GEOMETRY-GUIDED CROSS-VIEW REASONING IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view spatial reasoning requires MLLMs to associate observations of the same physical scene region across viewpoints. Yet existing geometry-aware MLLMs mainly enrich individual visual tokens with 3D or camera information, leaving such cross-view correspondences to emerge implicitly from appearance-driven attention. We introduce GeoMLLM, a geometry-guided MLLM that grounds cross-view information exchange in camera geometry. GeoMLLM equips each visual patch with a camera-ray representation and uses pairwise ray and epipolar cues to constrain which tokens can plausibly correspond across views. During training, depth-derived soft correspondences further supervise cross-view attention, teaching the model to route information between observations of the same physical scene region. Built on Qwen3-VL-4B, GeoMLLM achieves state-of-the-art performance on SPAR-Bench, reaching $\mathbf{83.79}$ overall, with $\mathbf{83.49}$ on multi-view tasks ($\mathbf{+20.86}$ points) and $\mathbf{88.13}$ on Cross-view Perception ($\mathbf{+25.12}$ points). GeoMLLM also consistently improves over its matched backbone across additional multi-image and 3D spatial reasoning benchmarks. These results demonstrate that explicitly modeling camera geometry in cross-view token interactions provides a strong inductive bias for multi-view spatial reasoning.

1 INTRODUCTION

Multimodal large language models (MLLMs) have extended language models from text-only reasoning to visual understanding by connecting pretrained vision encoders with large language models (Alayrac et al., 2022; Li et al., 2023; Bai et al., 2025; Team, 2025). Recent work further equips MLLMs with spatial and geometric knowledge through 3D representations, spatial instruction tuning, geometry-aware visual features, and camera-conditioned representations (Hong et al., 2023; Chen et al., 2024a; Miao et al., 2025; Yang et al., 2025; Zhang et al., 2026a). Yet multi-view reasoning requires more than representing geometry within individual observations: before information can be integrated across images, the model must determine which observations correspond to the same physical scene region. This association is inherently ambiguous because camera motion changes image location, scale, and appearance, occlusion limits visibility across views, and repeated structures create visually similar but geometrically distinct candidates. Robust multi-view reasoning therefore depends critically on resolving cross-view correspondences. As illustrated in Figure 1, camera geometry provides a principled signal for disambiguating such visually ambiguous matches.

Recent geometry-aware MLLMs have primarily introduced geometric information at the level of *token representation*. One line of work employs pretrained geometry encoders, such as VGGT (Wang et al., 2025), to extract multi-view geometric features and fuse them with conventional visual representations (Zheng et al., 2026; Zhao et al., 2025). Another explicitly encodes camera parameters into visual tokens; for example, Camera-Aware MLLM conditions each visual token on a dense camera-ray representation derived from camera intrinsics (Zhang et al., 2026a). These approaches provide useful geometric context for each observation, but do not directly model whether a particular pair of tokens across views depicts the same physical scene region. Once the tokens enter a Transformer, their interactions are still typically determined by learned query–key similarity (Vaswani et al., 2017). Geometry encoded in the token features may influence this similarity indirectly, yet cross-view association remains implicit and can be unreliable when viewpoint changes weaken appearance similarity or when different regions share similar semantics.

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) Implicit appearance-based association. (b) Geometry-guided cross-view correspondence.

Figure 1: **Motivation: resolving ambiguous cross-view correspondence.** Appearance-driven association can confuse visually similar candidate regions across viewpoints, whereas camera geometry helps identify the observation of the same scene region. Dashed arrows depict cross-view associations; colored outlines distinguish the source, ambiguous candidates, and selected correspondence.

Building on this view, we introduce *GeoMLLM*, a geometry-guided MLLM that explicitly promotes geometry-consistent correspondence across views. GeoMLLM associates each visual patch with a camera-ray embedding that specifies where and how it is observed in a shared reference frame. It then uses *Camera Geometry-aware Attention* to modulate query–key interactions according to pairwise ray and epipolar compatibility, suppressing cross-view connections that are inconsistent with the calibrated cameras while retaining content-based attention. *Geometry-aware correspondence supervision* back-projects source patches with depth and reprojects them into other views to construct soft attention targets. The two mechanisms play complementary roles: Camera Geometry-aware Attention determines where a correspondence can plausibly occur, while correspondence supervision teaches the model which plausible observation corresponds to the same physical scene region. Depth is used only to construct the training signal and is not required at inference.

The largest gains of GeoMLLM occur precisely where cross-view association matters most. Built on Qwen2.5-VL-3B (Bai et al., 2025), GeoMLLM achieves $\mathbf{82.18}$ overall on SPAR-Bench (Zhang et al., 2026b) ($\mathbf{+54.03}$ over the matched backbone; $\mathbf{+13.83}$ over the strongest prior spatially specialized model). The gains are particularly pronounced on tasks that require cross-view reasoning, reaching $\mathbf{82.38}$ on multi-view tasks ($\mathbf{+19.75}$) and $\mathbf{86.37}$ on Cross-view Perception ($\mathbf{+23.36}$). The improvements transfer to MMSI-Bench (Yang et al., 2026), ViewSpatial-Bench (Li et al., 2025), and CV-Bench (Tong et al., 2024), with especially strong gains on multi-image and 3D reasoning. Ablations further show complementary contributions from camera-ray conditioning, Camera Geometry Attention, and correspondence supervision, while camera-corruption interventions confirm that GeoMLLM meaningfully uses the provided calibration.

Our contributions are threefold:

- We identify reliable cross-view correspondence as a key bottleneck in multi-view MLLMs. Existing geometry-aware approaches primarily enrich individual token representations, leaving pairwise associations between observations of the same physical scene region only implicitly modeled.
- We introduce *Camera Geometry-aware Attention*, a geometry-conditioned visual connector that augments patch tokens with camera-ray representations and injects pairwise ray and epipolar compatibility directly into multi-head attention logits.
- We propose geometry-aware correspondence supervision that aligns cross-view attention with soft correspondences during training. Together with Camera Geometry-aware Attention, it promotes geometry-consistent information routing and substantially improves cross-view spatial reasoning.

2 Related Work

Multimodal Large Language Models (MLLMs). MLLMs extend the reasoning and language capabilities of large language models to visual modalities by connecting pretrained visual encoders with language models (Alayrac et al., 2022; Li et al., 2022; 2023; Liu et al., 2023a; 2024). They have demonstrated strong performance on image captioning, visual question answering, visual grounding, and general multimodal reasoning (Peng et al., 2024; Zhang et al., 2024a). Recent work further expands MLLMs to high-resolution images and long videos through dynamic visual tokenization, temporal modeling, and large-scale multimodal instruction tuning (Zhang et al., 2023; Lin et al.,

2024; Bai et al., 2025; Team, 2025). These advances establish visual tokens and attention-based interaction as a general interface between visual observations and language reasoning.

MLLMs for Spatial Intelligence. Recent work extends MLLMs from semantic perception to spatial understanding of 3D environments. Early approaches directly connect language models with 3D representations, including point clouds, object-centric scene features, and reconstructed geometry, for 3D question answering, grounding, captioning, and planning (Hong et al., 2023; Xu et al., 2024; Chen et al., 2024b; Huang et al., 2024; Zhu et al., 2025). To exploit scalable image and video data, studies construct spatial instruction data, lift 2D observations into 3D supervision, or distill geometric priors from pretrained vision models (Chen et al., 2024a; Miao et al., 2025; Zhang et al., 2026b; Wu et al., 2026). More recent efforts consider dynamic 4D scenes and cross-camera generalization, highlighting the importance of temporal geometry, camera parameters, and viewing rays for robust spatial reasoning (Zheng et al., 2026; Zhou et al., 2026; Chen et al., 2026; Zhang et al., 2026a). While these methods mainly improve the geometric information available to visual tokens, we study how camera geometry should govern the interaction between tokens across views.

Cross-view Correspondence and Camera Geometry. Camera intrinsics and extrinsics provide geometric constraints for relating observations across views. Multi-view perception methods encode projected 3D positions, camera coordinates, or view transformations into visual features and object queries (Liu et al., 2022; 2023b; Xiong et al., 2023; Li et al., 2024). Geometry-aware Transformers incorporate relative camera transformations into attention or positional encoding (Miyato et al., 2024; Li et al., 2026), while correspondence methods solve cross-image matching by combining feature similarity with epipolar constraints or learned matching (He et al., 2020; Sarlin et al., 2020; Sun et al., 2021; Jiang et al., 2021). These methods establish the value of geometry for matching and multi-view perception, but do not address geometry-grounded information routing inside a general-purpose MLLM. Conversely, the geometry-aware MLLMs discussed above mainly enrich individual token representations, leaving pairwise cross-view associations implicit. GeoMLLM bridges this gap by formulating connector attention as a pose-conditioned cross-view relation: we inject pairwise ray geometry into the attention logits and supervise the resulting distributions with depth-induced correspondences, allowing geometric matching to support language-based spatial reasoning.

3 METHOD

Given K multi-view images with known or estimated calibration, a vision encoder extracts P patch tokens from each view, yielding $\boldsymbol{X} \in \mathbb{R}^{K\times P\times C}$. We use *cross-view correspondence* to denote a patch-level association between observations of the same physical region. Our central premise is that cross-view information exchange should also depend explicitly on the pairwise camera geometry that constrains which tokens can plausibly correspond.

As illustrated in Figure 2, GeoMLLM addresses this correspondence problem through three complementary components. First, a *camera-ray embedding* conditions each visual token on the ray along which it observes the scene (Section 3.1). Second, *Camera Geometry-Aware Attention* uses pairwise ray and epipolar geometry to constrain which token interactions are geometrically plausible across views (Section 3.2). Third, *geometry-aware correspondence supervision* trains the aggregate cross-view attention distribution to concentrate on depth-derived correspondences (Section 3.3). The first two components constitute the inference-time connector; the correspondence objective is used only during training. This auxiliary geometry stream is separate from the proposed attention mechanism; its role, together with the remaining geometric constructions, is detailed in Appendix C.

3.1 CAMERA-RAY EMBEDDING

Let $\tilde{\boldsymbol{u}}_p = (u_p, v_p, 1)^\top$ be the center of patch p, and let $\boldsymbol{T}_k^{c2w} = [\boldsymbol{R}_k \mid \boldsymbol{t}_k]$ and $\boldsymbol{K}_k$ denote the pose and intrinsics of view k. We express its ray direction and camera origin in the coordinate system of the first view. The per-token geometric feature $\boldsymbol{g}_{k,p} \in \mathbb{R}^{15}$ concatenates camera- and reference-frame ray directions, the Plücker moment, normalized patch coordinates, and normalized intrinsics. This representation captures both the ray orientation and the camera baseline without depending on the choice of world frame.

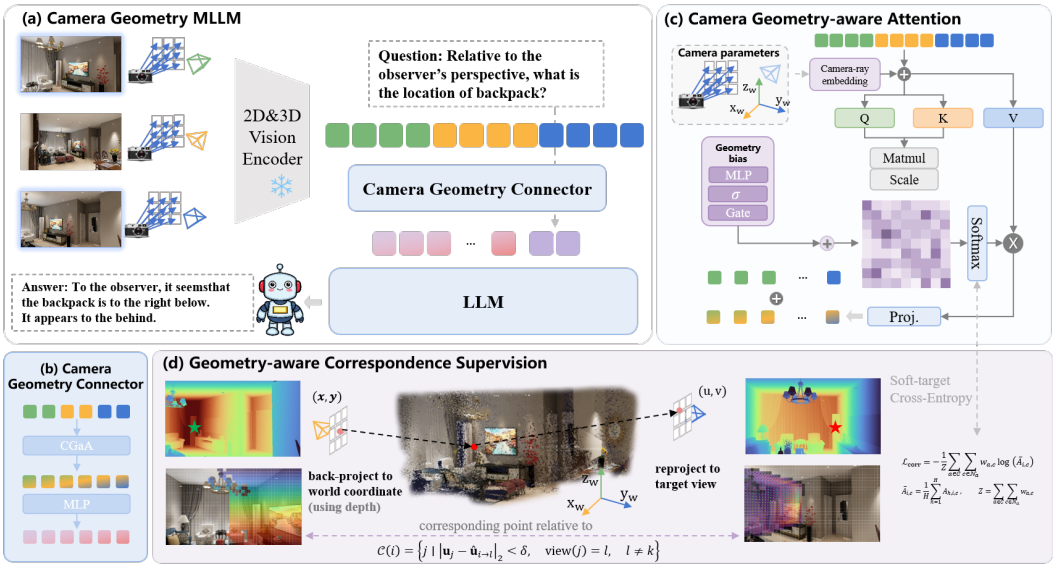


Figure 2: **Overview of GeoMLLM.** Given calibrated multi-view images, GeoMLLM promotes patch-level correspondence between observations of the same physical scene region. Each visual token is augmented with a camera-ray representation in a shared reference frame, while pairwise ray and epipolar geometry constrain geometrically plausible cross-view interactions. During training, depth-induced soft correspondences supervise the head-averaged attention maps to distinguish physical matches within this constrained space.

