controlled dataset.** (a) Controlled dataset is constructed from 16 combinations of four cube positions and four target orientations. The central target agent is varied across samples. (b) Qwen3-VL-4B-Instruct answer probabilities across cube positions and target orientations. Despite changes in target orientation, predictions remain strongly concentrated on the answer associated with the cube position, revealing a systematic egocentric bias.

containing a single target agent and a cube. As illustrated in Figure 2a, fixing the agent and cube positions while varying only the agent's orientation changes the altercentric relation without altering the overall scene configuration. By systematically varying cube position and agent orientation, we control egocentric and altercentric relations while minimizing unrelated visual variation. The dataset is balanced across front, back, left, and right relations under both perspectives and rendered in Blender using 13 visually distinct agents. This controlled setting supports the behavioral and representation-level analyses that follow.

2.2 Egocentric Bias in the Controlled Setting

Before proceeding to representation-level analysis, we verify that the controlled setting reproduces the egocentric bias observed in existing VSPT benchmarks. As shown in Figure 2b, answer probabilities largely follow the cube's image-space position, which determines the egocentric relation, while remaining relatively insensitive to changes in the target agent's orientation that determine the altercentric relation. This confirms that the controlled setting preserves the characteristic egocentric bias and provides a suitable basis for the analyses that follow.

3 Egocentric and Altercentric Information Are Linearly Decodable

A simple explanation for egocentric bias is that MLLMs fail to encode the altercentric spatial relation required by the task. We test this hypothesis by examining whether egocentric and altercentric relations are linearly decodable from the models' hidden representations.

3.1 Linear Probing

For each layer l, we train two linear probes on the last-token hidden representation $h^{(l)}$: one predicts the egocentric spatial relation and the other predicts the altercentric spatial relation. Each probe predicts one of four classes: front, back, left, or right. All model parameters are frozen, and probe performance is evaluated using 3-fold cross-validation with disjoint agent identities across folds.

3.2 Both Perspectives Are Linearly Decodable

As shown in Figure 3, egocentric spatial relations are nearly perfectly decodable from early layers. Altercentric relations are also decodable, reaching 70.2% accuracy in Qwen3-VL-4B-Instruct (Bai et al., 2025) and 67.7% in Molmo2-4B (Clark et al., 2026) at later layers, well above the 25% chance level. Notably, this altercentric probe performance remains substantially higher than the models' actual VSPT performance under perspective conflict.

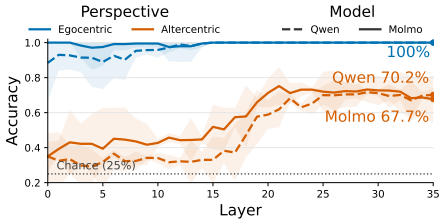


Figure 3: Linear probing reveals decodable egocentric and altercentric spatial relations across decoder layers.

These results show that egocentric bias cannot be explained solely by the absence of altercentric information in the hidden representations. Rather, both perspective-specific relations are linearly accessible even when model predictions remain egocentrically biased. This raises the next question: if both perspectives are represented, how does each contribute to the model's output decision?

4 Asymmetric Use of Perspective Information

Section 3 shows that both egocentric and altercentric spatial relations are linearly decodable from the hidden representations, despite strongly egocentric predictions. We therefore ask how information from the two perspectives is reflected in the model's output decision. To study this, we decompose each hidden representation into egocentric and altercentric components and quantify their contributions to the output.

4.1 Decomposing Representations by Perspective

Constructing Perspective Subspaces. Leveraging our controlled setting, we estimate a low-dimensional spatial subspace for each perspective from class-conditional mean representations. For each perspective $p \in \{\text{ego}, \text{alter}\}$ and spatial class $c \in \{\text{front}, \text{back}, \text{left}, \text{right}\}$, we compute the centered class-mean direction

$$\mathbf{d}p, c = \frac{1}{|\mathcal{D}p, c|} \sum_{i \in \mathcal{D}p,c} \mathbf{h}i - \boldsymbol{\mu}, \tag{1}$$

where $\mathcal{D}p, c$ denotes the samples whose spatial relation under perspective p is c, and $\boldsymbol{\mu}$ is the global mean representation. We stack these directions and apply singular value decomposition:

$$M_p = [\mathbf{d}_{p,\text{front}}, \mathbf{d}_{p,\text{back}}, \mathbf{d}_{p,\text{left}}, \mathbf{d}_{p,\text{right}}] \in \mathbb{R}^{d\times 4}, \qquad M_p = U_p \Sigma_p V_p^\top. \tag{2}$$

The left singular vectors spanning the class-mean variation form an orthonormal basis

$$B_p = [\mathbf{u}_{p,1}, \ldots, \mathbf{u}_{p,r_p}], \qquad r_p = \text{rank}(M_p) \le 3, \tag{3}$$

whose column space defines the empirical spatial subspace associated with perspective p.

