Eliciting Complex Spatial Reasoning in MLLMs through Wide-Baseline Matching

Supplementary Material

6. Appendix Overview

This supplementary material provides comprehensive details supporting our main paper, organized as follows:

Sec. 7 – Implementation Details: Complete specifications of data generation pipeline, experimental setup, prompt template, and curriculum progression schedules.

Sec. 8 – Additional Ablations: Extended analyses on curriculum design variants, overlap scheduling strategies, and detailed RL vs. SFT comparisons across training stages.

Sec. 9 – Benchmark Analysis: Detailed benchmark statistics, per-category model performance, failure mode analyses, and qualitative examples illustrating model capabilities and limitations.

Sec. 10 – Extended Discussion: Expanded analysis of limitations and concrete future research directions for advancing MLLM spatial intelligence.

7. Implementation Details

7.1. Dataset Generation

Correspondence Generation from RGB-D Data. Starting from RGB-D videos (or RGB with known intrinsics and camera-to-world poses), we back-project every pixel $\mathbf{x}$ with valid depth in image I_1 into a 3D point $\mathbf{X}$ in the world coordinate system under a static-scene assumption, and reproject it to the consecutive image I_2 using the known camera parameters to obtain the image location $\mathbf{y}^{\text{proj}}$ and its camera depth z^{proj}. At $\mathbf{y}^{\text{proj}}$ we bilinearly sample the observed depth z^{obs} and the color I_2. We source our data from diverse datasets including uCO3D [29], CO3D [39], and ScanNet [12]. We compute two verifiable consistency terms: (i) relative depth consistency

$$e_{\text{depth}} = \frac{\left|z^{\text{proj}} - z^{\text{obs}}\right|}{z^{\text{proj}} + \varepsilon}, \tag{6}$$

and (ii) photometric consistency defined as the channel-averaged RGB difference with bilinear sampling at $\mathbf{y}^{\text{proj}}$

$$e_{\text{photo}} = \tfrac{1}{3} \sum_{c\in\{r,g,b\}} \left|I_1^{(c)}(\mathbf{x}) - \hat{I}_2^{(c)}(\mathbf{y}^{\text{proj}})\right|. \tag{7}$$

A correspondence is considered valid if it meets basic visibility, boundedness criteria:

$$\mathbb{I}\big[\text{finite}(\mathbf{y}^{\text{proj}}) \wedge \text{in-bounds}(\mathbf{y}^{\text{proj}}) \wedge z^{\text{proj}} > 0 \wedge z^{\text{obs}} > 0\big], \tag{8}$$

and error criteria:

$$\mathbb{I}\big[e_{\text{depth}} < \tau_{\text{d}}\big] \cdot \mathbb{I}\big[e_{\text{photo}} < \tau_{\text{p}}\big]. \tag{9}$$

Here τ_{d} and τ_{p} are pre-defined thresholds for depth and photometric errors, respectively. Optional masks (e.g., dynamic-object or invalid-depth masks) can be applied on the image I_1 and I_2.

Correspondence Generation from SfM Data. For datasets lacking ground-truth depth, such as RealEstate10k [63] and DL3DV [27], we leverage 3D reconstructions from Structure-from-Motion (SfM). For DL3DV, we utilize the provided COLMAP [43] models directly. For RealEstate10k, which provides only RGB video streams, we first process the raw sequences with COLMAP to generate these 3D reconstructions. Ground-truth correspondences are then extracted by identifying shared 3D landmarks between two image views. Specifically, a pair of 2D keypoints $(\mathbf{x}_i, \mathbf{y}_j)$, where $\mathbf{x}_i$ is in image I_1 and $\mathbf{y}_j$ is in image I_2, is considered a valid match if they both correspond to the same 3D point $\mathbf{X}$ in the COLMAP model's sparse reconstruction. This inlier set, having already passed geometric verification within COLMAP, serves as our reliable matches, bypassing the consistency checks (Eq. (6) and (7)) used for RGB-D data.

Quantifying Viewpoint Change. To categorize the difficulty of a pair, we use a scalar overlap score $\omega \in [0, 1]$ as a proxy for co-visibility (higher means more similar views). For pairs derived from RGB-D data, the overlap ω is defined as the proportion of validly matched pixel pairs $\mathcal{M}$ relative to the total number of pixels in each frame with height H and width W:

$$\omega = \frac{|\mathcal{M}|}{H \times W} \tag{10}$$

For pairs derived from SfM data, we define this overlap based on the proportion of shared 3D landmarks. Let L_1 and L_2 be the sets of 3D landmarks (i.e., 3D points) visible in images I_1 and I_2, respectively. The overlap ω is computed as:

$$\omega = \frac{|L_1 \cap L_2|}{\min(|L_1|, |L_2|)} \tag{11}$$

where $|\cdot|$ denotes the cardinality of the set. The viewpoint-change magnitude is then defined as

$$\Delta_v \;=\; 1 - \omega, \tag{12}$$

which is monotonically larger for more disparate viewpoints. In curriculum-based generation, stages specify an

admissible overlap interval $[\underline{\omega}_s, \overline{\omega}_s]$, and a sample is routed to stage s if

$$\mathbb{I}\big[\underline{\omega}_s \le \omega \le \overline{\omega}_s\big] = 1. \tag{13}$$