We apply Fourier encoding to $\boldsymbol{g}_{k,p}$ and map the result to a C-dimensional ray embedding $\boldsymbol{e}^{\text{ray}}_{k,p}$. The embedding is injected into the visual token through a gated residual connection:

$$\boldsymbol{x}'_{k,p} = \boldsymbol{x}_{k,p} + \alpha\, \sigma\Big(\text{MLP}_{\text{gate}}\big[\boldsymbol{x}_{k,p}, \boldsymbol{e}^{\text{ray}}_{k,p}\big]\Big) \odot \boldsymbol{e}^{\text{ray}}_{k,p}, \tag{1}$$

where the gate adaptively controls the contribution of camera geometry to each visual token, and α determines its overall magnitude. This formulation preserves the appearance representation while conditioning each token on its viewing geometry.

3.2 Camera Geometry-Aware Attention

Camera-ray embeddings characterize how individual patches are observed, but cross-view correspondence is inherently a relation between token pairs. After flattening the view and patch dimensions, we therefore construct a geometric descriptor $\boldsymbol{\phi}_{ij}$ for every ordered token pair (i, j). The descriptor comprises three complementary groups of cues: relative ray incidence and camera baseline; the closest-approach geometry of the two viewing lines, including front-of-camera validity; and epipolar consistency in the image plane. Additional indicators distinguish same-view, cross-view, and geometrically degenerate pairs. Since $\boldsymbol{\phi}_{ij}$ depends only on the camera configuration, it is computed once and shared across all connector layers.

We convert $\boldsymbol{\phi}_{ij}$ into a per-head additive attention bias combining an interpretable epipolar penalty with a learned residual:

$$B^h_{ij} = \eta\left[-\lambda\, \psi(r_{ij}) v_{ij} + g^h(\boldsymbol{\phi}_{ij})\, \Delta^h(\boldsymbol{\phi}_{ij})\right], \tag{2}$$

where $r_{ij} = \min(\bar{e}_{ij}/\sigma, r_{\max})$ is the clipped, patch-normalized epipolar distance, $\psi(r) = \log(1 + r^2)$, and v_{ij} restricts the epipolar penalty to valid cross-view pairs. The first term explicitly suppresses interactions that violate epipolar geometry. The second term complements this fixed constraint with a gated, head-specific residual, allowing different heads to capture additional compatibility patterns from the pairwise ray geometry. Unlike the epipolar penalty, the learned residual is defined for both same- and cross-view pairs.

The resulting attention probabilities are

$$A^h_{ij} = \text{softmax}_j\Big((\boldsymbol{q}^h_i)^\top \boldsymbol{k}^h_j / \sqrt{d_h} + B^h_{ij}\Big). \tag{3}$$

This module retains content-based attention while making token interactions explicitly sensitive to camera geometry. Visual similarity identifies semantically plausible candidates, whereas the geometric bias discourages interactions that are inconsistent with the observed views.

3.3 Geometry-Aware Correspondence Supervision

Camera geometry restricts where a correspondence may occur, but does not determine which location observes the same 3D point. We therefore complement the geometric attention bias with direct correspondence supervision. Using the available training-time depth, we back-project each source patch into 3D and project the resulting point into every other view. We discard same-view, out-of-frame, behind-camera, and depth-inconsistent projections.

A continuous projection generally falls between patch centers. Rather than quantizing it to a single token, we distribute its target probability mass bilinearly among the valid neighboring tokens. Let $\mathcal{C}$ denote the set of valid source–target-view pairs. For a pair $a = (i, k') \in \mathcal{C}$, let $\mathcal{N}_a$ be the neighboring tokens in target view k', and let $w_{a,c}$ be their normalized bilinear weights.

For the final connector layer, we define the head-averaged attention and normalization factor as

$$\bar{A}_{i,c} = \frac{1}{H} \sum_{h=1}^{H} A_{h,i,c}, \qquad Z = \sum_{a \in \mathcal{C}} \sum_{c \in \mathcal{N}_a} w_{a,c}. \tag{4}$$

The correspondence loss is then

$$\mathcal{L}_{\text{corr}} = -\frac{1}{Z} \sum_{a \in \mathcal{C}} \sum_{c \in \mathcal{N}_a} w_{a,c} \log\big(\bar{A}_{i,c} + \epsilon\big)\,. \tag{5}$$

Supervising the head-averaged distribution allows individual heads to retain complementary interaction patterns while encouraging their aggregate attention to recover the physical correspondence. The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \beta_{\text{corr}} \mathcal{L}_{\text{corr}}. \tag{6}$$

This objective turns depth into a training signal for cross-view information routing, encouraging the connector to attend to patches that observe the same physical scene region rather than merely those with similar appearance. The depth maps and correspondence targets are discarded after training.

4 Experiments

We evaluate GeoMLLM along three complementary axes: (i) whether explicitly modeling camera geometry improves multi-view spatial reasoning, (ii) whether these improvements transfer to external benchmarks with distinct data sources and task formulations, and (iii) whether the proposed components contribute as intended to geometry-aware cross-view interaction.

4.1 Experimental Setup

Benchmarks. We use **SPAR-Bench** (Zhang et al., 2026b) as our primary benchmark, reporting the full split for the main comparison and the tiny split for comparison with prior work. Its coverage of depth and distance estimation, camera motion, viewpoint transformation, object relations, and spatial imagination directly tests geometry-aware multi-view reasoning. To assess cross-benchmark transfer, we further evaluate on **MMSI-Bench** (Yang et al., 2026) for general multi-image spatial reasoning, **ViewSpatial-Bench** (Li et al., 2025) for multi-view localization and reasoning, and **CV-Bench** (Tong et al., 2024) for single-image 2D and 3D spatial understanding, together distinguishing multi-view-specific gains from broader improvements in spatial understanding.

Implementation details. GeoMLLM is built on VG-LLM (Zheng et al., 2026) and follows its training-data configuration using SPAR-7M-RGBD (Zhang et al., 2026b) and LLaVA-Hound split of LLaVA-Video-178k (Zhang et al., 2024b). We freeze the visual encoder and 3D encoder and optimize the remaining model parameters jointly with the proposed geometry-aware modules. Training is performed for one epoch using AdamW with a global batch size of 64 and a learning rate of 10^{-5}. The correspondence loss is applied to the attention maps averaged across heads in the final connector layer, with $\beta_{\text{corr}} = 0.1$. Additional details are provided in Appendix B.

Table 1: Comparison of MLLMs' spatial reasoning performance on SPAR-Bench. We report the official aggregate scores and task-level results on both the tiny and full splits. MHSA replaces our Camera Geometry-Aware Attention with standard multi-head self-attention while retaining the remaining model and training configuration.

Method	Avg.	Low	Depth-OC	Depth-OC-MV	Depth-OO	Depth-OO-MV	Dist-OC	Dist-OC-MV	Dist-OO	Dist-OO-MV	Medium	PosMatch	CamMotion	ViewChg1	High	Dist1-OO	Dist1-OO-MV	ObjRel-OC-MV	ObjRel-OO	ObjRel-OO-MV	Spatmg-OC	Spatmg-OC-MV	Spatmg-OO	Spatmg-OO-MV
SPAR-Bench *(tiny)*																								
Human Level	67.27	55.31	72.75	74.25	28.75	36.25	78.25	52.25	66.50	33.50	72.32	92.00	64.00	60.97	76.22	80.00	94.00	70.00	92.00	80.00	78.00	82.00	50.00	60.00
GPT-4o	36.39	29.25	53.80	45.00	15.00	13.60	37.40	34.40	23.40	24.40	24.93	30.00	16.00	28.80	45.11	64.00	64.00	58.00	46.00	46.00	32.00	44.00	30.00	22.00
Claude-3.7-Sonnet	21.77	25.43	41.00	45.40	11.20	12.20	42.60	19.60	26.00	5.40	7.33	16.00	6.00	0.00	23.33	40.00	48.00	22.00	36.00	14.00	12.00	20.00	6.00	12.00
Qwen2-VL-72B	35.62	35.28	45.40	49.80	13.80	10.00	54.60	49.40	36.80	22.40	23.39	42.00	18.00	10.16	40.00	60.00	68.00	50.00	38.00	44.00	18.00	28.00	18.00	36.00
Qwen2.5-VL-72B	39.40	35.35	53.20	46.80	17.80	29.00	49.60	57.40	14.40	14.60	23.05	40.00	16.00	13.16	48.44	74.00	74.00	60.00	56.00	50.00	20.00	34.00	24.00	44.00
SPAR-Bench *(full)*																								
Generalist MLLMs																								
GPT-5	37.40	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Gemini-2.5-Pro	36.30	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Qwen2.5-VL-3B	28.15	23.65	29.06	32.31	20.59	14.97	19.19	19.70	26.97	26.40	26.34	30.79	36.75	11.47	32.77	56.18	58.33	35.75	31.59	19.94	20.97	26.16	23.84	22.13
Qwen3-VL-4B	34.94	29.14	31.42	40.97	19.27	20.38	10.66	18.63	50.74	41.08	32.86	52.16	28.75	17.68	40.78	65.29	58.93	45.75	52.75	31.58	26.88	32.85	25.83	27.17
InternVL2.5-8B	36.28	29.46	25.78	29.31	23.79	18.76	46.82	42.68	22.62	25.89	31.88	61.32	28.00	6.32	43.80	59.71	56.85	51.75	44.23	41.55	36.56	41.57	22.52	39.50
InternVL3-8B	33.80	23.85	26.47	26.75	17.66	18.82	14.66	13.32	41.85	31.25	38.62	59.54	38.25	18.06	41.04	65.29	62.80	40.00	52.20	32.13	25.27	37.79	22.19	31.65
LLaVA-OneVision-7B	31.20	21.79	30.33	26.94	18.58	13.87	10.43	13.64	31.24	29.29	26.13	38.68	30.25	9.47	40.14	56.47	55.06	37.25	48.63	38.23	30.38	33.72	26.49	35.01
Spatially Specialized MLLMs																								
VG-LLM-4B	60.36	52.81	83.97	44.75	34.27	16.96	79.08	60.63	66.20	36.63	51.35	39.69	63.00	25.82	72.92	89.12	63.69	81.25	83.79	77.84	68.82	60.17	59.60	71.99
SPAR-8B	63.25	65.53	81.53	79.22	38.25	35.51	78.93	79.18	68.13	63.50	63.01	78.88	73.00	37.14	61.29	86.47	79.76	64.00	69.23	59.00	47.31	50.00	42.38	53.50
CA-MLLM-4B	68.35	59.94	89.61	49.12	**57.07**	19.27	88.02	63.40	77.60	35.42	60.42	47.84	73.00	31.01	81.74	**92.35**	68.75	87.50	91.76	84.21	79.57	66.86	**83.44**	81.23
Ours$_{\text{Qwen2.5-3B}}$ (MHSA)	74.88	65.55	84.31	78.88	40.59	38.84	76.84	80.38	66.89	57.68	73.07	86.77	87.75	44.69	83.78	88.82	84.23	90.75	87.36	91.41	73.92	71.22	77.48	88.80
Ours$_{\text{Qwen2.5-3B}}$	82.18	75.18	88.56	86.69	52.63	**51.51**	86.82	86.79	75.23	73.23	86.37	91.35	**96.25**	71.51	87.01	91.18	**91.07**	94.75	90.93	**92.80**	75.00	73.26	82.78	91.32
Ours$_{\text{Qwen3-4B}}$	**83.79**	**76.70**	**90.00**	**87.44**	55.46	51.26	**88.37**	**88.49**	**78.10**	**74.48**	**88.13**	**92.37**	95.75	**76.27**	**88.65**	**92.35**	**91.07**	**95.75**	**93.13**	**92.80**	**81.18**	**78.20**	81.46	**91.88**

• Tasks are divided into low-, medium-, and high-level reasoning categories following the official SPAR-Bench partition. "MV" denotes multi-view evaluation, and "–" denotes unavailable or unreported results.

