LR-BENCH: DIAGNOSING AND MITIGATING LEFT-RIGHT CONFUSION IN MLLMS FOR HUMAN-CENTRIC SPATIAL UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Human subject-centered left-right (LR) discrimination, i.e., answering "which is *his* left hand" in the depicted person's own frame of reference (FoR) rather than in screen coordinates, is a fundamental capability for human-centric vision-language applications. We find that current multimodal large language models (MLLMs) fail at this task in a systematic way. Most open-source models score near chance, and for most mirror pairs their prediction stays the same after the input is horizontally flipped, which means the errors are systematic rather than random. To diagnose this failure, we introduce **LR-Bench**, a benchmark of 8,048 human-annotated binary QA pairs over 2,942 in-the-wild images and 1,809 video clips, spanning appearance, action & pose, spatial relation, and temporal dynamics, and evaluated with binary accuracy, a mirror-paired consistency score, and an LLM-judged caption protocol. Beyond evaluation, we propose **Spatial Latent Prompt (SLP)**, a plug-in that injects an input-derived human structural prior, a Mesh-Derived Part Map from an off-the-shelf mesh recovery model, into frozen visual tokens through learnable part embeddings and a person-relative positional channel. With fewer than 0.1M trainable parameters and a frozen backbone, SLP improves three MLLM backbones by 36.9–48.7 points on images and 27.9–35.8 points on videos, and clearly outperforms full-parameter SFT and LoRA trained on the same data. The gain therefore comes from how the structural prior is represented and injected, not simply from in-domain supervision. SLP thus establishes a lightweight, parameter-efficient paradigm for human-centric spatial perception in MLLMs.

1 INTRODUCTION

Human-centric spatial understanding is a fundamental topic in visual perception, and human subject-centered left-right (LR) discrimination is one of its most basic yet challenging components. Unlike viewer-centric spatial perception (e.g., "on the left side of the image"), subject-centered LR understanding requires anchoring spatial semantics in the human subject's own frame of reference (FoR) (e.g., "*his* left hand"). It matters for human-centric generation, sports analysis, healthcare, and embodied instruction following; e.g., LR hallucinations in MLLM captions become noisy supervision for human-centric generative models (Agrawal et al., 2026).

Although humans switch between FoRs effortlessly, most leading MLLMs fail systematically on this task: on our benchmark, most open-source models score near the 50% chance level, and for most mirror pairs their answer stays the same after the input is horizontally flipped, so a mirror-paired consistency score drops to 8–12%, well below the 25% expected from random guessing (Sec. 5.3). The errors are thus systematic rather than random, and they come from dataset priors and from answering in the viewer's frame. As illustrated in Figure 1, we analyze three underlying causes: 1) **Inconsistent Spatial Labels** (Figure 1a): VQA and caption datasets mix viewer-centric labels (screen-left/right) with subject-centered (anatomical) conventions, so a phrase like "the left side of this woman" is ambiguous and produces contradictory supervision. 2) **Weak Structural Laterality Cues** (Figure 1b): A viewpoint-balanced linear probe on frozen Qwen3-VL-8B (Bai et al., 2025) part features achieves near-chance accuracy, indicating that explicit anatomical laterality is not readily accessible from pretrained visual representations. 3) **Fragile Subject-Centered Reasoning** (Fig-

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

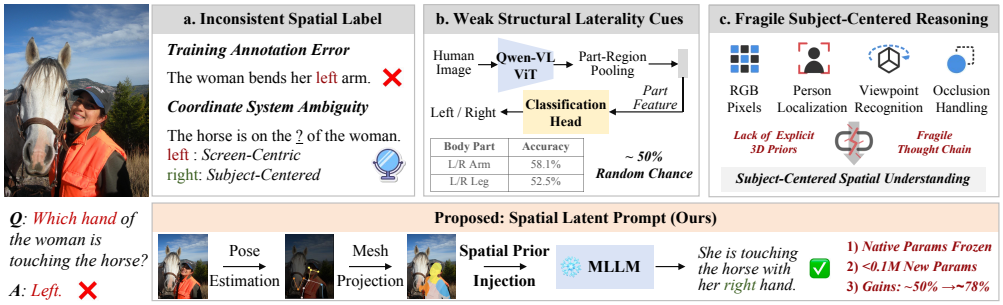


Figure 1: Analysis of human subject-centered LR confusion in MLLMs. Ambiguous LR references, weak laterality cues in visual representations, and fragile subject-centered reasoning jointly make this task challenging. SLP mitigates this issue by injecting human structural priors into frozen MLLM visual tokens.

ure 1c): Without an explicit prior about the body's 3D structure, the model must implicitly chain person localization, viewpoint recognition, and occlusion handling to infer laterality, a chain that collapses easily in the wild. All three causes suggest the same remedy, namely providing the model with an explicit human structural prior.