Joint Representation Decomposition. Given a hidden representation $\mathbf{h}$, we first center it as $\tilde{\mathbf{h}} = \mathbf{h} - \boldsymbol{\mu}$. Because the egocentric and altercentric subspaces need not be orthogonal, we jointly decompose the representation rather than projecting onto each subspace independently. We concatenate the two bases as

$$B_{\text{joint}} = [B_{\text{ego}}, B_{\text{alter}}] \in \mathbb{R}^{d\times(r_{\text{ego}}+r_{\text{alter}})}. \tag{4}$$

We then obtain the joint coefficient vector using the pseudoinverse:

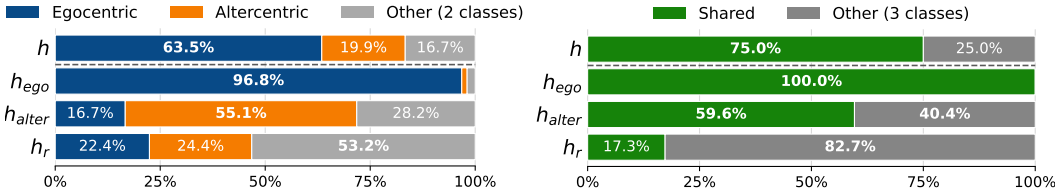
$$\boldsymbol{\alpha} = B_{\text{joint}}^{+} \tilde{\mathbf{h}} = [\boldsymbol{\alpha}_{\text{ego}}\ \boldsymbol{\alpha}_{\text{alter}}]. \tag{5}$$

The corresponding egocentric and altercentric components are reconstructed as

$$\mathbf{h}_{\text{ego}} = B_{\text{ego}} \boldsymbol{\alpha}_{\text{ego}}, \qquad \mathbf{h}_{\text{alter}} = B_{\text{alter}} \boldsymbol{\alpha}_{\text{alter}}. \tag{6}$$

We define the residual and reconstruct the original hidden representation as

$$\mathbf{h}_r = \tilde{\mathbf{h}} - \mathbf{h}_{\text{ego}} - \mathbf{h}_{\text{alter}}, \qquad \mathbf{h} = \boldsymbol{\mu} + \mathbf{h}_{\text{ego}} + \mathbf{h}_{\text{alter}} + \mathbf{h}_r. \tag{7}$$



(a) Conflict Samples (b) Non-conflict samples

Figure 5: **Semantic validation of decomposed perspective components for Molmo2-4B at layer 25.** Frozen LM-head readouts for (a) conflict and (b) non-conflict samples. The decomposed components favor the answers implied by their corresponding perspectives.

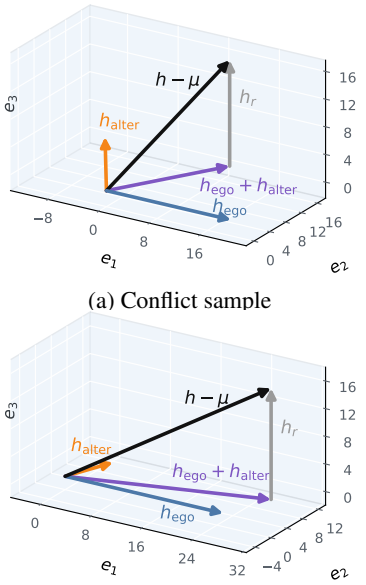
Here, $\mathbf{h}_r$ captures the remaining variation outside the joint reconstruction. Figure 4 provides a geometric illustration of this decomposition for representative conflict and non-conflict examples, showing how the egocentric, altercentric, and residual components combine to reconstruct the centered representation.

Semantic Validation of Decomposed Components. We next examine whether the decomposed components exhibit semantic preferences consistent with their corresponding perspectives. We pass the original representation and each decomposed component through the frozen LM head and record the highest-scoring answer among front, back, left, and right, and compute its selection rate.

For conflict samples, the egocentric and altercentric perspectives correspond to different answers. As shown in Figure 5a, $\mathbf{h}_{\text{ego}}$ strongly favors the egocentric answer, selecting it for more than 96% of samples. $\mathbf{h}_{\text{alter}}$ preferentially selects the altercentric answer in 55.1% of samples, a substantial increase from 19.9% with the original representation $\mathbf{h}$ and well above the 25% uniform baseline. The residual $\mathbf{h}_r$ shows no comparable preference for either perspective-relevant answer.

For non-conflict samples, both perspectives yield the same spatial relation. As shown in Figure 5b, $\mathbf{h}_{\text{ego}}$ selects this shared answer for all samples, while $\mathbf{h}_{\text{alter}}$ favors it in 59.6% of samples, again above the uniform baseline. The residual shows no clear preference for the shared answer.

Together, these results indicate that the decomposed components are semantically aligned with their corresponding perspectives, although the altercentric component exhibits weaker selectivity than the egocentric component. This provides the semantic basis for comparing how strongly the two components contribute to the model's output decision.



(a) Conflict sample

(b) Non-conflict sample

Figure 4: **Perspective decomposition examples.** Qwen3-VL-4B-Instruct at layer 25.

4.2 Perspective Contribution Asymmetry in Output Decisions

Having established that the decomposed components exhibit semantics consistent with their corresponding perspectives, we next ask how strongly each component contributes to the model's output decision when the two perspectives conflict, using representations from layer 35 for both models.