4.2 Main Results on Spatial Reasoning

GeoMLLM substantially improves SPAR-Bench performance. GeoMLLM achieves the best overall results on the SPAR-Bench full split with both the Qwen2.5-3B and Qwen3-4B backbones (Table 1). Relative to the corresponding backbone models, GeoMLLM improves by **+54.03** and **+48.85** points, respectively; it also surpasses the strongest prior spatially specialized model, CA-MLLM-4B, by **+13.83** and **+15.44** points. The improvement is not confined to one level of the benchmark: with the Qwen2.5-3B backbone, GeoMLLM gains **+51.53**, **+60.03**, and **+54.24** points on low-, medium-, and high-level reasoning. Under the same backbone and training configuration, GeoMLLM improves over the matched MHSA (Vaswani et al., 2017) configuration by **+7.30** points, isolating the contribution of geometry-conditioned interaction.

The gains transfer most strongly to multi-image and 3D reasoning. On MMSI-Bench, GeoMLLM improves over the corresponding Qwen2.5-3B and Qwen3-4B backbone models by **+7.46** and **+4.49** points, respectively (Table 2). With the Qwen2.5-3B backbone, the largest category-level gains occur in attribute (**+16.98**) and motion reasoning (**+8.91**). On ViewSpatial-Bench, GeoMLLM improves by **+8.14** and **+7.10** points with the two backbones. These gains are concentrated on person-centric reasoning (**+16.67** and **+10.86**), whereas camera-centric performance changes by **-4.66** and **+1.46** points, respectively. On CV-Bench, GeoMLLM with the Qwen2.5-3B backbone improves by **+11.25** points overall, with a larger gain on 3D (**+20.42**) than on 2D questions (**+2.08**). With the Qwen3-4B backbone, the transfer pattern is mixed: the 3D gain (**+2.42**) is offset by a 2D decrease (**-7.44**), resulting in an overall change of **-2.51** points. GeoMLLM with the Qwen2.5-3B backbone also improves over the matched MHSA configuration by **+4.00**, **+2.69**, and **+1.08** points on MMSI-Bench, ViewSpatial-Bench, and CV-Bench, respectively. Thus, the gains extend beyond SPAR-Bench but are not uniform: they are strongest for multi-image, 3D, and person-centric reasoning, while camera-centric localization and 2D recognition remain less consistent. Full task-level results are provided in Appendix D.

4.3 Analyzing Geometry-Aware Reasoning

Aggregate results alone do not reveal where GeoMLLM's improvements arise or what trade-offs they entail. We therefore analyze the method from four complementary perspectives: where the gains concentrate across task regimes, how performance responds to controlled camera-geometry perturbations, how each geometry-aware component contributes under matched settings, and how

Table 2: Category-level accuracy (%) on three spatial reasoning benchmarks. Category scores are macro-averaged over their corresponding sub-tasks. Avg. is the unweighted mean of the three benchmark overall scores. MHSA replaces our Camera Geometry-Aware Attention with standard multi-head self-attention while retaining the remaining model and training configuration. Best and second-best results are shown in **bold** and underline, respectively.

Methods	Avg.	MMSI-Bench					ViewSpatial-Bench			CV-Bench		
		Attr.	MSR	Motion	Relation	Overall	Camera	Person	Overall	2D	3D	Overall
Open-Source Models												
Qwen2.5-3B	43.90	20.00	21.21	20.50	28.24	24.70	38.10	34.05	35.67	71.56	71.08	71.32
Qwen2.5-7B	48.05	21.54	26.26	24.59	29.94	27.11	39.09	35.62	37.01	76.56	83.50	80.03
Qwen3-4B	49.85	30.18	22.73	28.02	25.70	26.67	38.41	36.49	37.26	78.65	92.58	85.62
Qwen3-8B	51.78	27.82	29.29	25.98	**32.93**	30.40	**42.26**	36.72	38.94	**80.11**	91.92	**86.01**
Open-Source Spatial Intelligence Models												
VG-LLM-4B	50.69	30.02	24.75	20.59	31.46	28.61	31.77	49.42	42.36	73.23	89.00	81.11
CA-MLLM-4B	47.14	26.16	26.26	17.96	30.87	27.24	34.30	42.59	39.27	59.39	90.42	74.90
Ours$_{\text{Qwen2.5-3B}}$ (MHSA)	50.26	23.08	25.76	21.32	32.54	28.16	32.05	47.16	41.12	72.74	90.25	81.49
Ours$_{\text{Qwen2.5-3B}}$	52.85	**36.98**	27.27	29.41	32.29	**32.16**	33.44	**50.72**	43.81	73.64	91.50	82.57
Ours$_{\text{Qwen3-4B}}$	**52.88**	29.29	**35.86**	**38.80**	28.45	31.16	39.87	47.35	**44.36**	71.21	**95.00**	83.11

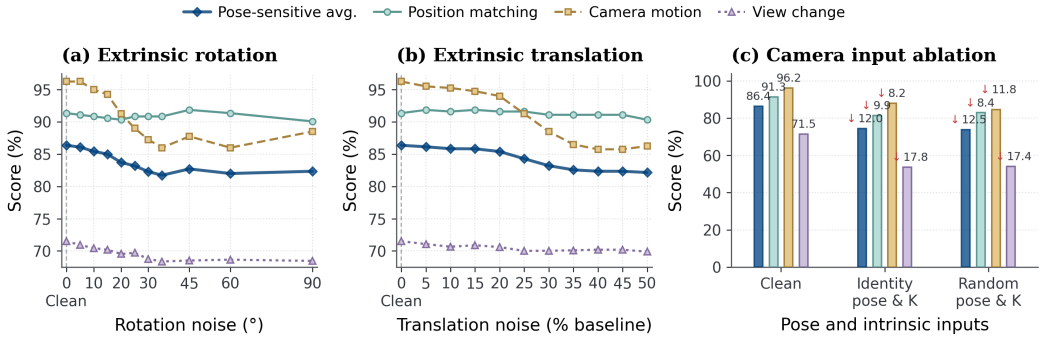


Figure 3: **Perturbation analysis on SPAR-Bench.** Panels (a) and (b) jointly perturb all camera extrinsics while keeping the images and intrinsics fixed. Panel (c) replaces both poses and intrinsics with identity or random matrices. The black series is the macro average of position matching, camera-motion inference, and viewpoint-change inference; the colored series show the three components. Red annotations in panel (c) denote the decrease from clean performance.

the resulting performance compares with inference cost. Together, these analyses examine whether the observed gains are consistent with the intended cross-view mechanism and characterize the contributions and computational overhead of its design choices.

The gains concentrate on cross-view tasks. Relative to its matched Qwen2.5-3B backbone, GeoMLLM gains **+60.56** points on position matching, **+59.50** points on camera-motion inference, and **+60.04** points on viewpoint-change inference. The same pattern holds for multi-view depth and distance reasoning, with gains of **+54.38** points on Depth-OC-MV and **+67.09** points on Dist-OC-MV.

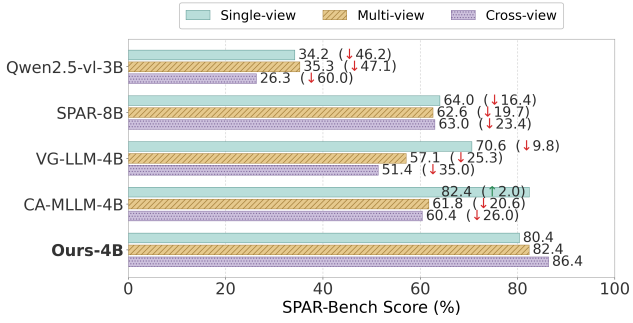


Figure 4: **SPAR-Bench task-group comparison.** Parentheses show within-group differences from Ours-4B.

Figure 4 provides a complementary grouped comparison. GeoMLLM exceeds SPAR-8B by **+19.75** points on multi-view tasks and by **+23.36** points on Cross-view Perception. From the single-view to the multi-view group, its performance changes by **+1.99** points, compared with **-13.51** for VG-LLM-4B and **-20.68** for CA-MLLM-4B. This comparison is descriptive rather than causal because the groups contain different tasks, but it consistently localizes the largest gains to cross-view association.

Performance depends on valid camera geometry. Figure 3 evaluates this dependence by keeping the images fixed and perturbing only the camera inputs. For pose-only interventions, we jointly apply fixed-magnitude local-axis rotations or translations to all camera extrinsics, with translation magnitudes normalized by the median non-zero inter-camera baseline. We further consider stronger interventions that replace the full camera specification with identity or random matrices.

Pose-only corruption produces moderate degradation. At the strongest rotation and translation perturbations, the pose-sensitive average changes by -**4.02** and -**4.21** points, while camera-motion inference changes by -**7.75** and -**10.00** points. This robustness is consistent with the partial invariance of pairwise geometric relations under joint perturbations and with content-based attention providing a fallback. Replacing the full camera specification has a substantially larger effect: the pose-sensitive average and viewpoint-change inference change by approximately -**12** and -**17** points, respectively. Identity and random replacements produce similar degradation, indicating that performance depends on preserving valid camera geometry rather than on the particular corruption used.

The geometry components provide complementary gains. Table 3 compares configurations that share the same two-block connector, initialization, training data, and optimization schedule. Combining the camera-ray embedding with Camera Geometry-Aware Attention improves the SPAR-Bench average by **+10.60** points over ray embedding alone and by **+4.15** points over geometry-aware attention alone. Their complementarity is most pronounced on viewpoint-change inference, where the corresponding gains are **+23.12** and **+25.14** points. Adding geometry-aware correspondence supervision on top of both components yields a further **+2.85** points overall and consistent improvements in position matching (**+3.82**), camera-motion inference (**+0.75**), viewpoint-change inference (**+3.06**), and multi-view reasoning (**+2.13**). The complete GeoMLLM configuration exceeds VG-LLM-4B by **+21.82** points overall; because VG-LLM-4B is an external baseline rather than a matched zero-component configuration, this final gap should not be interpreted as the joint causal effect of the three components.

Table 3: Component ablation on SPAR-Bench. We report the overall performance together with representative cross-view and camera-motion metrics.

Method Components			SPAR-Bench				
Ray Emb.	CGA Attn.	Corr.	Avg.	Pos.	Cam.	View	MV
✗	✗	✗	60.36	39.69	63.00	25.82	57.10
✓	✗	✗	68.73	69.72	88.00	45.33	71.06
✗	✓	✗	75.18	87.02	79.75	43.31	76.91
✓	✓	✗	79.33	87.53	95.50	68.45	80.25
✓	✓	✓	**82.18**	**91.35**	**96.25**	**71.51**	**82.38**

Ray Emb., **CGA Attn.**, and **Corr.** denote our three geometry components; Pos., Cam., View, and MV denote position matching, camera-motion inference, viewpoint-change inference, and the multi-view average. The component-free row reports VG-LLM-4B, not a matched ablation.