Previous visual-spatial benchmarks predominantly anchor spatial semantics to the viewer, to inanimate objects, or to the environment (Table 1). Left-right confusion has only been touched on briefly, for example on synthetic object scenes where the FoR is given in the prompt (Zhang et al., 2025), or as one grounding sub-task in a general human-centric benchmark (Liu et al., 2026b). No existing benchmark focuses on subject-centered laterality of *in-the-wild humans* in both *images and videos*, covers the temporal dynamics and person-relative *spatial relations* that make the problem difficult, or uses a mirror-paired protocol to separate real subject-centered reasoning from dataset priors, and none proposes a solution (Sec. 2.2).

To bridge this gap, we introduce **LR-Bench**, a benchmark dedicated to human subject-centered LR discrimination: 8,048 QA pairs from 4,751 sources (3,299 image and 4,749 video QA pairs) across four question types, with every label assigned by human annotators and the label distribution balanced. We evaluate with binary accuracy, a *Mirror-Consistency Score* that adapts the flip-paired protocol of FlipEval (Ma et al., 2025) to subject-centered laterality, and an LLM-judged caption protocol. We then propose **Spatial Latent Prompt (SLP)**, a plug-in that injects an input-derived human structural prior, a *Mesh-Derived Part Map* (MPM) rendered from an off-the-shelf mesh recovery model, into the visual tokens of a frozen MLLM through learnable part embeddings (a semantic channel) and a person-relative positional encoding for background tokens (a positional channel). Across three backbones, SLP improves accuracy by 36.9–48.7 points on images and 27.9–35.8 points on videos with fewer than 0.1M trainable parameters. Notably, full-parameter SFT and LoRA trained on the *same* LR data reach only 59–62%, compared with 79% for SLP, which shows that the gain comes from how the structural prior is represented and injected rather than from in-domain supervision. Our contributions are:

- We introduce **LR-Bench**, a benchmark explicitly designed to diagnose and evaluate human subject-centered LR awareness in MLLMs. Spanning diverse image and video sources across varied scenarios, it provides high-quality, multi-dimensional QA pairs and introduces robust evaluation protocols for comprehensive assessment.
- Methodologically, we propose a lightweight plug-in method as **Spatial Latent Prompt (SLP)**. SLP first derives novel Mesh-Derived Part Map representations, indicating subject-centered spatial layout, and then injects them into visual tokens as latent prompts through dual semantic-position encodings, to enable LR awareness in MLLMs.
- Extensive evaluations on LR-Bench reveal that current MLLMs suffer from serious LR confusion in subject-centered discrimination. In contrast, SLP brings significant improvements across three different backbones with minimal training and parameter (¡0.1M) cost.

Table 1: Comparison of LR-Bench with existing spatial understanding benchmarks. "Subject-centric" means left/right is defined in the depicted subject's own frame.

Benchmark	Modality	Data Scale	FoR	Core Spatial Focus
EgoOrientBench (Jung et al., 2025)	Image	8,365 images	Viewer-centric	Object orientation
3DSRBench (Ma et al., 2025)	Image	2,772 QAs (flip-paired)	Camera-centric	3D location / orientation / height
DORI (Tasnim et al., 2026)	Image	13,652 images / 33,656 QAs	Object-centric	Object orientation
COMFORT (Zhang et al., 2025)	Image (synthetic)	1,220 images	Multiple FoRs	FoR & cross-lingual ambiguities
VSI-Bench (Yang et al., 2024a)	Video	288 videos / 5,060 QAs	Environment-centric	General visual-spatial intelligence
SpatialBench (Xu et al., 2025)	Video	117 videos / 3,193 QAs	Environment-centric	Hierarchical spatial cognition
Human-MME (Liu et al., 2026b)	Image	LR as a sub-task	Subject-centric (grounding)	General human-centric understanding
LR-Bench (Ours)	**Image & Video**	**4,751 sources / 8,048 QAs**	**Subject-centric**	**Subject-centered LR understanding**