Query Construction. Given a raw pair routed to stage s via its overlap score (Eq. (13)), we take its set of candidate matches $\mathcal{M} = \{(\mathbf{x}_k, \mathbf{y}_k)\}$ generated from either the RGB-D or SfM pipeline. We then select a spatially diverse *core* subset $\mathcal{C} \subset \mathcal{M}$ under stage-specific sparsity constraints. By default, the tool uses a cluster-then-prune policy: matches are clustered in the joint space with DBSCAN at radius $\varepsilon = \alpha \cdot \tau_{\text{min-dist}}$ (with $\alpha \in (0, 1)$), one representative per cluster is kept (closest to the cluster centroid), and a greedy max-spacing pass further reduces to at most K points if needed. Core matches are given a one-to-one mapping by sampling label subsets $\mathcal{L}_A^{\text{core}}, \mathcal{L}_B^{\text{core}}$ of equal size, which satisfy the bijective constraint

$$f_{\text{core}} : \mathcal{L}_A^{\text{core}} \leftrightarrow \mathcal{L}_B^{\text{core}}. \tag{14}$$

To increase difficulty, we can optionally add *distractor* points $\mathcal{D}_A, \mathcal{D}_B$ sampled from the leftover matches in each view, ensuring no overlap with core matches. Shuffling is applied to the labels in each view to avoid positional bias. Finally, the benchmark sample is packaged as

$$(I_1, \mathcal{X} = \mathcal{C}_A \cup \mathcal{D}_A; \quad I_2, \mathcal{Y} = \mathcal{C}_B \cup \mathcal{D}_B), \tag{15}$$

with ground-truth mapping

$$f : \mathcal{L}_A \to \mathcal{L}_B \cup \{\varnothing\}, \tag{16}$$

where f extends f_{core} by assigning distractors to $\varnothing$.

7.2. Dynamic Curriculum Strategy

Verifiable Reward Design. Our reward function comprises two components that jointly encourage geometric accuracy and structured reasoning. Given the predicted mapping $\hat{f} : \{1, \ldots, n\} \to \{1, \ldots, m\} \cup \{\varnothing\}$ and ground-truth mapping f^* with the same domain, where n is the total number of query regions (including those with no correspondences in ground truth), we define:

Format compliance $r_{\text{format}} \in \{0, 1, 2\}$ verifies both structural and syntactic validity:

$$r_{\text{format}} = \mathbb{1}[\text{structure}] + \mathbb{1}[\text{JSON}], \tag{17}$$

where the structure component checks that the output follows the required format with reasoning confined within thinking tags, and the JSON component verifies that the content within answer tags forms a valid JSON mapping with $\hat{f}(i) \in \{1, \ldots, m\} \cup \{\varnothing\}$ for all $i \in \{1, \ldots, n\}$.

Matching correctness $r_{\text{match}} \in [0, 1]$ measures prediction accuracy over all n query regions:

$$r_{\text{match}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left[\hat{f}(i) = f^*(i)\right], \tag{18}$$

which counts exact agreements including correct predictions of unmatched regions ($\hat{f}(i) = \varnothing$ when $f^*(i) = \varnothing$). This formulation ensures that all regions are evaluated, rewarding comprehensive spatial reasoning.

The final reward combines these components:

$$r = w_f \cdot r_{\text{format}} + w_m \cdot r_{\text{match}}, \tag{19}$$

w_f and w_m being reward weights. Both components must be satisfied for high rewards, encouraging the model to produce well-structured and geometrically accurate predictions.

7.3. Curriculum Learning Implementation Details

Our curriculum design operates at two hierarchical levels: (1) three **cardinality settings** (L1, L2, L3) that define matching task configurations, and (2) three **training stages** that progressively combine these settings with increasing complexity.

Cardinality settings. We define three cardinality configurations corresponding to different matching scenarios:

- **L1 (Unambiguous matching):** $n \sim \mathcal{U}(3, 5)$ matchable points with $|\mathcal{X}^{\text{dist}}| = |\mathcal{Y}^{\text{dist}}| = 0$. This yields 3–5 one-to-one correspondences without distractors.
- **L2 (Selective matching):** $n \sim \mathcal{U}(1, 2)$ matchable points with $|\mathcal{X}^{\text{dist}}| = 0$ and $|\mathcal{Y}^{\text{dist}}| \sim \mathcal{U}(3, 6)$. The model must identify 1–2 correct matches among 4–8 candidates in $\mathcal{Y}$.
- **L3 (Partial matching):** $n \sim \mathcal{U}(3, 6)$ matchable points with $|\mathcal{X}^{\text{dist}}|, |\mathcal{Y}^{\text{dist}}| \sim \mathcal{U}(3, 6)$. Both views contain 6–12 points with bidirectional occlusion.