Performance gains come with additional inference cost. Figure 6 compares per-case inference latency against aggregate spatial-reasoning performance, measured as the mean of the overall scores on SPAR-Bench, MMSI-Bench, ViewSpatial-Bench, and CV-Bench. With the Qwen2.5-3B backbone, GeoMLLM achieves the highest aggregate performance among the compared models, improving over the matched backbone by **+20.22** points while adding **+1.00** seconds per case. Relative to the spatially specialized VG-LLM-4B and CA-MLLM-4B, GeoMLLM gains **+7.07** and **+7.74** points with latency increases of only **+0.12** and **+0.08** seconds per case, respectively. GeoMLLM therefore occupies the high-performance end of the trade-off, but its improvement over the base backbone is not computationally free.

4.4 Qualitative Analysis

Depth-derived supervision yields geometrically coherent targets. Figure 5 visualizes soft correspondence targets constructed from depth and calibrated camera geometry. Matching colors trace source patches to their projected locations on the corresponding scene regions in each target view, while invalid, out-of-frame, and depth-inconsistent projections remain blank. Bilinear splatting distributes each valid projection over neighboring target patches, producing spatially localized supervision rather than a single hard assignment. Because these targets are constructed directly from geometry, the visualization validates the supervision pipeline.

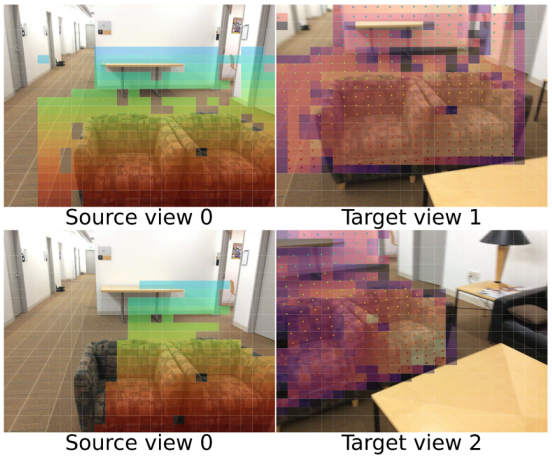


Figure 5: Geometry-aware correspondence targets.

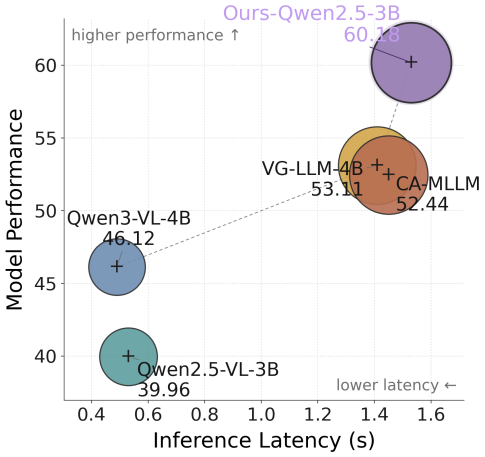
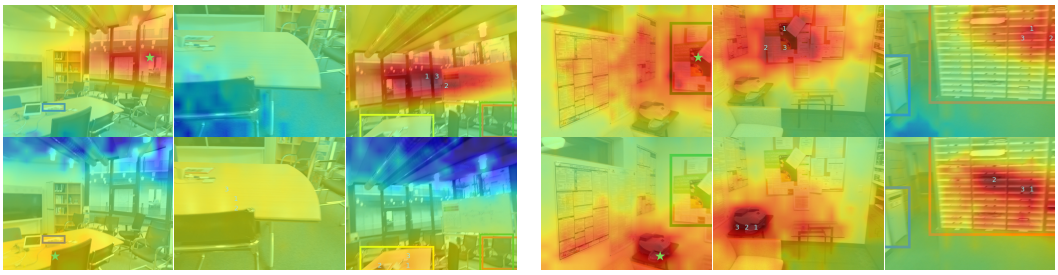


Figure 6: Performance–latency analysis.



Question: What is the initial relation of the window (red) to the table (yellow)? How does it change after moving to the chair (green) and facing the book (blue)?

Prediction: B: right, closer, front, front.

Question: What is the relation of the mailboxes (red) to the observer's initial position? How does it change after moving to the bulletin board (green) and facing the mailbox (blue)?

Prediction: A: left, front.

Figure 7: **Cross-view attention and spatial-reasoning examples on SPAR-Bench.** Stars mark source patches, heatmaps show connector attention in the other views, and numbered markers indicate the three highest-scoring target patches. The boxes report the corresponding questions and model predictions.

Learned attention concentrates on cross-view correspondences. Figure 7 visualizes connector attention for star-marked source patches in two SPAR-Bench examples. Despite changes in viewpoint and scale, high-attention regions and the three highest-scoring target patches cluster around corresponding scene structures in the other views. Notably, the attended regions shift to viewpoint-dependent image locations rather than remaining near the source patch's 2D coordinates, consistent with geometry-conditioned information routing. The question–prediction pairs connect these interactions to the final spatial judgments. Although qualitative, these patterns align with the quantitative gains on position matching and viewpoint-change inference.

5 Conclusion

Multi-view spatial reasoning requires more than geometry-aware visual representations: it requires models to explicitly determine how observations across views correspond to the same physical scene, despite substantial viewpoint and appearance changes. We introduced *GeoMLLM*, which addresses this interaction gap by grounding cross-view token exchange in camera geometry through ray-based representations, geometry-conditioned attention, and correspondence supervision. Across SPAR-Bench and additional spatial reasoning benchmarks, GeoMLLM consistently improves multi-view and cross-view reasoning, with ablations confirming that its components provide complementary benefits. Camera-corruption experiments further show that these gains arise from actively exploiting valid geometric calibration rather than from additional features alone. More broadly, our results suggest that explicit geometric structure is most useful when it governs *how* information is exchanged across views, rather than only *what* information each visual token represents.

LLM USAGE DISCLOSURE

Large Language Models (LLMs) were used as general-purpose assistance tools during the preparation of this manuscript. Specifically, LLMs were used to assist with drafting and editing portions of the manuscript, improve language clarity and overall readability, and provide assistance with software code, including implementation, debugging, and code refinement. All research ideas, methodological decisions, experimental design, analysis, and experimental results originated from the authors. All AI-assisted text and code were carefully reviewed, verified, and, where necessary, revised by the authors to ensure technical accuracy and consistency with the intended methodology. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide comprehensive details to facilitate the reproduction and verification of GeoMLLM. The model architecture and geometry-aware modules are described in Sections 3 and Appendix C. Details of the training data, benchmark splits, preprocessing and camera/depth handling, optimization schedule, hardware, random seed, and evaluation protocol are provided in Section 3 and Appendix B. We train on SPAR-7M-RGBD and the LLaVA-Hound split of LLaVA-Video-178K, and evaluate on SPAR-Bench, MMSI-Bench, ViewSpatial-Bench, CV-Bench, and VSI-Bench using their official evaluation protocols. Code and scripts for reproducing the reported experiments are provided in the accompanying anonymous repository.

ETHICS STATEMENT

This work studies geometry-aware multimodal language models for multi-view spatial reasoning in a controlled research setting. Our experiments use publicly available benchmark datasets, annotations, and pretrained models, and do not involve human participants or the collection of new personal data. We follow the terms and licenses of the data and model providers. The study focuses on improving geometric reasoning and supporting the evaluation of multimodal models; no identity recognition, biometric analysis, or sensitive-attribute inference is performed.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26428–26438, 2024b.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xiang An, Bo Li, Xin Xie, ZiDong Wang, Mingze Sun, Shuang Chen, Hongyu Li, et al. 4dthinker: Thinking with 4d imagery for dynamic spatial understanding. *arXiv preprint arXiv:2605.05997*, 2026.

Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 7779–7788, 2020.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024.

Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6207–6217, 2021.

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. *arXiv preprint arXiv:2505.21500*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *Advances in Neural Information Processing Systems*, 38:15984–16009, 2026.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European conference on computer vision*, pp. 531–548. Springer, 2022.

Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3262–3272, 2023b.

Xingyu Miao, Haoran Duan, Quanhao Qian, Jiuniu Wang, Yang Long, Ling Shao, Deli Zhao, Ran Xu, and Gongjie Zhang. Towards scalable spatial intelligence via 2d-to-3d data lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 945–959, 2025.

Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. Gta: A geometry-aware attention mechanism for multi-view transformers. In *International Conference on Learning Representations*, volume 2024, pp. 8172–8208, 2024.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, volume 2024, pp. 51575–51598, 2024.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4938–4947, 2020.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8922–8931, 2021.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in neural information processing systems*, 38: 13569–13597, 2026.

Kaixin Xiong, Shi Gong, Xiaoqing Ye, Xiao Tan, Ji Wan, Errui Ding, Jingdong Wang, and Xiang Bai. Cape: Camera view position embedding for multi-view 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21570–21579, 2023.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pp. 131–147. Springer, 2024.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026.

Gongjie Zhang, Wenhao Li, Quanhao Qian, Jiuniu Wang, Deli Zhao, Shijian Lu, and Ran Xu. On the generalization capacities of mllms for spatial intelligence. *arXiv preprint arXiv:2603.06704*, 2026a.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *Advances in Neural Information Processing Systems*, 38, 2026b.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14227–14238, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024b. URL https://arxiv.org/abs/2410.02713.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*, 2025.

Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems*, 38:20560–20586, 2026.

Shengchao Zhou, Yuxin Chen, Yuying Ge, Wei Huang, Jiehong Lin, Ying Shan, and Xiaojuan Qi. Learning to reason in 4d: Dynamic spatial understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9637–9646, 2026.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4295–4305, 2025.

A Code

Our code is available at *https://anonymous.4open.science/r/GeoMLLM-BEC5*.

B Supplementary Experimental Setup

B.1 Benchmark Descriptions

SPAR-Bench. SPAR-Bench (Zhang et al., 2026b) contains 7,207 manually verified questions from 20 SPAR-7M test tasks, spanning calibrated single- and multi-view inputs. The tasks progress from low-level depth and distance perception, through position matching, camera motion, and viewpoint-change inference, to high-level object relations and spatial imagination. Numerical predictions are evaluated by mean relative accuracy (MRA), whereas classification tasks use accuracy.

MMSI-Bench. MMSI-Bench (Yang et al., 2026) comprises 1,000 human-authored, four-option questions over 1,990 unique images, with 2.55 images per question on average and up to 10. Every question requires evidence from multiple images and is paired with a step-by-step reference rationale. Its taxonomy includes six positional relations among cameras, objects, and semantic regions; measurement and appearance attributes; camera and object motion; and a multi-step reasoning category. We group these 11 task types into relation, attribute, motion, and multi-step reasoning (MSR), and report exact-match accuracy.

ViewSpatial-Bench. ViewSpatial-Bench (Li et al., 2025) contains 5,712 multiple-choice examples from 1,338 scenes derived from ScanNet and MS-COCO. It evaluates five localization tasks: camera-centric relative direction and object orientation, and person-centric relative direction, object orientation, and scene-simulation relative direction. The benchmark combines 3D-assisted annotation with manual verification, distinguishes 18 directions, and is approximately balanced between camera- (48.4%) and person-centric (51.6%) questions. We report both viewpoint aggregates and their overall accuracy.

CV-Bench. CV-Bench (Tong et al., 2024) reformulates standard vision annotations into 2,638 manually inspected multiple-choice questions. Its 2D portion contains 650 spatial-relation and 788 object-counting examples sourced from ADE20K and COCO; its 3D portion contains 600 depth-order and 600 relative-distance examples sourced from Omni3D. The official overall score gives equal weight to the 2D and 3D accuracies rather than weighting tasks by sample count.

Compared models. We compare with Qwen2.5-VL, Qwen3-VL, InternVL2.5/3, LLaVA-OneVision, and available proprietary-model results, together with the spatially specialized VG-LLM (Zheng et al., 2026), SPAR (Zhang et al., 2026b), and CA-MLLM (Zhang et al., 2026a). A dash in a result table denotes an unavailable or unreported entry.