Ego–Alter Decision Direction For a conflict sample, let $\mathbf{w}_{\text{ego}}$ and $\mathbf{w}_{\text{alter}}$ denote the LM-head weight vectors corresponding to the egocentric and altercentric answers, respectively. We define the ego–alter logit margin as

$$M = \text{logit}_{\text{ego}} - \text{logit}_{\text{alter}} = (\mathbf{w}_{\text{ego}} - \mathbf{w}_{\text{alter}})^\top \mathbf{h}. \tag{8}$$

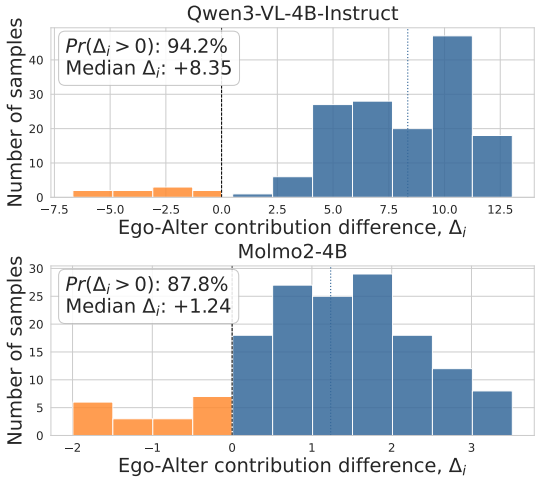


Figure 6: **Distribution of perspective-aligned contribution differences.** The dashed line denotes equal contribution ($\Delta_i = 0$), and the dotted line denotes the median.

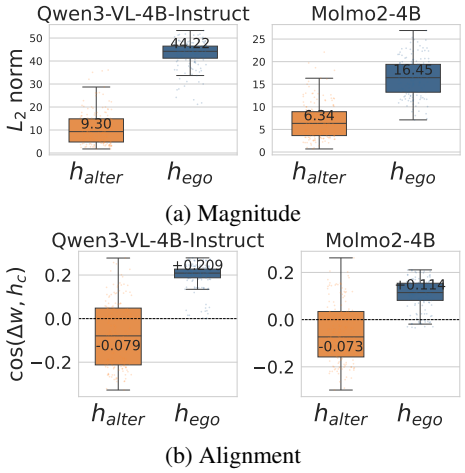


Figure 7: **Magnitude and Alignment of Perspective Components.** (a) Component ℓ_2 norms $||\mathbf{h}_c||$. (b) Perspective-aligned cosine similarity $\cos(\Delta\mathbf{w}, \mathbf{h}_c)$.

Letting $\Delta\mathbf{w} = \mathbf{w}_{\text{ego}} - \mathbf{w}_{\text{alter}}$, a positive margin favors the egocentric answer, whereas a negative margin favors the altercentric answer.

Component-wise Contribution. Substituting the decomposition in Eq. 7 gives

$$M = \Delta\mathbf{w}^\top \boldsymbol{\mu} + \Delta\mathbf{w}^\top \mathbf{h}_{\text{ego}} + \Delta\mathbf{w}^\top \mathbf{h}_{\text{alter}} + \Delta\mathbf{w}^\top \mathbf{h}_r. \tag{9}$$

We quantify the signed contribution of each component $\mathbf{h}_c$ along the ego–alter decision direction as

$$s_c = \frac{\Delta\mathbf{w}^\top \mathbf{h}_c}{\|\Delta\mathbf{w}\|_2} \tag{10}$$

Positive values indicate support toward the egocentric answer, whereas negative values indicate support toward the altercentric answer.

Results. For each conflict sample i, we compute $\Delta_i = s_{\text{ego},i} - (-s_{\text{alter},i})$, where each term measures the component's signed support toward its corresponding perspective-specific answer. Thus, $\Delta_i > 0$ indicates that the egocentric component provides stronger support toward the egocentric answer than the altercentric component provides toward the altercentric answer.

As shown in Figure 6, Δ_i is positive for 94.2% of Qwen3-VL-4B-Instruct samples and 87.8% of Molmo2-4B samples, with median differences of 8.35 and 1.24, respectively. This reveals a systematic imbalance under perspective conflict, with the egocentric component contributing more strongly toward its corresponding answer than the altercentric component.

4.3 Factors Underlying Perspective Contribution Asymmetry

We next examine what underlies the asymmetric contributions of the egocentric and altercentric components. From Equation 10, the signed contribution of a component can be expressed as

$$s_c = \|\mathbf{h}_c\|_2 \cos(\Delta\mathbf{w}, \mathbf{h}_c), \tag{11}$$

where $\|\mathbf{h}_c\|_2$ denotes the component magnitude and $\cos(\Delta\mathbf{w}, \mathbf{h}_c)$ denotes its alignment with the ego–alter decision direction. We analyze these two factors separately to identify which exhibits a consistent asymmetry between the two perspectives.

Component Magnitude. We first examine whether the two components differ in magnitude. As shown in Figure 7a, the egocentric component has a substantially larger norm than the altercentric component in both models. The median ℓ_2 norms are 44.22 versus 9.30 for the egocentric and altercentric components in Qwen3-VL-4B-Instruct, and 16.45 versus 6.34 in Molmo2-4B. Thus, the egocentric component exhibits a large and consistent magnitude advantage across models.

Component Alignment. We next examine whether a similar asymmetry appears in alignment with the output decision direction. Because the egocentric and altercentric answers lie on opposite sides of the ego–alter direction, we compare alignment toward each component's corresponding answer using $\cos(\Delta\mathbf{w}, \mathbf{h}_{\text{ego}})$ for the egocentric component and $-\cos(\Delta\mathbf{w}, \mathbf{h}_{\text{alter}})$ for the altercentric component.

As shown in Figure 7b, the median perspective-aligned values are 0.209 versus 0.079 for the egocentric and altercentric components in Qwen3-VL-4B-Instruct, and 0.114 versus 0.073 in Molmo2-4B. The alignment difference is therefore more pronounced in Qwen3-VL-4B-Instruct, whereas the two distributions substantially overlap in Molmo2-4B. Unlike the magnitude imbalance, alignment does not exhibit the same degree of egocentric asymmetry consistently across models.