Three-stage curriculum progression. Training progresses through three stages that combine cardinality settings with increasing complexity:

- **Stage 1 (L1 only):** The model trains exclusively on unambiguous matching (L1) to establish foundational geometric reasoning. Stage transition occurs when the sliding window reward $\bar{r} > 0.7$ for 10 consecutive evaluations.
- **Stage 2 (L2 only):** After mastering L1, training shifts entirely to selective matching (L2) to learn distractor rejection. The model progresses when $\bar{r} > 0.7$ is maintained.
- **Stage 3 (L1/L2/L3 mixed):** The final stage samples from all three settings with probability $p_{\text{L1}} : p_{\text{L2}} : p_{\text{L3}} = 0.3 : 0.3 : 0.4$, ensuring comprehensive exposure to diverse matching scenarios.

Adaptive demotion: If performance degrades ($\bar{r}_{match} < 0.2$ for 10 consecutive steps) during any stage, the curriculum temporarily reverts to the previous stage for stabilization before resuming progression.

Spatial distribution refinement within settings. Independent of the curriculum stage, each cardinality setting (L1, L2, L3) undergoes its own spatial refinement process. This creates a two-dimensional curriculum: the *stage* determines which cardinality settings are active, while *spatial refinement* modulates difficulty within each active setting.

For each cardinality setting, we progressively eliminate local spatial cues:

- **Initial sampling (clustered):** Points are sampled using DBSCAN clustering at radius $\varepsilon = \alpha \cdot \tau_{\text{min-dist}}$ followed by greedy max-spacing selection. Initially, $\alpha = 2.0$ and $\tau_{\text{min-dist}}$ equals the minimal non-overlap margin for annotations, producing spatially coherent clusters.
- **Progressive tightening:** Every time a level's internal reward threshold is satisfied, we update:

$$\tau_{\text{min-dist}} \leftarrow \max(\text{safe margin}, \tau_{\text{min-dist}} - 20) \text{ pixels} \quad (20)$$

  This gradually decreases minimum point separation and reduces clustering radius.
- **Final sampling (dispersed):** When the final threshold is achieved inside a level, sampling transitions to pure greedy max-spacing, maximizing spatial dispersion and eliminating local context cues.

Importantly, spatial refinement operates *independently* for each cardinality setting. When Stage 2 begins, L2 starts with clustered sampling even though L1 may have progressed to dispersed sampling. This ensures each matching configuration is learned from coarse to fine spatial distributions.

7.4. Prompt Design

We design a structured prompt that clearly specifies the cross-view matching task, input format, and expected output structure. As illustrated in Fig. 4, the prompt instructs the model to identify corresponding keypoint locations in a target image given marked query points in the source image. The prompt explicitly defines reasoning requirements, output format requirements (JSON structure with match validity flags). This design ensures consistent task interpretation across different models while enabling verifiable evaluation through structured output parsing.

8. Ablation Studies

Reinforcement Learning vs. Supervised Fine-tuning. To assess our curriculum-based reinforcement learning approach, we compare against supervised fine-tuning (SFT) on the same cross-view matching data. The SFT baseline uses 300 steps of supervised training with teacher-forced matching predictions, while our DCRL method employs reinforcement learning with holistic geometric rewards as described in Sec. 3.

Table 7. Ablation study on general vision-language understanding benchmarks. Both SFT and DCRL maintain strong performance on general tasks, with DCRL showing slight advantages in our evaluation.

Method	MME-RealWorld	MMStar	RealWorldQA	V*
Base Model	62.8	59.8	69.5	84.8
SFT	62.4	56.4	68.8	82.2
Ours (DCRL)	**63.8**	**62.5**	**70.5**	**85.9**

Table 8. Ablation study on spatial intelligence and geometric reasoning benchmarks. DCRL substantially outperforms SFT, particularly on tasks requiring fine-grained geometric correspondence (SAT, ReasonMatch), which is consistent with a benefit from curriculum-based reinforcement learning for spatial reasoning.

Method	OmniSpatial	MindCube	SAT	ReasonMatch
Base Model	43.6	40.0	70.0	27.5
SFT	42.6	45.1	41.3	51.0
Ours (DCRL)	**48.9**	**43.5**	**75.3**	**70.5**

General vision-language understanding. As shown in Table 7, DCRL maintains strong performance on general vision-language benchmarks and modestly improves over both the base model and the SFT baseline. While SFT degrades on MMStar (-3.4 points) and V* (-2.6 points), potentially due to catastrophic forgetting under domain shift, our RL-based approach preserves and slightly improves general capabilities ($+2.7$ on MMStar, $+1.1$ on V*). This result suggests that curriculum-based reinforcement learning with geometric rewards can strengthen spatial training without harming pre-existing vision-language understanding.

Spatial intelligence and geometric reasoning. Table 8 shows notable advantages of RL over SFT on spatial reasoning tasks. On our cross-view matching benchmark (ReasonMatch), DCRL achieves 70.5% compared with 51.0% for SFT, a 19.5-point improvement. The contrast is even larger on SAT (+34.0 points), where SFT drops from the base model (70.0 $\rightarrow$ 41.3) while RL improves over it (70.0 $\rightarrow$ 75.3). This difference is consistent with the view that teacher-forced imitation of specific matching patterns may be less robust when task distributions differ.