2 Related Work

2.1 Spatial Understanding in MLLMs

Modern MLLMs achieve strong general vision-language performance (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023b; Dai et al., 2023; Bai et al., 2023), yet persistently struggle with spatial understanding such as compositional relations and fine-grained localization (Yuksekgonul et al., 2023; Liu et al., 2023a; Kamath et al., 2023). Existing works strengthen spatial grounding via region-level alignment, pixel-level grounding, or explicit 3D-scene supervision (Li et al., 2022; Peng et al., 2024; Chen et al., 2024; Cheng et al., 2024; Li et al., 2024; Liu et al., 2025b; Hong et al., 2025). Human-centric LR discrimination is a distinct subject-centered challenge: the same image location denotes opposite anatomical sides depending on body orientation, viewpoint, and occlusion, so general 3D scene reasoning is insufficient.

SLP is related to methods that add small learnable inputs to a frozen VLM. PIN (Dorkenwald et al., 2024) inserts a learned, input-independent positional prompt into frozen visual features to enable localization, and SpatialVLM and SpatialRGPT (Chen et al., 2024; Cheng et al., 2024) feed depth or 3D-scene cues as extra inputs. Different from them, the prompt in SLP is *input-conditioned*, derived from a human mesh for each image, and *anatomically structured*, encoding part identity and person-relative position instead of a global positional bias.

2.2 Visual-Spatial Intelligence Benchmarks

As summarized in Table 1, most spatial benchmarks (Tasnim et al., 2026; Jung et al., 2025; Kamath et al., 2023; Yang et al., 2024a; Ma et al., 2025) anchor coordinates to the observer's screen (*viewer-centric*), to the canonical orientation of rigid objects (*object-centric*), or to the environment (*environment-centric*). Human anatomical laterality is fundamentally different: rigid objects have lateral components bound to a canonical orientation, whereas the articulated human body has limbs that cross the midline or become occluded, so LR discrimination requires part-level grounding under pose, occlusion, and viewpoint variation.

A few works touch on human laterality. COMFORT (Zhang et al., 2025) reports that models struggle to adopt an intrinsic FoR when asked to, but on synthetic scenes that relate an object to a reference object, with the FoR given in the prompt. Human-MME (Liu et al., 2026b) is a general human-centric benchmark whose LR items are grounding-based, asking the model to localize a named part such as the left hand instead of answering a natural LR question. The concurrent MirrorQA (Liu et al., 2026a) studies subject-centered left-right questions on still images. In contrast, LR-Bench focuses on subject-centered LR discrimination of in-the-wild humans in both *images and videos*. It includes temporal dynamics, which only video can express, and person-relative *spatial relations* to external objects, is stratified by visibility, viewpoint, and person count, comes with a mirror-paired protocol, and, unlike all of the above, is paired with a token-level mitigation method. Paired scoring itself is not new: FlipEval (Ma et al., 2025) pairs each question with its horizontally flipped version to remove laterality priors, and CircularEval (Liu et al., 2024) counts an item as correct only if all permuted variants are answered correctly. Our Mirror-Consistency Score applies this idea to subject-centered human laterality and, together with the same-answer rate, uses it to explain why models fail (Sec. 5.3).