B.2 Training Data

Following the training-data configuration of VG-LLM (Zheng et al., 2026), we train GeoMLLM on a mixture of SPAR-7M-RGBD (Zhang et al., 2026b) and the LLaVA-Hound split of LLaVA-Video-178K (Zhang et al., 2024b). SPAR-7M-RGBD provides calibrated single- and multi-view RGB-D examples for spatial instruction tuning and correspondence-target construction, while the LLaVA-Hound split provides general video instruction data. Since LLaVA-Hound does not provide depth maps or camera parameters, we use Pi3 (Wang et al., 2026) and UniDepthV2 (Piccinelli et al., 2025) to estimate the missing camera parameters and per-frame depth maps, respectively. These estimates supply the camera geometry and depth supervision required by the proposed modules. Videos are represented as multi-image sequences, with at most eight frames sampled from each video.

B.3 Implementation Details

Architecture and geometry inputs. We instantiate GeoMLLM with Qwen2.5-VL-3B-Instruct and Qwen3-VL-4B-Instruct backbones. Both backbone configurations use a two-block visual connector with eight attention heads, and the pairwise-geometry MLP has hidden dimension 128. The

ray encoder uses six Fourier bands and a hidden dimension of 256; both the ray residual scale and the geometry-bias output scale start at 10^{-3}. The first input view defines the reference frame.

The frozen VGGT-1B branch always extracts patch-level geometry features, which are spatially merged by a trainable MLP and added after the pose-aware connector. Valid dataset intrinsics are retained per view, while invalid ones are filled from VGGT. Dataset poses are retained only when every pose in the group is valid; otherwise, the complete VGGT trajectory is used. For these fallback groups, paired VGGT depth is used for correspondence supervision even when dataset depth exists, preserving a common scale and coordinate gauge. Only the visual encoder and VGGT-1B branch are frozen; the remaining model parameters, including the proposed geometry-aware modules, are optimized jointly.

Optimization and correspondence supervision. We supervise the attention probabilities averaged over all heads in the final connector layer with $\beta_{\text{corr}} = 0.1$. Bilinear splatting produces the soft targets. We reject depths below 10^{-4} and use target-view depth consistency with $\tau_{\text{abs}} = 0.10$ and $\tau_{\text{rel}} = 0.05$. We train for one epoch with AdamW, global batch size 64, learning rate 10^{-5}, weight decay 0.01, 3% linear warmup, and cosine decay. Training uses gradient checkpointing, DeepSpeed ZeRO-2, and random seed 0. All experiments are conducted on eight NVIDIA A800 GPUs. The per-image pixel budget ranges from 16×28^2 to 576×28^2.

Evaluation protocol. We apply the same evaluation protocol to all benchmarks. Each model is evaluated on the official test split with the benchmark-provided prompts. Decoding is greedy with temperature 0, top-$p = 1$, one beam, and no sampling. Multiple-choice questions are scored using exact-match accuracy, whereas numerical-answer questions are scored using mean relative accuracy (MRA), averaged over relative-accuracy thresholds from 0.50 to 0.95 in increments of 0.05. Category-level results are macro-averaged over task types. All values are reported as percentages, and higher is better.

C Additional Method Details

For view k, let $\boldsymbol{K}_k$ be the intrinsic matrix and $\boldsymbol{T}_k^{c2w} = [\boldsymbol{R}_k \mid \boldsymbol{t}_k]$ the camera-to-world transform. Each image contains P visual tokens, and the first input view is used as the reference frame. We write $\boldsymbol{T}_k^{w2c} = (\boldsymbol{T}_k^{c2w})^{-1}$.

C.1 Camera-Ray Construction

Let $\tilde{\boldsymbol{u}}_p = (u_p, v_p, 1)^\top$ denote the homogeneous pixel coordinate of patch center p. Its camera-frame direction, world-frame direction, and reference-frame ray are

$$\boldsymbol{d}_{k,p}^{\text{cam}} = \text{norm}\big(\boldsymbol{K}_k^{-1}\tilde{\boldsymbol{u}}_p\big), \qquad \boldsymbol{d}_{k,p}^{\text{world}} = \text{norm}\big(\boldsymbol{R}_k\boldsymbol{d}_{k,p}^{\text{cam}}\big), \tag{7}$$

$$\boldsymbol{d}_{k,p}^{\text{ref}} = \text{norm}\big(\boldsymbol{R}_r^\top\boldsymbol{d}_{k,p}^{\text{world}}\big), \qquad \boldsymbol{o}_k^{\text{ref}} = \boldsymbol{R}_r^\top(\boldsymbol{t}_k - \boldsymbol{t}_r)/s_{\text{ray}}. \tag{8}$$

Here $\text{norm}(\cdot)$ denotes ℓ_2 normalization and s_{ray} is the median non-zero distance from the reference camera to the remaining cameras; we use $s_{\text{ray}} = 1$ for single-view inputs. With the Plücker moment $\boldsymbol{m}_{k,p}^{\text{ref}} = \boldsymbol{o}_k^{\text{ref}} \times \boldsymbol{d}_{k,p}^{\text{ref}}$, the per-token ray feature is

$$\boldsymbol{g}_{k,p} = \big[\boldsymbol{d}_{k,p}^{\text{cam}}, \boldsymbol{d}_{k,p}^{\text{ref}}, \boldsymbol{m}_{k,p}^{\text{ref}}, \boldsymbol{u}_p^{\text{norm}}, \boldsymbol{\iota}_k\big] \in \mathbb{R}^{15}, \tag{9}$$

where $\boldsymbol{u}_p^{\text{norm}} \in [-1, 1]^2$ and $\boldsymbol{\iota}_k = [f_x/W_{\text{img}}, f_y/H_{\text{img}}, c_x/W_{\text{img}}, c_y/H_{\text{img}}]$. We apply a six-band Fourier encoding, project the result with MLP_{ray}, and inject it through Equation 1.

C.2 Pairwise Geometry and Attention Bias

After flattening all views, let tokens i and j correspond to world-space rays $(\boldsymbol{o}_i, \boldsymbol{d}_i)$ and $(\boldsymbol{o}_j, \boldsymbol{d}_j)$. Translation-dependent quantities are normalized by s_{pair}, the median non-zero cross-view camera baseline. Define $\boldsymbol{w}_{ij} = \boldsymbol{o}_i - \boldsymbol{o}_j$, $c_{ij} = \boldsymbol{d}_i^\top\boldsymbol{d}_j$, $d_{ij} = \boldsymbol{d}_i^\top\boldsymbol{w}_{ij}$, and $e_{ij}^{\text{ray}} = \boldsymbol{d}_j^\top\boldsymbol{w}_{ij}$. The closest points on the two viewing lines are parameterized by

$$\alpha_{ij} = \frac{c_{ij}e_{ij}^{\text{ray}} - d_{ij}}{1 - c_{ij}^2}, \qquad \beta_{ij} = \frac{e_{ij}^{\text{ray}} - c_{ij}d_{ij}}{1 - c_{ij}^2}, \tag{10}$$

where $\boldsymbol{p}_i = \boldsymbol{o}_i + \alpha_{ij}\boldsymbol{d}_i$ and $\boldsymbol{p}_j = \boldsymbol{o}_j + \beta_{ij}\boldsymbol{d}_j$. We use their distance $\delta_{ij} = \|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2$, the signs of α_{ij} and β_{ij}, and a parallel-ray indicator; the denominator is clamped by ϵ for nearly parallel rays. The image-plane epipolar distance is

$$e_{ij}^{\text{epi}} = \left|\tilde{\boldsymbol{u}}_j^\top \boldsymbol{F}_{ij}\tilde{\boldsymbol{u}}_i\right| / \left\|(\boldsymbol{F}_{ij}\tilde{\boldsymbol{u}}_i)_{1:2}\right\|_2 . \tag{11}$$

Here $\boldsymbol{F}_{ij}$ is the fundamental matrix. We normalize the distance by $s_{\text{patch}} = \sqrt{H_{\text{img}}^2 + W_{\text{img}}^2}/\sqrt{P}$ and apply the fixed penalty only to valid, non-degenerate cross-view pairs.

The 18-dimensional descriptor $\boldsymbol{\phi}_{ij}$ concatenates three view-type and validity indicators; normalized image displacement and epipolar distance; relative-ray angle and log-normalized baseline; and normalized closest-approach distance, its baseline ratio, positive log-normalized α_{ij} and β_{ij}, front-of-camera indicators, and degeneracy flags. It is shared by all connector layers.

C.3 Depth-Based Correspondence Supervision

For source patch p in view k, we sample depth $z_{k,p}$ at the patch center and project its 3D point into each target view k':

$$\boldsymbol{X}_{k,p}^w = \boldsymbol{R}_k\left(z_{k,p}\boldsymbol{K}_k^{-1}\tilde{\boldsymbol{u}}_p\right) + \boldsymbol{t}_k, \quad (\boldsymbol{u}_{k\to k'}, z_{k\to k'}) = \pi\left(\boldsymbol{K}_{k'}, \boldsymbol{T}_{k'}^{w2c}, \boldsymbol{X}_{k,p}^w\right). \tag{12}$$

We retain cross-view projections that have valid source depth, lie in front of the target camera and inside its image, and satisfy the depth-consistency thresholds in Appendix B. For a valid projection (g_x, g_y) on the target token grid, bilinear splatting assigns weights to the four neighboring cells $c = (x_c, y_c)$:

$$\widetilde{w}_{a,c} = (1 - |g_x - x_c|)_+(1 - |g_y - y_c|)_+, \qquad w_{a,c} = \frac{\widetilde{w}_{a,c}}{\sum_{c'\in\mathcal{N}_a}\widetilde{w}_{a,c'}}. \tag{13}$$

Invalid out-of-grid corners are removed before the remaining weights are renormalized. Equation 5 is applied to the head-averaged, pre-dropout attention probabilities of the final connector layer, with $\beta_{\text{corr}} = 0.1$ in Equation 6.

D Task-Level Results on Additional Benchmarks

Task-level gains are broad but not uniform. Tables 4 and 5 expand the category-level results in Table 2. With the Qwen2.5-VL-3B backbone, GeoMLLM improves MMSI-Bench camera motion by **+27.03** points, attribute measurement by **+20.32** points, and the camera–object relation by **+13.96** points, while object motion changes by **-9.21** points. On ViewSpatial-Bench, the three person-centric tasks improve by **+30.22**, **+11.29**, and **+8.51** points, whereas camera-centric object orientation changes by **-20.79** points. With the Qwen3-VL-4B backbone, the largest gains occur on MMSI-Bench camera motion (**+18.92**), multi-step reasoning (**+13.13**), and the camera–region relation (**+13.25**), as well as ViewSpatial-Bench person-centric object orientation (**+28.92**); declines remain on attribute measurement (**-9.38**), camera-centric object orientation (**-10.04**), and person-centric relative direction (**-4.75**). These task-level results show that the aggregate transfer arises from multiple reasoning capabilities but remains dependent on task type and backbone.

Geometry-aware attention contributes under a matched comparison. Ours$_{\text{Qwen2.5-3B}}$ (MHSA) replaces Camera Geometry-Aware Attention with standard multi-head self-attention while retaining the remaining model and training configuration. Relative to this configuration, the complete model gains **+17.19** points on MMSI-Bench attribute measurement, **+14.87** points on camera motion, and **+8.14** points on ViewSpatial-Bench person-centric simulated relative direction. The effect is again non-uniform, including changes of **-7.52** points on the MMSI-Bench camera–camera relation and **-4.62** points on ViewSpatial-Bench camera-centric object orientation. CV-Bench provides only the official 2D and 3D component scores rather than task-level results; these are reported in Table 2 and are therefore not repeated here.