Taken together, the egocentric component exhibits a large and consistent magnitude advantage, whereas alignment differences are weaker and more model-dependent. These results identify component magnitude as the more consistent factor associated with the contribution asymmetry, motivating an intervention that directly reduces the egocentric component's magnitude.

5 METHOD

The analysis in Section 4 shows that the perspective-aligned contribution of a component depends on two geometric factors: its magnitude and its alignment with an output decision direction. We therefore construct two inference-time interventions that manipulate these factors separately.

Egocentric Magnitude Suppression. Given the decomposition in Equation 7, we reduce the magnitude of the egocentric component:

$$\mathbf{h}' = \mathbf{h} - \lambda_{\text{mag}}\mathbf{h}_{\text{ego}}. \tag{12}$$

The suppression strength λ_{mag} controls how much of the egocentric component is removed, ranging from no intervention at $\lambda_{\text{mag}} = 0$ to complete removal at $\lambda_{\text{mag}} = 1$. By reducing this component, we test whether the magnitude asymmetry identified in Section 4.3 contributes to egocentric bias.

Norm-Preserving Alignment Rotation. We next manipulate the alignment of the egocentric component while keeping its magnitude fixed. Following geometry-aware activation rotation (You et al., 2026), we use spherical linear interpolation (Slerp; Shoemake 1985) to rotate the component away from its preferred ego–alter output direction.

For each sample, let $\hat{\mathbf{w}}_{\text{ego}}$ and $\hat{\mathbf{w}}_{\text{alter}}$ denote the LM-head answer vectors most aligned with $\mathbf{h}_{\text{ego}}$ and $\mathbf{h}_{\text{alter}}$, respectively. We define the corresponding output decision direction $\mathbf{v}$ and a target direction $\mathbf{t}$ orthogonal to $\mathbf{v}$ as

$$\mathbf{v} = \frac{\hat{\mathbf{w}}_{\text{ego}} - \hat{\mathbf{w}}_{\text{alter}}}{\|\hat{\mathbf{w}}_{\text{ego}} - \hat{\mathbf{w}}_{\text{alter}}\|_2}, \qquad \mathbf{t} = \frac{\mathbf{h}_{\text{ego}} - (\mathbf{h}_{\text{ego}}^\top \mathbf{v})\mathbf{v}}{\left\|\mathbf{h}_{\text{ego}} - (\mathbf{h}_{\text{ego}}^\top \mathbf{v})\mathbf{v}\right\|_2}. \tag{13}$$

We then rotate $\mathbf{h}_{\text{ego}}$ toward $\mathbf{t}$ using Slerp while preserving its norm:

$$\mathbf{h}'_{\text{ego}} = \|\mathbf{h}_{\text{ego}}\|_2 \,\text{Slerp}\left(\frac{\mathbf{h}_{\text{ego}}}{\|\mathbf{h}_{\text{ego}}\|_2}, \mathbf{t}; \lambda_{\text{align}}\right), \qquad \mathbf{h}' = \mathbf{h} - \mathbf{h}_{\text{ego}} + \mathbf{h}'_{\text{ego}}. \tag{14}$$

The rotation strength λ_{align} interpolates between the original egocentric component at $\lambda_{\text{align}} = 0$ and a component orthogonal to $\mathbf{v}$ at $\lambda_{\text{align}} = 1$, while leaving its magnitude unchanged. When the two components favor the same LM-head answer, the corresponding ego–alter decision direction is undefined. In these cases, we leave the representation unchanged.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

We evaluate on our controlled dataset and two natural-image benchmarks, 3DSR-Bench (Ma et al., 2025) and ViewSpatial-Bench (Li et al., 2026), using Qwen3-VL-4B-Instruct (Bai et al., 2025) and

Benchmark	Model	Intervention	Overall	Conflict	Non-conflict
Controlled	Qwen3-VL-4B	Baseline	29.33	7.69	94.23
		Magnitude	43.75 (+14.42)	42.95 (+35.26)	46.15 (−48.08)
		Alignment	35.58 (+6.25)	16.03 (+8.33)	94.23 (+0.00)
	Molmo2-4B	Baseline	29.81	8.97	92.31
		Magnitude	43.75 (+13.94)	37.82 (+28.85)	61.54 (−30.77)
		Alignment	40.38 (+10.58)	25.64 (+16.67)	84.62 (−7.69)
3DSR	Qwen3-VL-4B	Baseline	44.72	25.29	77.73
		Magnitude	54.27 (+9.55)	49.31 (+24.02)	62.70 (−15.04)
		Alignment	49.57 (+4.85)	38.51 (+13.22)	68.36 (−9.38)
	Molmo2-4B	Baseline	43.42	26.09	72.85
		Magnitude	52.39 (+8.97)	49.89 (+23.79)	56.64 (−16.21)
		Alignment	51.30 (+7.89)	45.17 (+19.08)	61.72 (−11.13)
ViewSpatial	Qwen3-VL-4B	Baseline	36.46	29.67	75.81
		Magnitude	37.41 (+0.95)	37.88 (+8.22)	34.68 (−41.13)
		Alignment	37.53 (+1.07)	31.48 (+1.81)	72.58 (−3.23)
	Molmo2-4B	Baseline	34.44	29.39	63.71
		Magnitude	35.75 (+1.30)	31.06 (+1.67)	62.90 (−0.81)
		Alignment	36.70 (+2.26)	33.15 (+3.76)	57.26 (−6.45)

Table 1: **Effects of factor-specific interventions on VSPT accuracy (%).** Intervention rows report accuracy, with changes from the corresponding baseline in percentage points in parentheses.