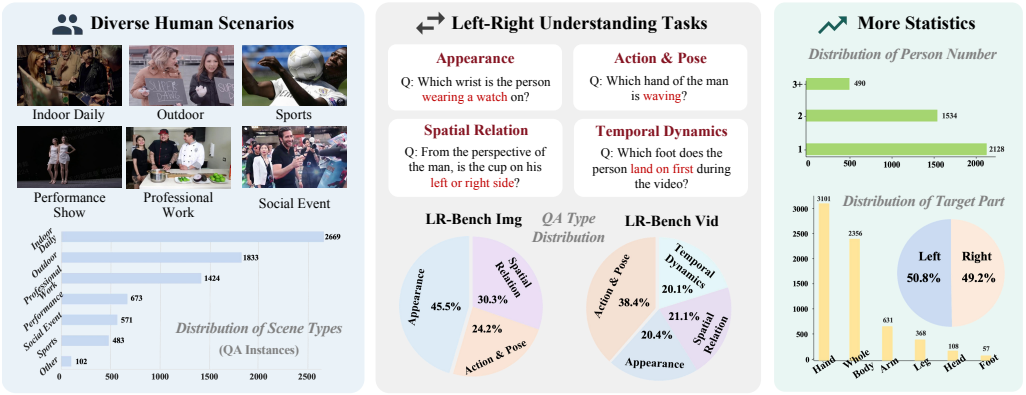


Figure 2: Overview of LR-Bench. The benchmark covers image and video sources, four LR question types, and diverse human scenarios, with all annotations following the subject-centered FoR rather than screen coordinates.

3 LR-Bench: Benchmarking Subject-Centered LR Discrimination

3.1 Benchmark Overview

LR-Bench comprises **8,048 QA pairs** over **4,751 sources**, split into LR-Bench$^{\text{Img}}$ (2,942 images, 3,299 QA pairs) and LR-Bench$^{\text{Vid}}$ (1,809 video clips, 4,749 QA pairs). All annotations follow the subject-centered FoR: left and right are defined from the target human subject, not the observer. Every question has a binary left/right ground truth and belongs to one of four categories: **Appearance (Appear.)**, grounding a visible attribute to the correct side (*"Which wrist is the person wearing a watch on?"*); **Action & Pose (A & P)**, identifying which limb performs an action or pose (*"Which hand of the man is waving?"*); **Spatial Relation (Spatial)**, localizing an object or another person in the target's FoR (*"From the perspective of the man, is the cup on his left or right side?"*); **Temporal Dynamics (Temporal, video only)**, tracking LR-specific changes over time (*"Which foot does the person land on first?"*). We deliberately keep the answer space binary. Chance performance is then exactly 50%, the mirror-paired protocol below is well defined, and annotation stays unambiguous, since questions whose true answer would be "both" or "neither" are removed during curation. LR-Bench is therefore a focused test of laterality discrimination, which general human-centric benchmarks only report as part of aggregate scores.

3.2 Data Curation Pipeline

Source collection. We collect 1,809 video clips and 2,942 images from public datasets (Liu et al., 2025a; Zheng et al., 2024; Li et al., 2025; 2024; Hong et al., 2025; Xing et al., 2025; Schuhmann et al., 2021), stratified by body visibility, camera viewpoint, person count, and scene type. SLP's training data (Sec. 4.3) is drawn from disjoint source pools.

Question pre-generation without answers. For each source, Qwen3-VL-235B-A22B (Bai et al., 2025) proposes candidate LR questions *without answers*. This filters sources lacking salient LR content and reduces annotator workload, while no model-produced laterality judgment enters the ground truth. The candidates only serve as a starting point. About 70% of the final questions were rewritten or newly written by annotators, and every answer is given by a human.

Two-pass human annotation. First-pass annotators (i) remove incorrect or hallucinated questions, (ii) remove ambiguous questions lacking a prominent LR feature or whose answer would be "both"/"neither", (iii) add LR questions missed by pre-generation, and (iv) rewrite phrasing to be unambiguously subject-centered ("the man's left hand", not "the hand on the left"), then assign the label. A second, independent annotator then re-answers and re-checks every item. The two passes agree on 94% of the items, and the disagreements are resolved by a further check.

3.3 Benchmark Statistics

Video clips range from 2 s to 1 min (over 90% shorter than 20 s), yielding 2.63 QA pairs per source versus 1.12 for LR-Bench$^{\text{Img}}$. As shown in Figure 2, the benchmark spans varied scenes (predominantly *indoor daily* and *outdoor*), camera viewpoints (front 48.6%, side 41.1%, other 10.3%), a wide range of person counts and queried targets (most frequently *hand* and *whole body*), and balanced labels (50.8% left vs. 49.2% right). These attributes are used again in Sec. 5.6 to analyze where the gains of SLP come from.

3.4 Evaluation Protocols and Metrics

1) LR binary QA accuracy over all QA pairs, where chance is 50%.