E Additional Evaluation on VSI-Bench

VSI-Bench (Yang et al., 2025) contains more than 5,000 questions generated from 288 egocentric indoor videos in the ScanNet, ScanNet++, and ARKitScenes validation sets. Its eight tasks cover

Table 4: Task-level accuracy (%) on MMSI-Bench. APPR. and MEAS. denote appearance and measurement; CAM., OBJ., and REG. denote camera, object, and semantic region. Relation columns specify the ordered entity pair. Overall is the macro-average over all 11 tasks.

METHOD	ATTRIBUTE		MSR	MOTION		POSITIONAL RELATION						OVERALL
	APPR.	MEAS.		CAM.	OBJ.	CAM.–CAM.	CAM.–OBJ.	CAM.–REG.	OBJ.–OBJ.	OBJ.–REG.	REG.–REG.	
OPEN-SOURCE MODELS												
Qwen2.5-VL-3B	19.70	20.31	21.21	8.11	32.89	25.81	30.23	32.53	26.60	25.88	28.40	24.70
Qwen2.5-VL-7B	21.21	21.88	26.26	18.92	30.26	24.73	30.23	31.33	29.79	37.65	25.93	27.11
Qwen3-VL-4B	18.18	42.19	22.73	29.73	26.32	17.20	38.37	19.28	26.60	31.76	20.99	26.67
Qwen3-VL-8B	19.70	35.94	29.29	24.32	27.63	29.03	32.56	33.73	28.72	36.47	37.04	30.40
OPEN-SOURCE SPATIAL INTELLIGENCE MODELS												
VG-LLM-4B	28.79	31.25	24.75	14.86	26.32	30.11	39.53	32.53	29.79	24.71	32.10	28.61
CA-MLLM-4B	25.76	26.56	26.26	14.86	21.05	30.11	41.86	34.94	25.53	31.76	20.99	27.24
$\text{Ours}_{\text{Qwen2.5-3B}}$ (MHSA)	22.73	23.44	25.76	20.27	22.37	33.33	44.19	25.30	30.85	28.24	33.33	28.16
$\text{Ours}_{\text{Qwen2.5-3B}}$	33.33	40.63	27.27	35.14	23.68	25.81	44.19	32.53	29.79	30.59	30.86	32.16
$\text{Ours}_{\text{Qwen3-4B}}$	25.76	32.81	35.86	48.65	28.95	25.81	37.21	19.28	30.85	34.12	23.46	31.16

Table 5: Task-level accuracy (%) on ViewSpatial-Bench. OBJ. ORI., REL. DIR., and SIM. REL. DIR. denote object-view orientation, relative direction, and scene-simulation relative direction, respectively. Overall is the macro-average over the five tasks.

METHOD	CAMERA-CENTRIC		PERSON-CENTRIC			OVERALL
	OBJ. ORI.	REL. DIR.	OBJ. ORI.	REL. DIR.	SIM. REL. DIR.	
OPEN-SOURCE MODELS						
Qwen2.5-VL-3B	31.43	44.78	43.78	28.50	29.86	35.67
Qwen2.5-VL-7B	31.02	47.15	40.76	37.41	28.69	37.01
Qwen3-VL-4B	27.91	48.90	40.96	40.74	27.78	37.26
Qwen3-VL-8B	30.32	54.20	42.47	36.94	30.77	38.94
OPEN-SOURCE SPATIAL INTELLIGENCE MODELS						
VG-LLM-4B	14.36	49.18	75.50	38.00	34.75	42.36
CA-MLLM-4B	10.44	58.15	63.05	34.32	30.41	39.27
$\text{Ours}_{\text{Qwen2.5-3B}}$ (MHSA)	15.26	48.84	73.49	37.77	30.23	41.12
$\text{Ours}_{\text{Qwen2.5-3B}}$	10.64	56.23	74.00	39.79	38.37	43.81
$\text{Ours}_{\text{Qwen3-4B}}$	17.87	61.87	69.88	35.99	36.20	44.36

spatial configuration (object count, relative distance and direction, and route planning), metric estimation (absolute distance, object size, and room size), and spatiotemporal memory (appearance order). Following the official protocol, numerical-answer tasks use MRA and multiple-choice tasks use accuracy. We use VSI-Bench only for the additional evaluation in Table 6; it is not part of the main SPAR-Bench experiment.

Transfer is strongest on metric-estimation tasks. Relative to the matched Qwen2.5-VL-3B backbone, GeoMLLM improves the VSI-Bench average by **+19.10** points (Table 6). The largest gains occur on object counting (**+35.70**), object size (**+35.10**), room size (**+33.30**), and absolute distance (**+18.70**), with an additional **+15.90** points on appearance order. Transfer is weaker on relative distance (**+11.20**) and route planning (**+2.60**), while relative direction remains unchanged. In aggregate, GeoMLLM exceeds VG-LLM-4B and CA-MLLM-4B by **+0.40** and **+0.90** points, respectively, but trails GPT-5 and Gemini-2.5-Pro by **1.40** and **5.00** points. VSI-Bench therefore provides additional evidence of transfer beyond the main benchmarks while showing that the gains are concentrated in metric estimation rather than uniformly distributed across spatial tasks.

F TRAINING CONVERGENCE

The language and correspondence objectives decrease jointly. Figure 8 shows the total training objective together with the language-model cross-entropy loss $\mathcal{L}_{\text{CE}}$ and the unweighted correspondence loss $\mathcal{L}_{\text{corr}}$. All three curves decrease sharply at the beginning of training and then enter a slower, stable regime. Although $\mathcal{L}_{\text{corr}}$ exhibits greater batch-to-batch variation than $\mathcal{L}_{\text{CE}}$, neither component shows late-stage divergence, and the total objective continues to decline. For the reported training run, these trajectories show that language modeling and geometry-aware correspondence supervision can be optimized jointly under the stated schedule; they do not establish convergence behavior across random seeds or alternative configurations.

Table 6: Additional evaluation on VSI-Bench. Numerical-answer tasks use MRA, multiple-choice tasks use accuracy, and Avg. is the unweighted mean over all eight tasks.

Methods	Avg.	Numerical Answer				Multiple-Choice Answer			
		Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Baseline									
Chance Level (Random)	–	–	–	–	–	25.0	36.1	28.3	25.0
Chance Level (Frequency)	34.0	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2
Proprietary Models (API)									
GPT-4o	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5-Flash	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5-Pro	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
GPT-5	49.1	52.4	36.6	68.0	55.5	49.9	47.4	33.0	49.8
Gemini-2.5-Pro	52.7	48.2	35.6	71.3	51.7	58.9	42.4	46.8	66.7
Open-Source Models									
LLaVA-OneVision-7B	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
LLaVA-NeXT-Video-7B	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
InternVL2-8B	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
Qwen2.5-VL-3B	28.6	32.7	19.5	17.3	25.1	37.3	44.9	30.4	21.8
Qwen2.5-VL-7B	29.3	25.2	10.5	36.4	29.6	38.4	38.0	29.8	26.8
Qwen2.5-VL-72B	37.0	25.1	29.3	54.5	38.8	38.2	37.0	34.0	28.9
Open-Source Spatial Intelligence Models									
SPAR-8B	44.1	–	–	–	–	–	–	–	–
SpatialLadder-3B	44.8	–	–	–	–	–	–	–	–
VG-LLM-4B	47.3	66.0	37.8	55.2	59.2	44.6	45.6	33.5	36.4
CA-MLLM-4B	46.8	71.3	39.3	50.4	65.9	50.7	45.5	31.4	20.2
$\text{Ours}_{\text{Qwen2.5-3B}}$	47.7	68.4	38.2	52.4	58.4	48.5	44.9	33.0	37.7

"–" denotes an unavailable or unreported entry. Obj., Abs., Rel., and Appr. abbreviate object, absolute, relative, and appearance, respectively.

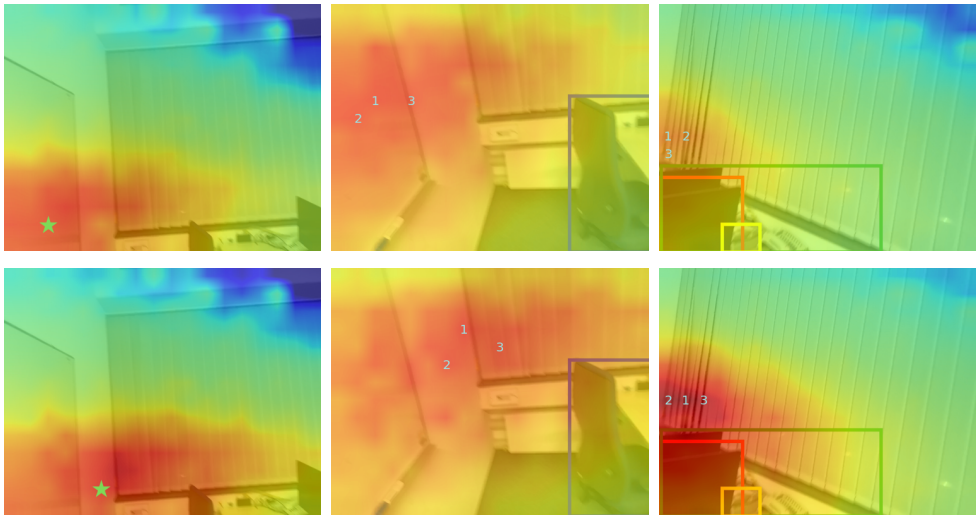


Figure 8: Training trajectories of the total objective, language-model cross-entropy loss ($\mathcal{L}_{\text{CE}}$), and unweighted correspondence loss ($\mathcal{L}_{\text{corr}}$). The total objective weights $\mathcal{L}_{\text{corr}}$ by $\beta_{\text{corr}} = 0.1$.

G Additional Cross-View Attention Visualizations

Correct predictions exhibit spatially localized cross-view attention. Figures 9, 10, 11, 12, 13, and 14 extend Figure 7 across additional scenes and viewpoint changes. Each row selects a star-marked source patch; heatmaps show its connector attention over three views, and numbered markers identify the three highest-scoring target patches. High-attention regions remain localized around spatially related content in these successful cases, although the visualizations alone do not establish causal sufficiency.

Failure cases expose a gap between cross-view association and final spatial judgment. Figures 15 and 16 show that localized attention does not guarantee a correct answer. Ambiguous content or weak relative-depth evidence leaves multiple relations plausible, so the final judgment still depends on resolving object identity, depth ordering, and the viewpoint-conditioned relation. Green denotes ground truth and red an incorrect prediction.



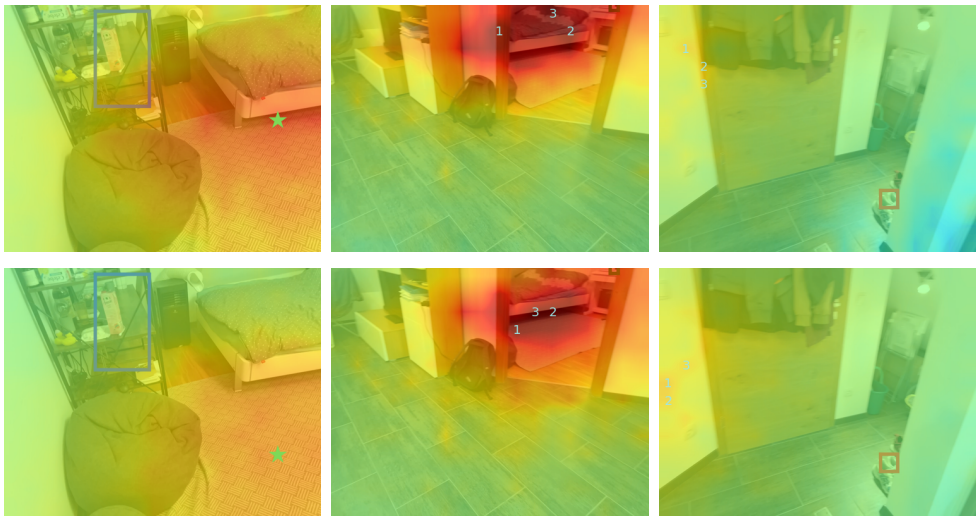
Question: Describe the spatial configuration of monitor (red bbox) relative to telephone (yellow bbox) as initially seen by the observer and how this changes after observer moving to window sill (green bbox) and facing office chair (blue bbox).
A. right, above, closer, , behind B. right, below, farther, front, behind
C. right, above, farther, front, front D. left, above, farther, front, behind
Prediction: C. right, above, farther, front, front

Figure 9: **Correct case 007101.** GeoMLLM selects the green ground truth.