One possible explanation for the performance contrast is the difference in training dynamics. SFT encourages the model to imitate exact correspondence patterns from training data, which may create rigid associations that generalize poorly and can override useful pre-existing knowledge. In contrast, RL permits exploration of diverse matching strategies guided by holistic geometric feedback, which may help preserve prior capabilities while improving the target task. The dynamic curriculum may further support this process by progressively exposing harder scenarios rather than rely-

Matching Prompt

- You are given two images of the same physical scene, each having several regions annotated with circles and IDs (e.g., ”1”, ”2”, ”3”).Your task is to identify the underlying correspondence between regions in the two images. Note that **maybe not** every region in the first image has a match in the second.Please provide your response in **two parts**: thinking process and final answer, each wrapped by some special tags.
- `thinking process`: Your analysis where you show your analyzing and thinking process wrapped in `<thinking>` `</thinking>` tags, which includes but not limited to:
  - 1. **Describe Visual Regions**: For each annotated area in both images, describe the key visual characteristics and **spatial context within the scene**, including: - **Intrinsic properties**: color, shape, texture, size, material - **Spatial relationships in the 3D scene** (NOT pixel coordinates): * What objects/structures are directly adjacent to this region? (e.g., ”attached to a wooden door”, ”sitting on a metal shelf”) * What is this region positioned relative to in the physical space? (e.g., ”below the window”, ”behind the chair”, ”left side of the bookshelf”) * Semantic context: What functional area or object group does it belong to? (e.g., ”part of the dining area”, ”on the workspace desk”) - **Avoid** describing regions by their pixel locations (top-left, center-right, etc.) unless necessary for disambiguation - Focus on **scene-level landmarks** as reference points (e.g., ”near the entrance”, ”opposite to the main table”, ”in the corner with the lamp”)
  - 2. **Compare Viewpoints**: Analyze the geometric or perspective relationship between the two images, such as: - Overview of the two images’ contents, focusing on **how the physical scene layout appears** in each view - **Camera transformation**: rotation angle (e.g., ”camera rotated 90° clockwise around the room center”), translation (e.g., ”camera moved closer to the left wall”), zoom/scale differences - **Occlusion changes**: which objects/regions become visible or hidden due to viewpoint change? - **Perspective distortion**: how do spatial relationships appear to change due to different viewing angles? (e.g., ”objects on the right side now appear more frontal”)
  - 3. **Infer Matching Relations**: Based on region appearance and **scene-relative spatial relationships**, establish correspondences: - Iterate through each annotated region in image 1, and for each one, compare it sequentially with every annotated region in image 2 to infer the matching likelihood and reasonableness of region correspondences between the two images - Prioritize matching based on **what surrounds each region** and **3D spatial context** rather than 2D image positions - Use **stable scene anchors** (walls, large furniture, architectural features) to reason about region identity across views - Consider how the viewpoint change transforms the **spatial relationships** you identified in step 1 - Example reasoning: ”Region A-1 is next to a red door and below a window. Region B-3 is also adjacent to the same red door (now seen from a different angle) and below the same window structure, so A-1 matches B-3”
- `final answer`: Your final answer based on your thinking part wrapped by `<answer>` `</answer>` tags. The JSON object is a mapping of region IDs from image A (as string keys) to the corresponding region IDs in image B (as string values). **For regions in A that have no match in B, use ”none” as the value.**

Figure 4. Prompt For Wide-Baseline Matching Task.

ing on a single fixed difficulty level.

Across spatial benchmarks, DCRL improves over the base model on all four reported tasks (+43.0 on ReasonMatch, +5.3 on SAT, +5.3 on OmniSpatial, and +3.5 on MindCube). By contrast, SFT shows mixed behavior: it improves on ReasonMatch (+23.5) and MindCube (+5.1), is roughly flat on OmniSpatial (−1.0), and degrades substantially on SAT (−28.7). This pattern suggests that supervised imitation may be less robust for transferable spatial reasoning in this setting, while our RL approach produces gains that transfer more consistently across the reported geometric benchmarks.

Curriculum Learning. To assess the importance of our dynamic three-dimensional curriculum, we compare against three ablated training strategies: (1) uniform sampling without curriculum, (2) training only on easy samples (first quartile by viewpoint overlap), and (3) training only on hard samples (last quartile with minimal overlap). All variants use identical RL training setup and total training steps.

As shown in Table 9, our dynamic curriculum achieves 70.5% F1, outperforming uniform sampling (65.3%, +5.2 points). This result suggests that progressively increasing viewpoint divergence, correspondence complexity, and spatial distribution challenges can be more effective than treating all samples equally. A plausible explanation is that the curriculum allows the model to build useful intermediate competence on simpler configurations before tackling extreme viewpoint changes.

Training exclusively on easy or hard samples yields sig-

Table 9. Ablation study on curriculum learning. Progressive difficulty adjustment across viewpoint divergence and correspondence complexity improves performance over uniform sampling and fixed-difficulty training.