2) Mirror-Consistency Score (MCS) and same-answer rate. Accuracy alone is insufficient: a model exploiting dataset priors (watches are usually on the left wrist) or answering in screen coordinates can score well above zero without following the subject's FoR. Following FlipEval (Ma et al., 2025), we pair each item with its horizontally mirrored version. Mirroring swaps the subject's left and right, and hence the ground truth, while the question stays the same. As in CircularEval (Liu et al., 2024), an item is counted as correct only if both versions are answered correctly: $\mathrm{MCS} = \frac{1}{N}\sum_i c_i\, c_i^{\mathrm{flip}}$, with $c_i, c_i^{\mathrm{flip}} \in \{0, 1\}$ the correctness indicators. Two independent random guesses give an expected MCS of 25%. This is a reference value rather than a lower bound, because two kinds of failures both score close to zero: a model that answers from image-independent priors gives the *same* answer to both versions and is correct on exactly one, while a model that answers in the viewer's frame is wrong on *both* whenever the subject faces the camera. To tell them apart, we also report the *same-answer rate*, the fraction of mirror pairs that receive the same answer, which reflects prior-driven, mirror-*invariant* answering. The task instead requires a high MCS and a low same-answer rate.

3) Caption-based evaluation. To test whether LR perception carries over to open-ended generation, we prompt each model for a detailed human-centric caption with explicit LR details (Appendix A) and use an LLM judge to label each benchmark QA pair against the caption as *correct*, *wrong*, or *not mentioned*. With counts $N_{\mathrm{c}}, N_{\mathrm{w}}, N_{\mathrm{m}}$ ($|\mathcal{Q}| = N_{\mathrm{c}} + N_{\mathrm{w}} + N_{\mathrm{m}}$), we report $\mathrm{Coverage} = (N_{\mathrm{c}} + N_{\mathrm{w}})/|\mathcal{Q}|$, i.e., how often the caption mentions a side at all, and $\mathrm{Precision} = N_{\mathrm{c}}/(N_{\mathrm{c}} + N_{\mathrm{w}})$, i.e., how often the mentioned side is correct. The judge is Qwen3-235B-A22B (Yang et al., 2025), a *text-only* LLM. It only sees the caption, the question, and the ground-truth answer, never the image. Judging is thus a different task from answering LR questions: it only checks whether the caption text agrees with the given answer and needs no visual LR ability, so even a model that fails on LR-Bench would be an unbiased judge (and our judge is in any case a different model from the Qwen3-VL-235B-A22B baseline in Table 2).

4 Method

As analyzed in Sec. 1, LR confusion stems primarily from the absence of explicit human structural priors in pretrained visual representations, compounded by noisy LR supervision. We address both with **Spatial Latent Prompt (SLP)**, which injects part-identity information into the visual tokens of a frozen backbone (Figure 3), trained on a curated LR dataset.

4.1 Mesh-Derived Part Map as Spatial Prior

The central design question is *which* representation of human structure to inject. **Human parsing maps** (Khirodkar et al., 2024; Yang et al., 2024b) are dense but label regions by clothing category ("upper clothes", "pants"), which discards limb identity and anatomical LR. **Pose keypoint heatmaps** carry explicit LR semantics (L_wrist, R_wrist) and are robust to occlusion, but their sparse keypoints do not align with the dense token grid. Our **Mesh-Derived Part Map (MPM)** combines the strengths of both: dense coverage, explicit anatomical LR semantics, and improved handling of partial occlusion through a 3D body prior.

Construction. We run an off-the-shelf human mesh recovery model, SAM3D-Body (Yang et al., 2026b), accelerated with Fast SAM3D-Body (Yang et al., 2026a). Each mesh vertex is pre-assigned

to one of six foreground part categories, i.e., head, torso, left/right arm (shoulder to wrist), and left/right leg (hip to ankle). We project the labeled mesh onto the image plane and label uncovered pixels as *background*, giving a pixel-level part map $M \in \{0, \dots, 6\}^{H_m \times W_m}$ with $K = 7$ categories in total. For videos, we apply the same procedure *independently* to each of the 16 frames fed to the MLLM. Since laterality is defined per person and per frame, no cross-frame identity tracking is needed. When several people appear in a frame, their part maps are merged into one frame-level MPM in decreasing order of mask area. We do not select a "target" person, so the MPM carries no information about which person the question refers to; the MLLM has to identify the queried person from the question and the visual content, exactly as the baseline does (Appendix B.2).