Molmo2-4B (Clark et al., 2026). For the natural-image benchmarks, we evaluate only question types that require reasoning from a specified target-agent viewpoint. Detailed question-type selection and filtering criteria are provided in Appendix C.

For in-domain evaluation, we estimate the perspective-specific decomposition and evaluate the intervention using grouped K-fold cross-validation, with samples grouped by image to prevent the same visual scene from appearing across estimation and evaluation folds.

For cross-domain transfer, we estimate the perspective-specific decomposition using the full controlled dataset and transfer the resulting subspaces to 3DSR-Bench and ViewSpatial-Bench.

6.2 In-Domain Results

For each intervention and model–benchmark pair, we sweep the intervention layer and strength, λ_{mag} or λ_{align}. The main results report the highest overall accuracy across the sweep, and the corresponding configuration is used for all subgroup analyses. Full sweep results are provided in Appendix D.2.

Overall Accuracy. Table 1 shows that both interventions improve overall VSPT accuracy across model–benchmark pairs. Magnitude suppression yields the largest gains on the controlled dataset and 3DSR-Bench, reaching improvements of up to 14.42 and 9.55 percentage points, respectively, while gains on ViewSpatial-Bench remain modest.

Conflict vs. non-conflict. Both interventions yield their largest gains when the two perspectives conflict (Table 1), with magnitude suppression improving accuracy by up to 35.26 and 24.02 percentage points on the controlled dataset and 3DSR-Bench, respectively. However, magnitude suppression can substantially degrade performance on non-conflict samples, where the egocentric component already supports the shared correct answer. Figure 8 helps explain this trade-off: among samples that are correct at baseline but become incorrect after suppression, the isolated egocentric component remains correct in all cases, whereas the altercentric component is correct in only 4.85% of cases for Molmo. Thus, suppressing the

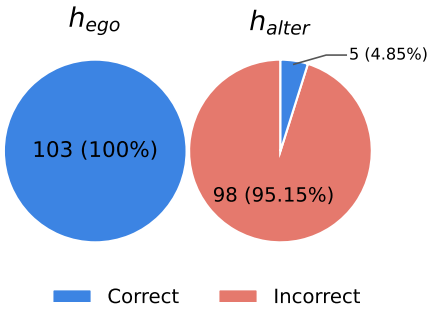


Figure 8: **Component readouts on degraded 3DSR-Bench non-conflict samples.** The $\mathbf{h}_{\text{ego}}$ readout remained correct in all cases, whereas the $\mathbf{h}_{\text{alter}}$ readout was correct in only 4.85% of cases.

egocentric component removes highly reliable evidence for the correct answer that the altercentric component alone often cannot replace. This pattern is consistent with the earlier semantic validation in Figure 5b, where the altercentric component exhibits weaker selectivity than the egocentric component, suggesting that the trade-off partly reflects limited fidelity in the current perspective decomposition. A more accurate separation of altercentric information could therefore preserve the conflict-case benefits of suppression while reducing its degradation on non-conflict cases.

6.3 Domain Transfer

Our preceding analysis highlights the importance of decomposition quality, raising the question of whether the controlled dataset can provide a reusable source for estimating perspective subspaces. We therefore estimate the decomposition only on the controlled dataset and transfer it to 3DSR-Bench and ViewSpatial-Bench without target-benchmark re-estimation. As shown in Table 2, magnitude suppression improves 3DSR-Bench by 5.72 and 7.02 percentage points for Qwen and Molmo, respectively, and ViewSpatial-Bench by 7.72 points for Qwen, although Molmo shows no gain. Notably, for Qwen on ViewSpatial-Bench, the transferred decomposition outperforms direct benchmark-specific estimation, consistent with the balanced factorial structure of the controlled dataset providing cleaner perspective subspaces. Additional analysis is provided in Appendix D.1. Overall, these results suggest that a controlled estimation set can support transferable perspective decomposition, although transfer remains model- and benchmark-dependent.

Benchmark	Model	Intervention	Overall
3DSR	Qwen3-VL-4B	Baseline	44.72
		Magnitude	50.43 (+5.72)
		Alignment	46.02 (+1.30)
	Molmo2-4B	Baseline	43.42
		Magnitude	50.43 (+7.02)
		Alignment	51.01 (+7.60)
ViewSpatial	Qwen3-VL-4B	Baseline	36.46
		Magnitude	44.18 (+7.72)
		Alignment	39.67 (+3.21)
	Molmo2-4B	Baseline	34.44
		Magnitude	34.44 (+0.00)
		Alignment	34.44 (+0.00)

Table 2: **Cross-domain transfer results.**

7 Related Work

Recent benchmarks show that MLLMs struggle with viewpoint-sensitive spatial reasoning, particularly when reasoning from perspectives that differ from the camera view (Ma et al., 2025; Li et al., 2026). Subsequent studies identify persistent egocentric failures even when visual grounding is simplified (Zhang et al., 2026) and suggest that perspective taking remains difficult despite stronger performance on its constituent reasoning steps (Wang et al., 2026). Prior mitigation approaches address these failures by transforming spatial representations into the target reference frame (Lee et al., 2025; Bu et al., 2025) or injecting explicit orientation and perspective information through specialized training or tokens (Leonard & Murray, 2026; Fan et al., 2026). In contrast, our work investigates the internal representational source of egocentric bias, showing how concurrently encoded egocentric and altercentric information contribute asymmetrically to model predictions.