Training Strategy	**ReasonMatch**
No Curriculum (uniform cardinality)	65.3
Only 1/4 minimal Δ_v samples (Easy)	59.9
Only 1/4 maximum Δ_v samples (Hard)	62.3
Dynamic Curriculum (Ours)	**70.5**
Δ (Ours vs. No Curriculum)	+5.2
Δ (Ours vs. Easy)	+10.6
Δ (Ours vs. Hard)	+8.2

nificantly worse results. The easy-only strategy (59.9%) underperforms uniform sampling by 5.4 points, suggesting that limiting training to small viewpoint changes and simple correspondence structures is insufficient in this setup. Conversely, training only on hard samples (62.3%) also underperforms uniform sampling (-3.0 points), suggesting that starting with extreme viewpoint divergence and complex correspondences is also less effective than gradual difficulty scheduling.

The substantial gaps between curriculum and single-difficulty training (+10.6 over easy-only, +8.2 over hard-only) are consistent with the value of progressive learning. Our curriculum's adaptive adjustment mechanism monitors sustained performance improvements before increasing difficulty, helping the model stabilize before moving to harder settings. This staged progression across three dimensions (viewpoint span, correspondence complexity, and spatial distribution) is also consistent with the transfer gains observed on OmniSpatial, MindCube, and SAT.

Interestingly, hard-only training (62.3%) slightly outperforms easy-only (59.9%), suggesting that exposure to challenging cases, even without curriculum structure, may provide more useful learning signals than trivial examples. However, both approaches remain well below curriculum learning, indicating that progressive difficulty scheduling is beneficial in this setting.

9. ReasonMatch-Bench Results

In this section, we report detailed results on ReasonMatch-Bench and analyze how different multimodal LLMs behave under our cross-view region matching task. We first summarize the overall quantitative performance across scenarios and difficulty levels and then provide a fine-grained comparison along several error dimensions where models exhibit the largest discrepancies. Finally, we briefly discuss a visualization that highlights these differences.

9.1. Overall Quantitative Performance

Table 1 reports F1, Precision, and Recall on ReasonMatch-Bench for a broad range of multimodal LLMs, together with per-scenario scores on three difficulty levels (Indoor/Outdoor/Object, L1–L3).

ReasonMatch-Bench is challenging for strong general-purpose MLLMs. Even the best general-purpose models only achieve F1 scores in the 50–60 range on our benchmark. For example, GPT-5-mini obtains an overall F1 of 57.9, while GPT-5-Chat and Qwen3-VL-235B reach 51.5 and 49.2, respectively. Other strong closed-source systems such as Gemini-2.5-Pro, Claude-4.5-Sonnet, and GPT-4o remain in the 33–43 F1 range. This indicates that ReasonMatch-Bench is substantially harder than standard VQA-style benchmarks, even for large commercial MLLMs.

DCRL turns a weak open-weight model into the best performer. The baseline performs poorly on ReasonMatch-Bench (27.5 %), substantially below both commercial models and the larger Qwen3-VL-235B (49.2 %). After applying our DCRL training, the same 8B backbone achieves 70.5% F1, outperforming all other models in the table by a large margin. This suggests that targeted training on ReasonMatch-Bench-style tasks can, in this setting, be more effective than simply scaling up model size.

Performance consistently drops with difficulty, especially on Object L3. Across models, scores tend to decrease from L1 to L3 within each scenario, indicating that our difficulty annotation captures genuine changes in task hardness. For instance, GPT-5-mini drops from 68.2 to 47.0 on Indoor (L1→L3) and from 75.8 to 51.4 on Outdoor, and similar declines appear for other models. The Object L3 column is consistently the hardest setting, and our Qwen3-VL+DCRL model still struggles on this regime (33.7) compared to easier levels, but remains clearly above all baselines.

9.2. Fine-Grained Error Analysis Across Models

Although Table 1 reflects how often models select the correct match, it does not explain *why* they fail. To better characterize model behavior, we use Qwen3-VL-235B as a blind evaluator to score the complete ReasonMatch-Bench outputs of the five models shown in Figure 5. For each example, the evaluator receives the two annotated images, the ground-truth mapping, the verbatim reasoning trace, and the parsed final answer, and assigns discrete scores in $\{0, 1, 2\}$ to a broader rubric, where 2 indicates no issue and 0 indicates a severe issue. Below, we report the five dimensions

Table 10. Human study on the 90 largest view divergence samples across three datasets. Human annotators achieve substantially higher performance than all tested models, with the performance gap most pronounced on object-centric scenes (uCO3D).