MPM is a structural prior, not answer leakage. Because the MPM encodes anatomical left/right regions, one may ask whether it leaks the answer. It does not. *(i)* It is an input-derived feature computed on-the-fly at inference by an off-the-shelf model, available for any unseen sample at deployment, like the depth or 3D-scene cues of prior spatial VLMs (Chen et al., 2024; Cheng et al., 2024); no test-time ground truth or benchmark annotation is used anywhere in the pipeline, including the identity of the queried person. *(ii)* It supplies the *frame of reference*, i.e., where the subject's left and right sides are and hence which way the subject faces, but nothing about the queried content (which wrist bears the watch, where the cup is, what happens first); the model must still perceive the target and bind it to the correct side. For Spatial-Relation questions, the MPM says nothing about the external object, but the facing direction it provides is what allows a screen-space offset to be converted into a subject-centered one. This is why the semantic channel alone already helps on this category (Sec. 5.5). *(iii)* Merely exposing the same L/R labels as visual prompts yields only marginal gains (Sec. 5.4); the benefit comes from *how* the prior is injected.

4.2 SLP: Dual Semantic-Position Encoding

SLP adds two channels to the visual tokens, a **semantic channel** encoding body-part identity and a **positional channel** that re-expresses background tokens relative to the human. For the i-th visual token $\mathbf{v}_i \in \mathbb{R}^D$,

$$\hat{\mathbf{v}}_i = \mathbf{v}_i + \boldsymbol{\delta}_i^{\text{sem}} + \boldsymbol{\delta}_i^{\text{pos}}, \quad i = 1, \dots, N. \tag{1}$$

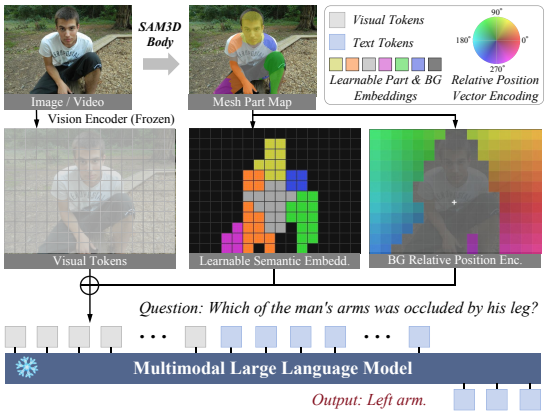


Figure 3: Overview of SLP: an MPM from off-the-shelf mesh recovery is injected into frozen visual tokens via a semantic channel (part embeddings) and a positional channel (person-relative offsets for background tokens).

1) Part weights with lateral sharpening. For each token we average-pool the MPM over its receptive patch to obtain an occupancy vector $\mathbf{w}_i \in \Delta^K$ whose k-th entry is the fraction of the patch covered by category k (background included). Tokens on body-part boundaries mix a lateral part (L/R arm or leg) with non-lateral regions, diluting the laterality signal. Let $\mathcal{L} = \{\text{L-arm}, \text{R-arm}, \text{L-leg}, \text{R-leg}\}$ denote the lateral categories. Whenever a token contains any lateral part, we keep only the lateral weights and renormalize them:

$$\tilde{w}_{ik} = \begin{cases} w_{ik}\,\mathbb{I}[k \in \mathcal{L}] \,/\, \sum_{k' \in \mathcal{L}} w_{ik'}, & \text{if } \sum_{k' \in \mathcal{L}} w_{ik'} > 0, \\ w_{ik}, & \text{otherwise.} \end{cases} \tag{2}$$

This step has no parameters, and tokens that contain only head, torso, or background are left unchanged.

2) Part-aware semantic encoding. With K learnable part embeddings $\mathbf{E} \in \mathbb{R}^{K \times D}$ (one per category including background), $\boldsymbol{\delta}_i^{\text{sem}} = \tilde{\mathbf{w}}_i \mathbf{E}$. $\mathbf{E}$ is zero-initialized so that training starts from the unmodified backbone and the prior is introduced gradually.