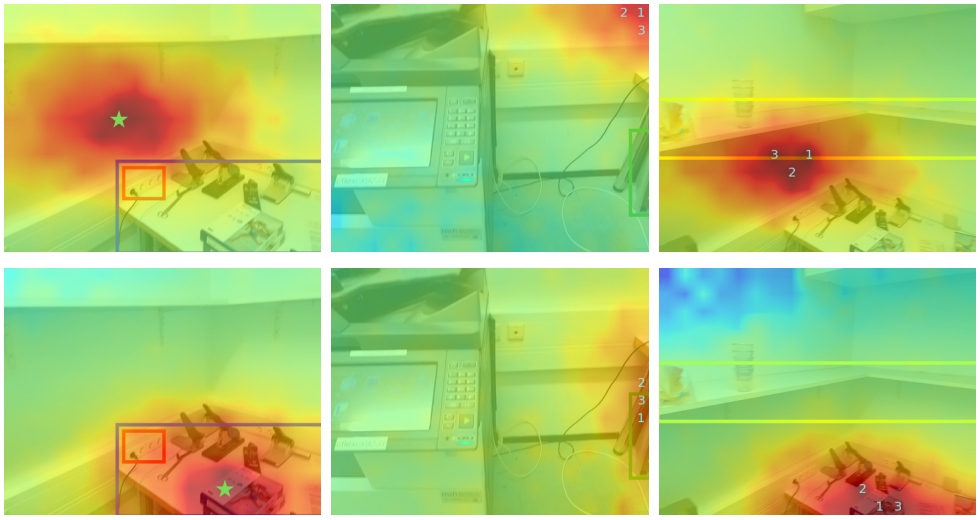
Question: How does the observer movement to picture (green bbox) and orientation towards lamp (blue bbox) influence the observer's view of the spatial relationship between machine (red bbox) and picture (yellow bbox)?
A. left, below, front, front B. right, above, front, behind
C. right, farther, front, front D. right, farther, front, behind
Prediction: C. right, farther, front, front

Figure 10: **Correct case 007105.** Localized attention accompanies the correct prediction.



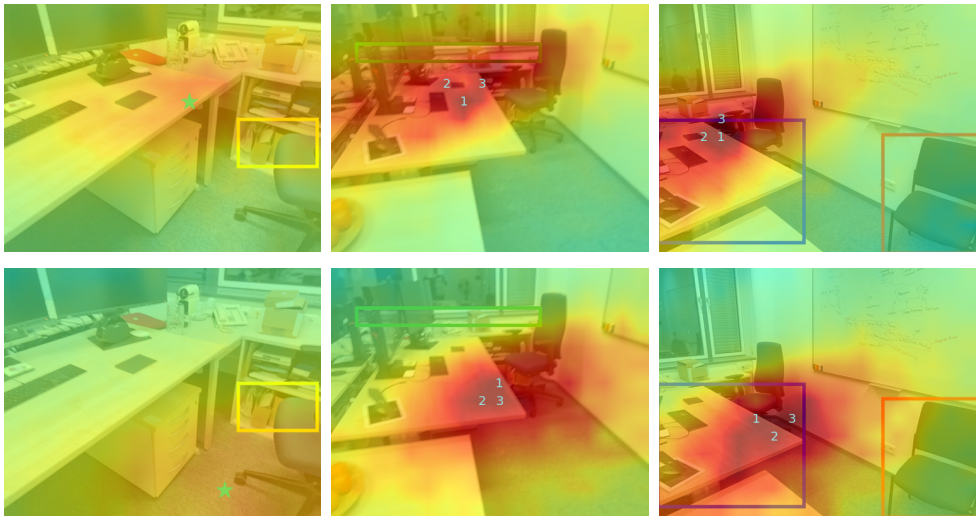
Question: From the observer's point of view, what is the spatial positioning of shoe (red bbox) and power socket (yellow bbox) initially, and how does this relationship change after observer relocating to headset (green bbox) and facing chair (blue bbox)?
A. left, above, farther, behind B. left, below, farther, behind, behind
C. left, below, farther, behind D. right, above, farther, behind, behind
Prediction: B. left, below, farther, behind, behind

Figure 11: **Correct case 007106.** GeoMLLM selects the green ground truth.



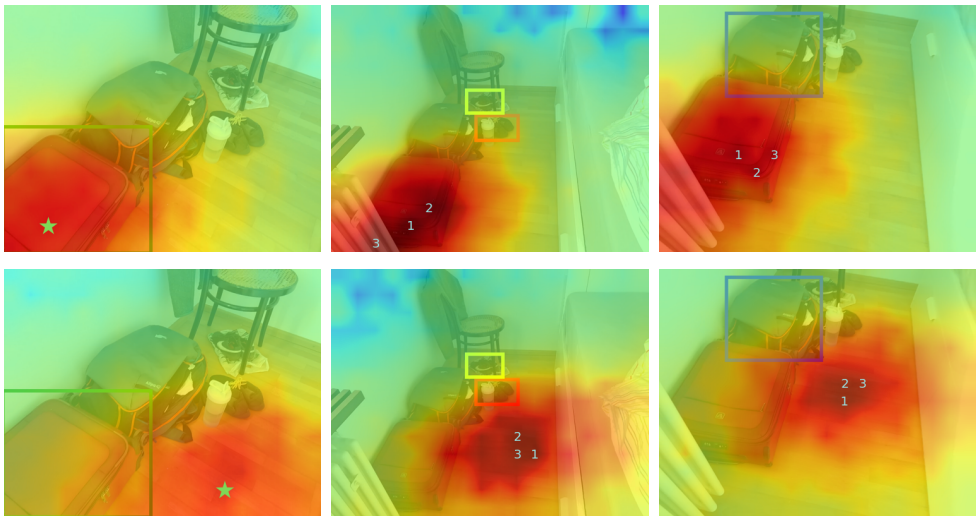
Question: How does the spatial orientation of power socket (red bbox) and shelf (yellow bbox) shift in the observer's view after observer moving to board (green bbox) and focusing on table (blue bbox)?
A. right, above, farther, behind B. right, below, closer, front, front
C. left, above, front, behind D. right, below, closer, front
Prediction: D. right, below, closer, front

Figure 12: **Correct case 007128.** Attention remains localized after the viewpoint change.



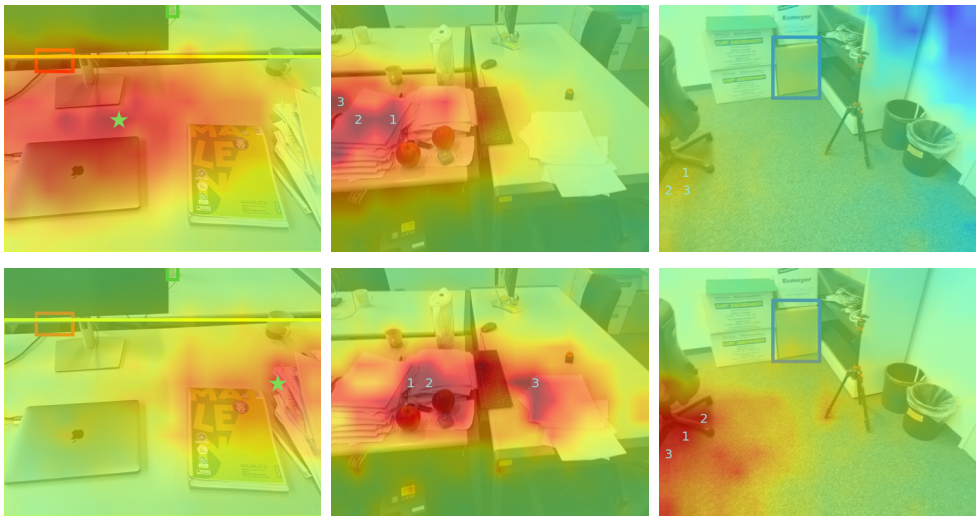
Question: How does the observer's perception of the spatial positioning of chair (red bbox) and objects (yellow bbox) change when observer relocates to windowsill (green bbox) and face table (blue bbox)?
A. left, farther, behind, front B. right, above, farther, behind
C. below, farther, behind, front D. left, below, farther, behind
Prediction: A. left, farther, behind, front

Figure 13: **Correct case 007154.** GeoMLLM selects the green ground truth.



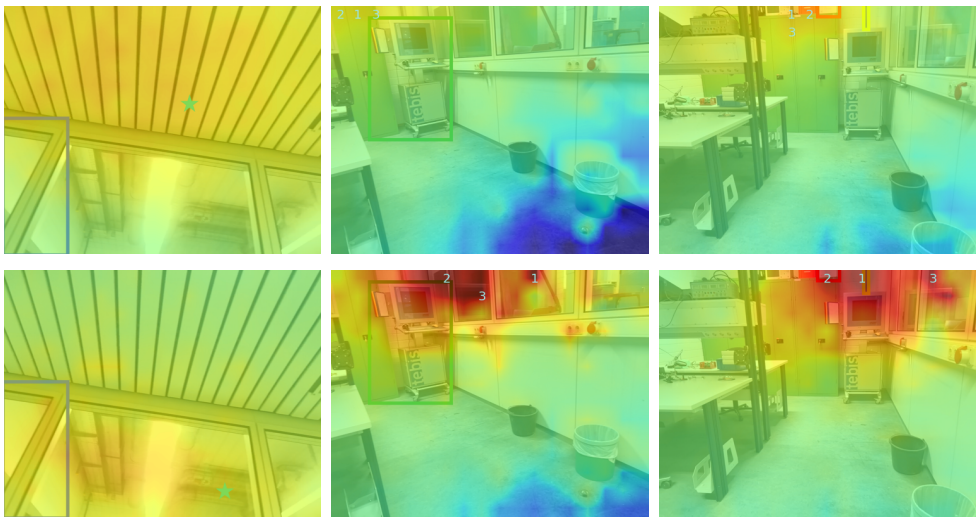
Question: What is the observer's perspective of shoes (red bbox) and shoes (yellow bbox) before and after observer relocating to suitcase (green bbox) and facing blanket (blue bbox)?
A. left, above, farther, behind, front B. left, above, closer, front, front
C. left, below, closer, front, front D. right, below, farther, front
Prediction: B. left, above, closer, front, front

Figure 14: **Correct case 007157.** Localized attention accompanies the correct prediction.



Question: From the observer's perspective, what is the relationship between computer tower (red bbox) and table (yellow bbox) before moving? How does it evolve after camera moving to 3d center of mug (green bbox) and facing 3d center of box (blue bbox)?
A. left, above, front, behind B. left, above, closer, behind, behind
C. left, closer, front, front D. left, closer, front, behind
Prediction: B. left, above, closer, behind, behind

Figure 15: **Failure case 007123.** Red: prediction; green: ground truth.