Method	Overall			DL3DV			RE10K			uCO3D		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Closed-Source Models												
GPT-5-mini	34.5	40.7	37.2	32.5	40.4	35.9	47.4	52.7	49.7	23.6	29.1	25.8
Gemini-2.5-Pro	28.3	31.1	29.5	25.1	28.3	26.5	42.8	45.6	44.1	17.1	19.3	18.0
Claude-4.5-Sonnet	22.1	26.7	24.0	20.6	24.8	22.2	28.1	33.6	30.5	17.4	21.6	19.2
Open-Source Models												
Qwen3-VL-235B	27.5	33.0	29.9	23.2	28.1	25.3	42.3	50.2	45.7	17.1	20.9	18.7
Ours (DCRL)	**48.3**	**56.8**	**52.0**	**53.3**	**63.2**	**57.7**	**65.7**	**76.6**	**70.6**	**25.9**	**30.4**	**27.8**
Human	**87.5**	**81.3**	**84.0**	**94.3**	**93.3**	**93.5**	**96.1**	**93.9**	**94.7**	**72.2**	**56.6**	**62.1**
Δ (Human vs. Ours)	+39.2	+24.5	+32.0	+41.0	+30.1	+35.8	+30.4	+17.3	+24.1	+46.3	+26.2	+34.3

with the clearest cross-model differences: Local-Cue Reliance (F1), Global Layout Misalignment (F2), Reasoning-Answer Mismatch (F4), Overuse of "None" (F5), and Reasoning Coherence.

Local-Cue Reliance (F1). This score measures how well a model can integrate correct local observations into a globally consistent cross-view mapping. Low F1 reflects the typical failure pattern where the chain-of-thought describes the correct shelf, tier, or surrounding objects, but the final correspondence still points to a wrong region. Models with higher F1 scores are better at binding fine-grained visual cues into a coherent global decision.

Global Layout Misalignment (F2). F2 captures large-scale geometric errors such as swapped left/right ordering, incorrect vertical tiers, or inconsistent depth relations between regions. This dimension directly measures a model's ability to maintain a viewpoint-consistent global layout rather than reasoning in a purely local, patch-based manner.

Reasoning-Answer Mismatch (F4). F4 quantifies inconsistencies between the model's chain-of-thought and its final structured output. A low score indicates cases where the reasoning text identifies the correct target region, but the final JSON or `<answer>` block encodes a different ID or swaps indices. This dimension probes how reliably a model can translate its internal reasoning into a precise, machine-usable prediction.

Overuse of "None" (F5). ReasonMatch-Bench includes regions that legitimately have no correspondence in the other view. F5 measures whether the model uses the "none" option in a calibrated way. Over-using "none"—especially on regions that do have correct matches—reveals an overly conservative behavior: the model can often describe the local content but refuses to commit to a specific correspondence.

Reasoning Coherence. Finally, we rate the overall logical coherence of the chain-of-thought: whether observations are ordered in a reasonable way, whether there are self-contradictions, and whether the final decision is clearly connected to previous observations. This dimension complements F1/F2 by highlighting cases where the answer is correct but the reasoning is disorganized, as well as cases where an apparently detailed reasoning trail actually does not support the final mapping.

Figure 5 summarizes these five dimensions across all evaluated models. We report mean rubric scores in $[0, 2]$, so that higher values consistently correspond to better behavior (fewer or milder errors). Several patterns emerge:

- Models differ most strongly on **F1** and **F2**, suggesting that binding local cues into a globally consistent spatial alignment is the central bottleneck on ReasonMatch-Bench.
- **F4** reveals that some models frequently "think correctly but answer incorrectly", exposing weaknesses in reasoning-to-action alignment that are invisible to raw F1 / Precision / Recall.
- Differences in **F5** highlight distinct calibration behaviors: some models are overly cautious and tend to abstain by predicting "none", while others are more willing to commit to concrete correspondences.
- The **logical coherence** scores show that even when overall accuracy is similar, models can vary substantially in how structured and self-consistent their reasoning is.

Overall, these error dimensions provide a complementary perspective to the aggregate metrics in Table 1. They show that current multimodal LLMs are limited not only by perception, but also by their ability to maintain a globally