3) Person-relative position encoding for background tokens. Native token positions are expressed in absolute image coordinates, agnostic to where the human is. We instead re-express each background token's position relative to the centroid $(\bar{r}, \bar{c})$ of all foreground (human) tokens: with

Table 2: LR binary QA accuracy (%) on LR-Bench. Video inputs are 16 uniformly sampled frames, and chance is 50%.

Model	LR-Bench^{Vid}					LR-Bench^{Img}			
	Overall	Appear.	A & P	Spatial	Temporal	Overall	Appear.	A & P	Spatial
Closed-source models									
Gemini-3-Flash (Google, 2026)	61.30	70.76	64.33	52.90	54.71	70.35	73.47	64.83	70.10
Gemini-2.5-Pro (Team et al., 2024)	53.02	61.23	55.99	37.50	50.73	54.53	61.00	59.82	40.60
GPT-4.1 (Achiam et al., 2023)	39.90	34.40	43.45	30.20	48.85	37.28	41.53	41.55	27.50
Open-source models									
Qwen3-VL-32B (Bai et al., 2025)	54.31	57.64	65.04	25.50	60.56	44.68	55.80	54.57	20.10
Qwen3-VL-235B-A22B (Bai et al., 2025)	57.89	62.60	71.18	27.30	59.73	51.59	66.93	60.95	21.10
MiniCPM-V-4.5 (Yu et al., 2025)	48.26	51.65	53.32	32.10	52.09	39.50	45.53	45.93	25.30
LLaVA-OV-1.5-8B (An et al., 2025)	47.69	34.50	60.27	30.90	54.60	37.04	39.47	45.81	26.40
Ovis2.5-9B (Lu et al., 2025)	48.37	56.30	52.27	26.70	55.54	42.68	51.20	51.94	22.50
MiMo-VL-7B-SFT (Team et al., 2025a)	49.80	47.73	59.73	26.10	57.74	43.86	54.07	51.94	22.10
InternVL3.5-14B (Wang et al., 2025)	49.08	62.09	52.82	32.40	46.23	46.68	58.93	52.19	23.90
Frozen backbones with SLP									
Qwen3-VL-8B (Bai et al., 2025)	50.16	54.96	60.05	25.00	52.72	41.68	51.67	51.19	19.10
+ **SLP (Ours)**	**78.69**	**84.09**	**86.74**	**71.20**	**65.69**	**79.48**	**81.47**	**84.36**	**72.60**
InternVL3.5-8B (Wang et al., 2025)	50.68	60.23	54.68	28.70	56.38	42.38	52.60	47.81	22.70
+ **SLP (Ours)**	**78.54**	**86.67**	**89.53**	**61.50**	**67.15**	**79.23**	**84.33**	**89.36**	**62.76**
GLM-4.1V-9B (Team et al., 2025b)	41.44	44.73	50.74	23.20	39.44	31.71	38.87	35.04	18.30
+ **SLP (Ours)**	**77.24**	**84.50**	**87.12**	**65.30**	**63.49**	**80.45**	**82.87**	**86.11**	**72.30**

$(r_i, c_i) \in [0,1]^2$ the token's normalized coordinates, the offset $\boldsymbol{\rho}_i = (r_i - \bar{r},\ c_i - \bar{c})$ is encoded with a sinusoidal basis $\gamma(\cdot)$ (L=4 bands per axis, see Appendix B.1) and projected by a learnable linear layer:

$$\boldsymbol{\delta}_i^{\text{pos}} = \mathbf{W}_{\text{pos}}\,\gamma(\boldsymbol{\rho}_i) \ \text{ for background tokens}, \qquad \boldsymbol{\delta}_i^{\text{pos}} = \mathbf{0} \ \text{ for foreground tokens.} \tag{3}$$

This translates the image-plane axes to the person centroid rather than rotating into the anatomical frame, while laterality itself comes from the semantic channel. In total, $\mathbf{E}$ ($7\times D$) and $\mathbf{W}_{\text{pos}}$ ($16\times D$ plus a bias) have fewer than 0.1M parameters for D up to 4096.