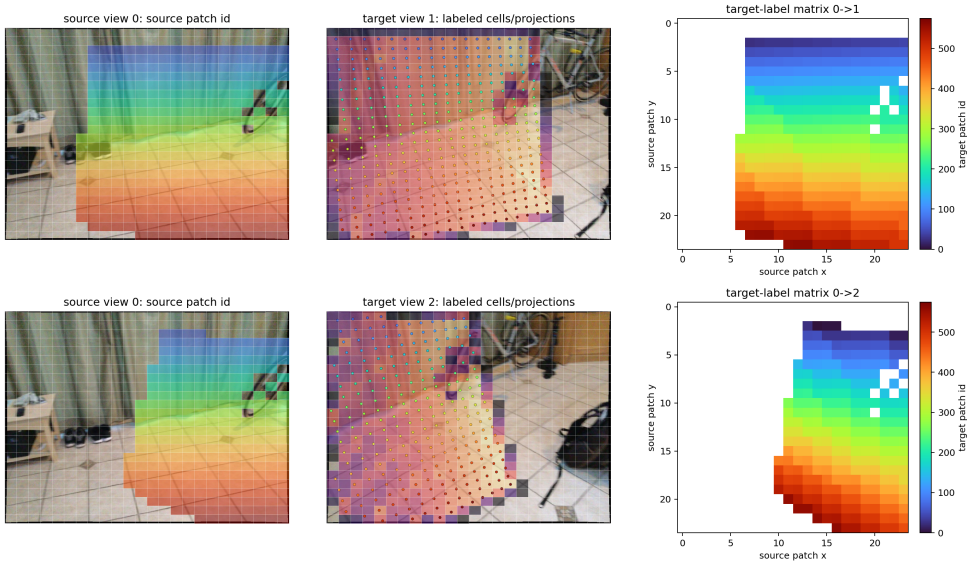
Question: Describe the spatial dynamics between cardboard box (red bbox) and pipe (yellow bbox) as seen initially and after the observer moves to laboratory equipment (green bbox) and focuses on window frame (blue bbox).
A. right, below, closer, behind B. right, above, closer, behind, front
C. right, farther, front, front D. right, closer, front, front
Prediction: C. right, farther, front, front

Figure 16: **Failure case 007153.** Attention does not resolve relative distance.

H CORRESPONDENCE TARGETS ACROSS DATASETS

The projection pipeline produces spatially coherent targets across data sources. Figures 17, 18, and 19 extend the correspondence-target visualization to real ScanNet and ScanNet++ scenes and synthetic Structured3D scenes. Across datasets and viewpoint changes, neighboring valid source patches generally project to neighboring target regions, while missing overlap, out-of-frame projections, and failed depth-consistency checks produce structured invalid regions. In each row, the left panel colors source patches by flattened patch index, the middle panel transfers those colors to valid projections in a target view, and the right panel records the assigned target-patch index at each source-grid location; white cells indicate that no valid target is retained. These visualizations verify the outputs of the depth- and calibration-based supervision pipeline across heterogeneous scene sources, but do not depict attention predicted by GeoMLLM.

(a) `scene0000_00`



(b) `scene0013_01`

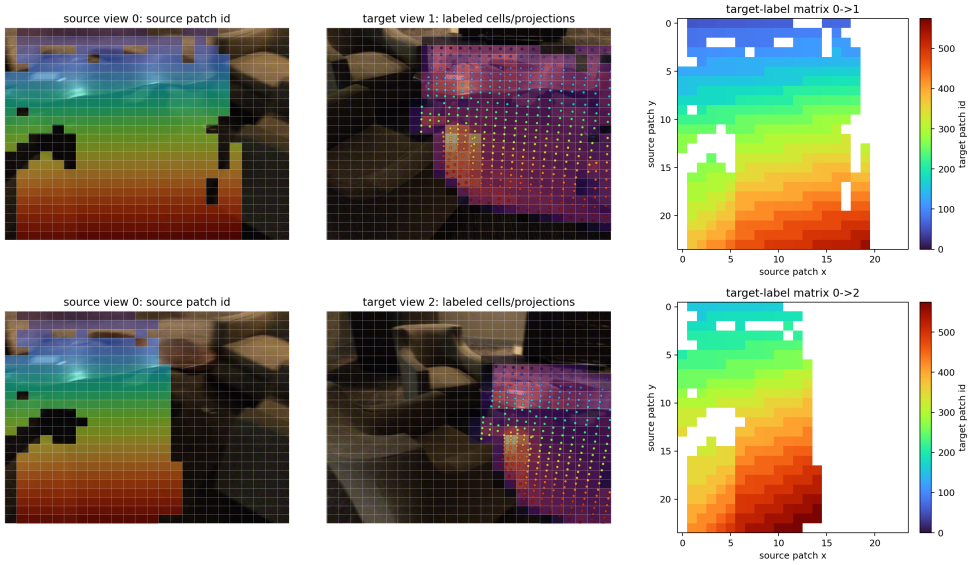
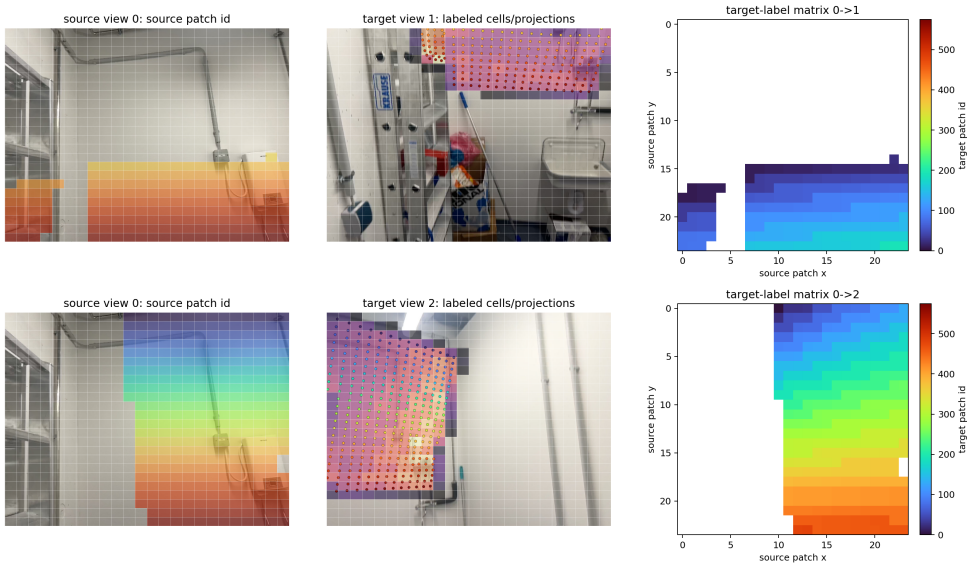


Figure 17: **Correspondence targets for two ScanNet scenes.** Colors identify source patches and their projected target locations; white matrix cells denote invalid correspondences.

(a) `0a5c013435`



(b) `05d3c8c6f0`

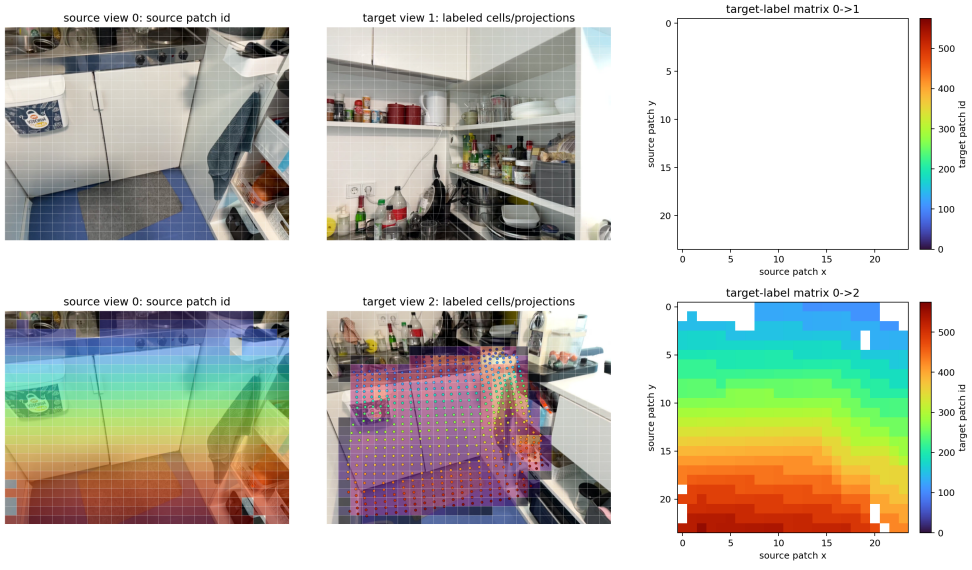
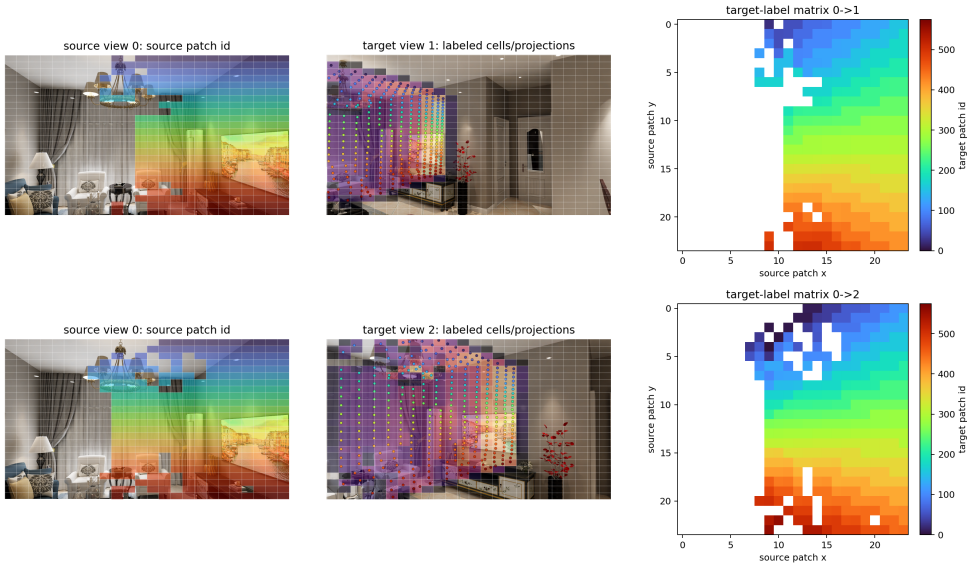


Figure 18: **Correspondence targets for two ScanNet++ scenes.** Colors identify source patches and their projected target locations; white matrix cells denote invalid correspondences.

(a) `scene_00003`



(b) `scene_01500`

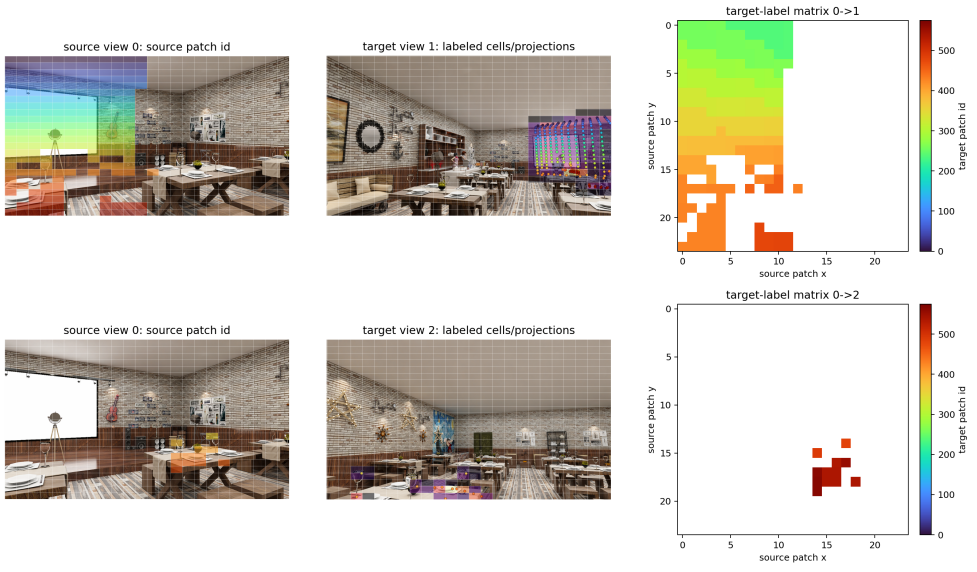


Figure 19: **Correspondence targets for two Structured3D scenes.** Colors identify source patches and their projected target locations; white matrix cells denote invalid correspondences.