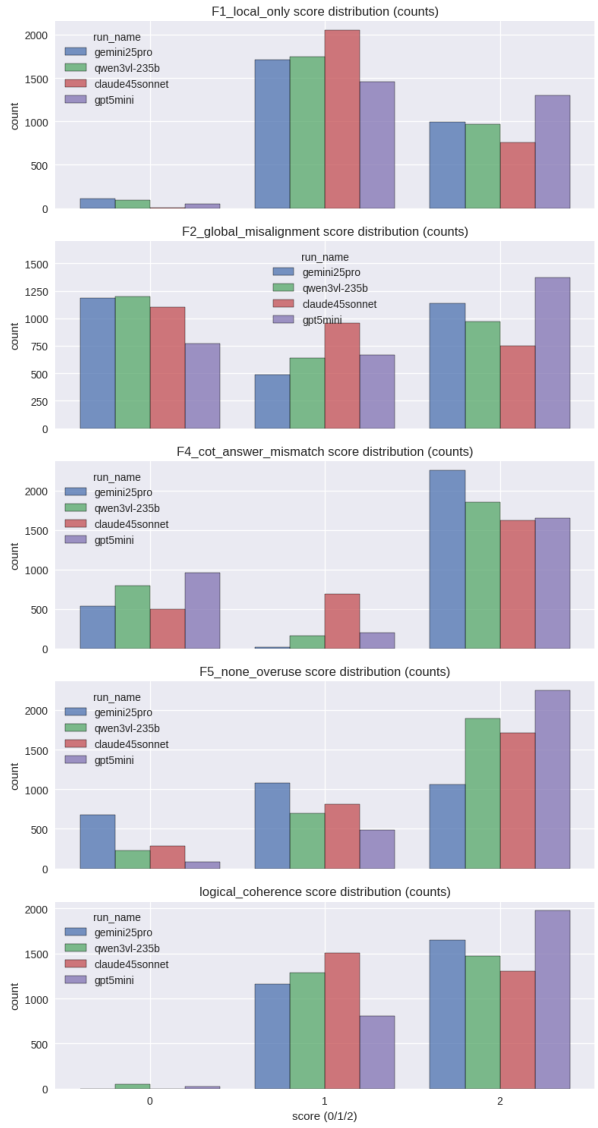


Figure 5. Comparison of five models across key error dimensions: Local-Cue Reliance (F1), Global Layout Misalignment (F2), Reasoning-Answer Mismatch (F4), Overuse of "None" (F5), and Reasoning Coherence. Higher bars indicate fewer or milder errors.

consistent 3D picture of the scene, propagate that picture through a coherent reasoning process, and translate it into a precise region-level correspondence.

Qualitative examples. To make the above error dimensions more concrete, we highlight a representative case for `F2` from ReasonMatch-Bench. In a ScanNet indoor scene (`scannet_000751`), several floor markers form a small cluster that appears in both views under essentially the same physical configuration. The ground truth correspondence in Image B is region B-1, which is the leftmost marker in this cluster. However, the model often selects region B-3 and describes it as the leftmost marker. This behavior indicates that the model has roughly identified the correct cluster, but fails to maintain a consistent left–right ordering when projecting the 3D layout into the two viewpoints, which is precisely what `F2` is designed to capture.

9.3. Visualization of Model Behaviors

To make the above differences more concrete, Figure 5 provides a compact visual summary of model behaviors along five error dimensions: Local-Cue Reliance (F1), Global Layout Misalignment (F2), Reasoning-Answer Mismatch (F4), Overuse of "None" (F5), and Reasoning Coherence.

This visualization highlights which aspects of spatial reasoning each model handles well and where it fails:

- Models with high scores on F1 and F2 are able to build a coherent 3D understanding from two wide-baseline views, rather than relying on local texture cues alone.
- Low F4 and logical coherence expose brittle reasoning pipelines, where correct evidence is not reliably turned into the correct structured prediction.
- Variations in F5 reveal different risk profiles: some models trade recall for precision by overusing "none", while others achieve better coverage at the cost of more misaligned matches.

In combination with the quantitative scores in Table 1, this figure shows that ReasonMatch-Bench is not just a harder matching dataset, but also a targeted probe of multi-view spatial understanding, calibration, and reasoning quality in multimodal LLMs.

To further illustrate how models behave on individual instances, Figures 6 and 7 show full chain-of-thought traces and JSON outputs on two representative successful examples from ReasonMatch-Bench.

9.4. Human Study

To assess how the task aligns with natural human spatial reasoning and to measure the gap between human performance and current models, we conduct a human study on a stratified 90-sample subset with the largest view divergence. Two non-expert annotators, i.e., ordinary participants without task-specific training, complete the task independently using the same instructions as the models. They do not discuss their answers during annotation, and we report the average of their scores as the final human result.

Despite not having specialized training, the human annotators achieve an overall F1 of 84.0%, substantially outperforming all tested models, including frontier systems (Table 10). The 32-point gap between the non-expert annotators (84.0%) and our best model (52.0%) highlights a large remaining challenge in fine-grained geometric correspondence.

Performance patterns differ markedly across scene types.

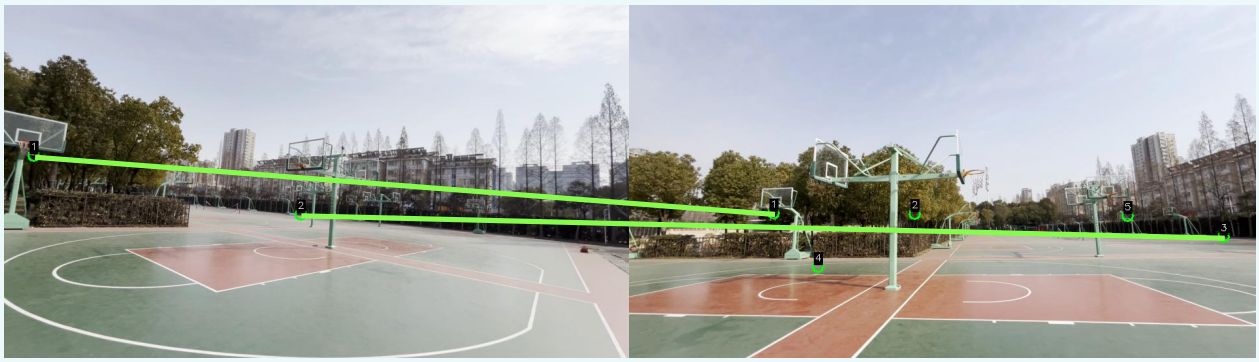
On structured indoor and outdoor environments, humans achieve near-perfect accuracy (F1 $>$ 93%), likely aided by rich contextual cues and stronger geometric regularity. Both humans and models struggle more on object-centric scenes (uCO3D), where humans achieve 62.1% F1 compared with 27.8% for our best model. Analysis of human annotations suggests that errors primarily occur when objects exhibit severe self-occlusion or lack distinctive surface features, which are also challenging conditions for current models. This pattern is consistent with the view that the benchmark captures genuine spatial reasoning challenges rather than artifacts of imperfect ground truth.