8 Conclusion

We investigate the representational basis of egocentric bias in visuospatial perspective taking. Although both egocentric and altercentric spatial relations are linearly decodable, under perspective conflict the egocentric component contributes systematically more strongly to the output decision. This imbalance is accompanied by a large and consistent egocentric advantage in component magnitude, while alignment differences are weaker and more model-dependent. Factor-specific interventions on magnitude and alignment improve perspective-taking performance, particularly under conflict, with effects that transfer to natural-image benchmarks in several settings. Together, our results suggest that egocentric bias arises not simply from missing altercentric information, but from an egocentric-favoring imbalance in how represented perspectives contribute to model predictions.

AI USE STATEMENT

We used generative AI tools, including ChatGPT and Codex, to assist with brainstorming and refining research directions, supporting implementation and experimental execution, sourcing and summarizing relevant background information, and drafting and polishing portions of the manuscript. These tools served only as assistive aids. All research decisions, experimental designs, implementations, analyses, claims, and manuscript content were reviewed and verified by the authors. We did not use generative AI to generate synthetic datasets or to formulate or prove mathematical claims. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed experimental and evaluation settings in the Appendix. Controlled-dataset construction is described in Appendix A. Model checkpoints, activation extraction, decoding settings, and representation-level analysis procedures are provided in Appendix B. Benchmark filtering, cross-validation protocols, and intervention hyperparameters are reported in Appendix C. We also provide full layer–strength sweeps to characterize intervention sensitivity in Appendix D.

REFERENCES

Natania Ang, Birgit Brucker, David Rosenbaum, Martin Lachmair, Thomas Dresler, Ann-Christine Ehlis, and Peter Gerjets. Exploring the neural basis and modulating factors of implicit altercentric spatial perspective-taking with fnirs. *Scientific Reports*, 13(1):20627, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yuqi Bu, Xin Wu, Zirui Zhao, Yi Cai, David Hsu, and Qiong Liu. Walk in others' shoes with a single glance: Human-centric visual grounding with top-view perspective transformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26904–26923, 2025.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

Wenxiao Fan, Hang Yin, and Kan Li. Orientsam: Mitigating camera-centric shortcut in multimodal spatial reasoning via orientation-aware spatial alignment. *arXiv preprint arXiv:2607.17657*, 2026.

Xavier E Job, Louise P Kirsch, and Malika Auvray. Spatial perspective-taking: insights from sensory impairments. *Experimental brain research*, 240(1):27–37, 2022.

Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9241–9251, 2025.

Bridget Leonard and Scott O. Murray. Cognitively-inspired tokens overcome egocentric bias in multimodal models. *arXiv preprint arXiv:2601.16378*, 2026.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. In *European Conference on Computer Vision*, pp. 95–111. Springer, 2026.

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

Elisabetta Mazzarella, Richard Ramsey, Massimiliano Conson, and Antonia Hamilton. Brain systems for visual perspective taking and action perception. *Social neuroscience*, 8(3):248–267, 2013.

François Quesque, Ian Apperly, Renée Baillargeon, Simon Baron-Cohen, Cristina Becchio, Harold Bekkering, Daniel Bernstein, Maxime Bertoux, Geoffrey Bird, Henryk Bukowski, et al. Defining key concepts for mental state attribution. *Communications Psychology*, 2(1):29, 2024.

Steven Samuel, Thorsten M Erle, Louise P Kirsch, Andrew Surtees, Ian Apperly, Henryk Bukowski, Malika Auvray, Caroline Catmur, Klaus Kessler, and Francois Quesque. Three key questions to move towards a theoretical framework of visuospatial perspective taking. *Cognition*, 247:105787, 2024.

Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pp. 245–254, 1985.

Maijunxian Wang, Yijiang Li, Bingyang Wang, Tianwei Zhao, Ran Ji, Qingying Gao, Emmy Liu, Hokin Deng, and Dezhi Luo. Egocentric bias in vision-language models. *arXiv preprint arXiv:2602.15892*, 2026.

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical steering: Geometry-aware activation rotation for language models. *arXiv preprint arXiv:2602.08169*, 2026.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and DING ZHAO. Spinbench: Perspective and rotation as a lens on spatial reasoning in vlms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 70072–70141, 2026.

APPENDIX

A CONTROLLED DATASET CONSTRUCTION

We construct a controlled dataset of 208 rendered scenes using 13 agent assets obtained from Objaverse (Deitke et al., 2023). Because the canonical forward direction differs across assets, we visually inspect the rendered scenes and verify the orientation assigned to each agent before constructing the perspective-taking questions. The dataset contains 156 conflict examples, in which the egocentric and altercentric answers differ, and 52 non-conflict examples, in which they agree. We use three predefined agent-disjoint folds containing 80, 64, and 64 examples, respectively. The question template used for the controlled dataset is shown in Table 3.

B EXPERIMENTAL DETAILS

B.1 MODELS AND INFERENCE

We use the publicly released Qwen3-VL-4B-Instruct and Molmo2-4B checkpoints (Qwen/Qwen3-VL-4B-Instruct and allenai/Molmo2-4B, respectively). For interventions, we extract activations at the last input (prefill) token from the post-block residual stream, before final normalization, at each of the 36 decoder layers. Model outputs are generated using greedy decoding with sampling disabled, and the generated direction word is used as the prediction.

For both linear probing and representation decomposition, all quantities are estimated using only the training portion of each fold and evaluated on the corresponding held-out fold.