The human-model gap underscores substantial limitations in current MLLMs' spatial reasoning. While DCRL improves over strong baselines (+14.8% over GPT-5-mini) and reaches 62% of non-expert human performance overall, there remains considerable room before models approach human-level geometric reasoning. The fact that non-expert annotators can solve many of these tasks reliably further motivates the benchmark and continued work on spatial understanding in MLLMs.

10. Limitations and Future Work

Despite substantial improvements over baselines, our approach reveals both the promise and challenges of developing human-level spatial intelligence in MLLMs. Our best model achieves 52.0% F1 compared to untrained humans' 84.0%, with the gap most pronounced on object-centric scenes where even humans struggle (27.8% vs. 62.1%). This persistent performance gap highlights fundamental limitations in current architectures' ability to perform fine-grained geometric reasoning under extreme viewpoint changes and limited contextual cues.

Our work focuses on pairwise cross-view matching as a foundational capability, but comprehensive spatial intelligence demands reasoning over multiple views simultaneously, integrating geometric correspondence with 3D scene understanding, temporal dynamics, and semantic knowledge. Future research should extend beyond pairwise matching toward holistic multi-view reasoning that mirrors human spatial cognition—synthesizing information across viewpoints to construct coherent 3D mental models.



Two images show the same scene from different viewpoints. Match the region IDs from Image A to Image B

<thinking>
The two images depict the same outdoor basketball court from different camera angles. The first image is a wide shot from the side of the court, while the second image is a closer, more centered perspective looking down the length of the court.
Region 1 in the first image is the backboard and hoop assembly on the far left side of the court, near the hedge. In the second image, this same structure is now visible from the front, and is labeled as region 1.
Region 2 in the first image is the base of a basketball hoop on the far side of the court, near the fence. In the second image, this same base is visible from the front and is labeled as region 3.
The second image provides a clearer view of the court's layout, showing multiple hoops and their bases. Region 4 in the second image is the base of the hoop on the left side of the court, which corresponds to the base of the hoop in the first image (region 1).
Region 5 in the second image is the backboard and hoop assembly on the far right side of the court, which corresponds to the hoop in the first image (region 2).
The spatial relationships are consistent: the hoop bases are positioned along the court's perimeter, and the backboards are mounted on top of the poles. The camera has rotated to face down the court, making previously side-viewed objects now appear more frontal.
</thinking>
<answer>
```json
{ "1": "1", "2": "3" }
```
</answer>

Figure 6. Qualitative example on ReasonMatch-Bench (Example 1). The model is given two views of the same scene and asked to match region IDs from Image A to Image B. Its chain-of-thought correctly identifies the corresponding regions across viewpoints and produces a consistent JSON mapping for the cross-view region matching task.

Two images show the same scene from different viewpoints. Match the region IDs from Image A to Image B

<thinking>
The two images depict the same industrial ceiling scene from different camera angles. The first image is a wider shot looking up at the ceiling, while the second image is a closer, more centered view of the same area, showing more details of the crane and ceiling structure.
Region 1 in the first image is a small green circle on the corrugated metal ceiling, located near the bottom right of the frame, adjacent to a long fluorescent light fixture. In the second image, this same region is now visible as region 3, which is also on the ceiling, directly below the crane's hoist cable and near the same light fixture.
Region 2 in the first image is a green circle on the red crane beam, specifically highlighting the point where the hoist motor is attached. In the second image, this corresponds to region 1, which is also on the crane beam, directly above the hoist motor and next to the Chinese characters "品质优先" (Quality First).
Region 3 in the second image is a green circle on the ceiling, located directly below the crane's hoist cable. This region is not present in the first image.
Region 4 in the second image is a green circle on the ceiling, located above the crane beam and near the top of the frame. This region is not present in the first image.
Region 5 in the second image is a green circle on the ceiling, located directly below the crane's hoist cable and near the same light fixture as region 3. This region is not present in the first image.
Based on the spatial relationships and scene context, the matching is as follows:
- Region 1 in the first image matches region 3 in the second image, as they are the same physical location on the ceiling, viewed from different angles.
- Region 2 in the first image matches region 1 in the second image, as they are the same physical point on the crane beam, viewed from different angles.
</thinking>
<answer>
```json
{ "1": "3", "2": "1" }
```
</answer>

Figure 7. Qualitative example on ReasonMatch-Bench (Example 2). Another successful case where the model aligns regions across wide-baseline views, using textual reasoning over geometric layout and shared objects to produce a correct JSON correspondence. Together with Figure 5, these examples illustrate how multimodal LLMs solve our pixel-anchored cross-view matching task when their spatial reasoning is reliable.