B.2 LINEAR PROBING DETAILS

We train separate linear probes to predict egocentric and altercentric spatial-relation labels from the activation at each layer. Each probe consists of feature standardization followed by ℓ_2-regularized multinomial logistic regression. We use $C = 0.01$, the LBFGS solver, and a maximum of 5,000 optimization iterations. Probe performance is reported using held-out predictions under the predefined agent-disjoint folds.

B.3 DECOMPOSITION DETAILS

Within each training fold, we first center the activations and compute class-mean directions for each egocentric and altercentric spatial label. We apply singular value decomposition to the corresponding direction matrices and retain singular directions whose singular values exceed the numerical rank threshold determined by machine precision. We then use the Moore–Penrose pseudoinverse of the concatenated egocentric and altercentric bases to decompose each activation into egocentric, altercentric, and residual components.

For 3DSR-Bench, we use task-wise centering, estimating separate activation means for the left/right and front/behind question types. For the controlled dataset and ViewSpatial-Bench, we instead use a single global centering mean across questions.

C EVALUATION PROTOCOL

For 3DSR-Bench, we evaluate the `in_front_of` and `on_the_left` relation types within the `orientation` category. For ViewSpatial-Bench, we evaluate the `Person perspective - Relative Direction` question type. These subsets correspond to questions that require reasoning about spatial relations from a specified target-agent viewpoint.

To identify conflict and non-conflict examples in the natural-image benchmarks, we manually annotate the egocentric answer for each example from the camera-centered viewpoint.

For 3DSR-Bench, we retain 691 examples after excluding cases whose spatial relation is difficult to determine from the mirrored view. The resulting evaluation set contains 435 conflict and 256

Table 3: Question template used for controlled dataset

Question Template for Controlled Dataset
Where is the cube from the { target agent }'s perspective?
Answer with exactly one word: front, back, left, or right.

Table 4: **Selected layer and intervention strength for factor-specific interventions.** For each evaluation benchmark, subspace source, model, and intervention type, we report the decoder layer and intervention strength used in evaluation. "Subspace Source" denotes the dataset used to estimate the intervention subspace. Dashes indicate that no layer–strength setting improved over the baseline.

Evaluation Benchmark	Subspace Source	Model	Intervention	Layer	Strength
Controlled	Controlled	Qwen3-VL-4B	Magnitude	35	1.0
			Alignment	35	1.0
		Molmo2-4B	Magnitude	35	0.75
			Alignment	35	0.75
3DSR-Bench	3DSR-Bench	Qwen3-VL-4B	Magnitude	30	1.0
			Alignment	35	0.5
		Molmo2-4B	Magnitude	35	1.0
			Alignment	35	1.0
ViewSpatial-Bench	ViewSpatial-Bench	Qwen3-VL-4B	Magnitude	30	1.0
			Alignment	30	0.5
		Molmo2-4B	Magnitude	30	0.5
			Alignment	30	1
3DSR-Bench	Controlled	Qwen3-VL-4B	Magnitude	35	1.0
			Alignment	35	1.0
		Molmo2-4B	Magnitude	35	1.0
			Alignment	35	1.0
ViewSpatial-Bench	Controlled	Qwen3-VL-4B	Magnitude	35	1.0
			Alignment	35	1
		Molmo2-4B	Magnitude	–	–
			Alignment	–	–

non-conflict examples. To reduce sensitivity to answer-option order, each example is evaluated twice with the order of the two candidate answers reversed (e.g., "left or right" and "right or left"), and performance is averaged across the two orders. We use five folds for evaluation.

For ViewSpatial-Bench, we evaluate 842 examples, consisting of 718 conflict and 124 non-conflict examples, using five folds.

For each model–benchmark–intervention setting, we sweep the intervention layer and strength λ and report the configuration with the highest overall accuracy. The selected configurations for all evaluation settings are reported in Table 4.

D Additional Analysis

D.1 Class Imbalance in ViewSpatial-Bench

As shown in Figure 9, ViewSpatial-Bench exhibits a substantially imbalanced joint distribution of egocentric and altercentric spatial labels. For egocentric labels, left and right occur most frequently, whereas front-left and unclear are relatively rare. Altercentric labels are concentrated primarily

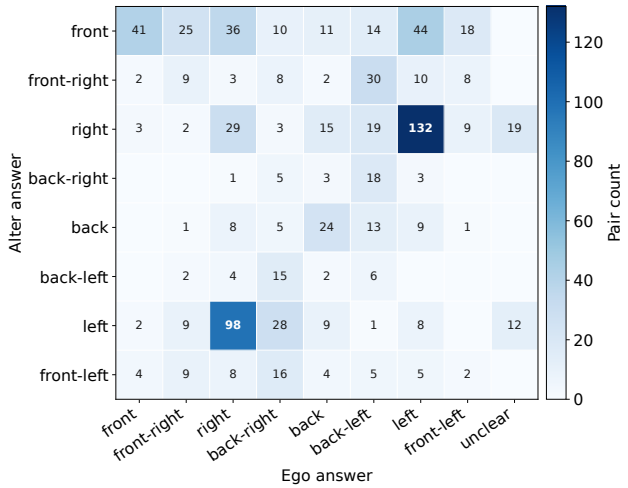


Figure 9: **Class imbalance in ViewSpatial-Bench.** Joint distribution of egocentric and altercentric spatial labels in the evaluated subset.

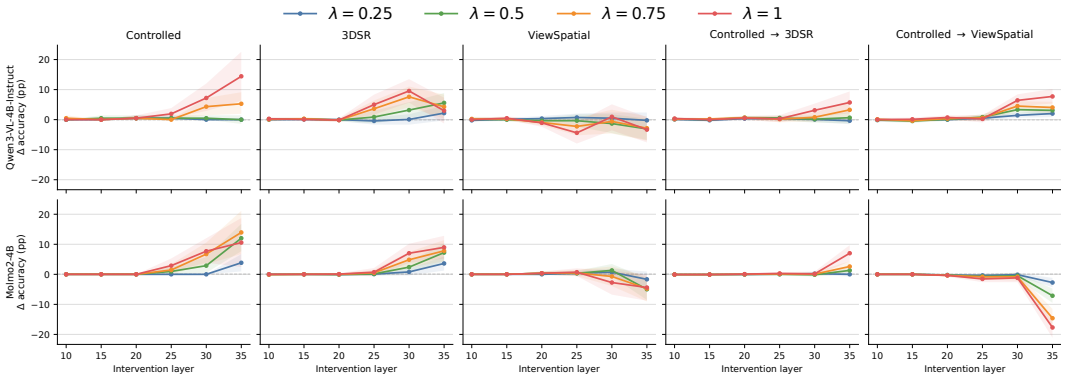


Figure 10: **Magnitude intervention layer–strength sweep.** Accuracy change relative to the baseline (percentage points) across intervention layers and strengths λ_{mag}, for Qwen3-VL-4B and Molmo2-4B. Columns show in-domain and cross-domain transfer settings; the dashed horizontal line denotes the baseline.

in right, front, and left, while back-right and back-left occur less frequently. The joint ego–alter distribution is similarly uneven, with a small number of label pairs accounting for a large fraction of the data.

This imbalance may affect the class-mean estimates used to construct the perspective-specific subspaces, because directions associated with frequent labels and label pairs are estimated from substantially more examples than those associated with rare configurations.

D.2 Full Layer-Strength Sweeps

We conduct layer–strength sweeps over layers $\{10, 15, 20, 25, 30, 35\}$. The full sweep results for magnitude suppression and alignment rotation are shown in Figure 10 and Figure 11, respectively.

D.3 Matched-Gating Ablation

Alignment rotation is applied only when the egocentric and altercentric components favor different LM-head answers. To control for this difference in intervention coverage, we additionally apply the same gating criterion to magnitude suppression. For samples that do not satisfy the criterion, the original representation is left unchanged.

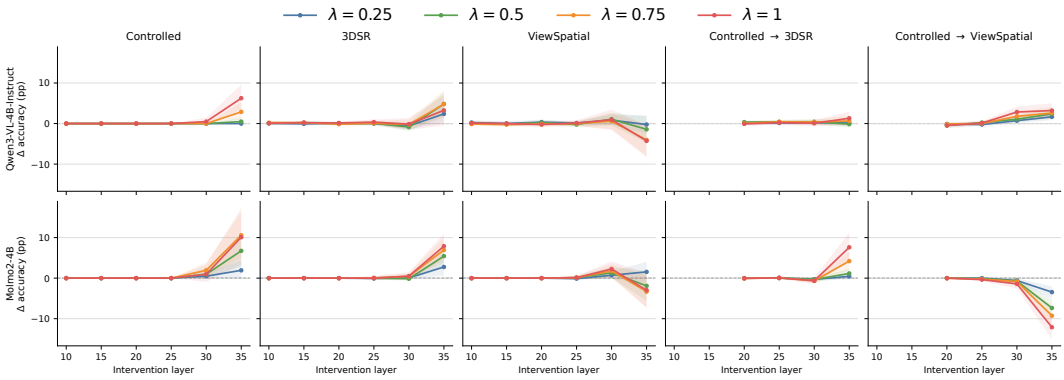


Figure 11: **Alignment intervention layer–strength sweep.** Accuracy change relative to the baseline (percentage points) across intervention layers and strengths λ_{align}, for Qwen3-VL-4B and Molmo2-4B. Columns show in-domain and cross-domain transfer settings; the dashed horizontal line denotes the baseline.

Table 5: **Magnitude suppression under matched alignment gating.** We apply magnitude suppression only when the egocentric and altercentric components favor different LM-head answers, matching the applicability criterion of alignment rotation.

Setting	Model	Layer	Strength	Overall	Conflict	Non-conflict
Controlled	Qwen3-VL-4B	35	1.00	44.71 (+15.38)	39.10 (+31.41)	61.54 (-32.69)
	Molmo2-4B	35	0.75	42.79 (+12.98)	35.90 (+26.92)	63.46 (-28.85)
3DSR	Qwen3-VL-4B	30	1.00	54.12 (+9.41)	48.51 (+23.22)	63.67 (-14.06)
	Molmo2-4B	35	1.00	52.24 (+8.83)	49.66 (+23.56)	56.64 (-16.21)
ViewSpatial	Qwen3-VL-4B	30	1.00	37.29 (+0.83)	37.60 (+7.94)	35.48 (-40.32)
	Molmo2-4B	30	0.50	35.75 (+1.31)	31.06 (+1.67)	62.90 (-0.81)
Controlled → 3DSR	Qwen3-VL-4B	35	1.00	46.96 (+2.24)	32.99 (+7.70)	70.70 (-7.03)
	Molmo2-4B	35	1.00	49.64 (+6.22)	39.66 (+13.56)	66.60 (-6.25)
Controlled → ViewSpatial	Qwen3-VL-4B	30	1.00	42.99 (+6.53)	36.21 (+6.55)	82.26 (+6.45)
	Molmo2-4B	30	0.25	34.32 (-0.12)	29.11 (-0.28)	64.52 (+0.81)