4.3 Training Data and Strategy

Each training sample is an image or clip, its MPM, and an LR QA pair. Training and LR-Bench sources are disjoint at the *source* level (Appendix A.1). Because model-generated LR labels are themselves noisy, we build two complementary subsets. **Diverse natural QAs (∼50K):** we draw pose-grounded visual prompts (e.g., blue/red boxes over left/right limbs) as cues and prompt Gemini 3.1 Pro (Google, 2026) to write LR questions and answers, keeping only high-confidence samples. **Clean templated QAs (∼70K):** a single pose-grounded box on a body part with the template "Which body part is highlighted by the box?" and an answer such as "*left* knee", providing clean LR supervision at scale. We freeze the visual encoder, connector, and LLM and train only $\mathbf{E}$ and $\mathbf{W}_{\text{pos}}$ (<0.1M parameters, <0.001% of the backbone). One adapter is trained for each backbone. Since SLP only depends on the spatial layout of visual tokens, the same design applies to backbones with different token grids. As the trainable parameters are zero-initialized, a relatively large learning rate (10^{-2}) is stable and speeds up convergence.

5 Experiments

5.1 Setup

We apply SLP to Qwen3-VL-8B (Bai et al., 2025), InternVL3.5-8B (Wang et al., 2025), and GLM-4.1V-9B (Team et al., 2025b), training one adapter per backbone for 2 epochs with AdamW (learning rate 10^{-2}, batch size 192, warm-up ratio 0.05) on 24 A100 GPUs. All models take one image or 16 uniformly sampled frames and are asked to output the option directly (Appendix A). The other MLLMs are evaluated zero-shot.

5.2 Main Results on LR-Bench

Existing MLLMs fail at subject-centered LR discrimination. In Table 2, most open-source models, including the 235B-parameter Qwen3-VL, are within a few points of chance on both subsets, and scaling up helps little (Qwen3-VL-32B vs. 235B: 54.3 vs. 57.9 on videos). Gemini-3-Flash is the strongest zero-shot model, yet it reaches only 61.3% on videos.

Table 3: Mirror-paired analysis on LR-Bench$^{\text{Vid}}$ (%).

Model	Acc	Acc$^{\text{flip}}$	MCS↑	Same answer↓
Qwen3-VL-8B	50.16	29.29	9.18	61.09
+ **SLP**	**78.69**	**70.29**	**59.21**	**30.55**
InternVL3.5-8B	50.68	32.66	8.11	67.13
+ **SLP**	**78.54**	**72.71**	**60.39**	**30.47**
GLM-4.1V-9B	41.44	34.79	11.98	52.26
+ **SLP**	**77.24**	**70.84**	**58.52**	**31.04**

Table 4: Ablation of the semantic (SE) and positional (PE) channels on LR-Bench$^{\text{Vid}}$ with Qwen3-VL-8B. Accuracy (%).

SE	PE	Overall	Appear.	A & P	Spatial	Temporal
		50.16	54.96	60.05	25.00	52.72
✓		74.67	81.30	83.89	63.00	62.55
✓	✓	**78.69**	**84.09**	**86.74**	**71.20**	**65.69**

Table 5: Controlling for training data (Qwen3-VL-8B). SFT and LoRA are trained on the *same* 120K LR data without MPM. Overall accuracy (%).

Method	Trainable	LR-Bench$^{\text{Img}}$	LR-Bench$^{\text{Vid}}$
Frozen backbone	–	41.68	50.16
+ Full SFT	∼8B	59.02	60.25
+ LoRA	∼20M	62.26	61.48
+ **SLP (Ours)**	<**0.1M**	**79.48**	**78.69**

Table 6: Controlling for the interface (LR-Bench$^{\text{Vid}}$, Qwen3-VL-8B): the same L/R labels supplied as visual prompts vs. latent injection. Accuracy (%).

Injection	Overall	Appear.	A & P	Spatial	Temp.
Box visual prompt	53.11	53.00	60.22	38.30	55.13
Pose visual prompt	51.59	56.82	52.88	38.40	57.64
MPM visual prompt	57.32	60.74	64.33	41.20	57.32
Latent (SLP)	**78.69**	**84.09**	**86.74**	**71.20**	**65.69**

Spatial Relation is systematically *below* chance. For all open-source models and GPT-4.1, Spatial-Relation accuracy is between 18% and 33%, far below chance. The models answer in the *viewer's* frame, which is the mirror image of the subject's frame when the subject faces the camera, and are therefore consistently wrong in such cases. Simply flipping the answers would not fix this, because Appearance and Action & Pose are *above* chance for the same models.

SLP yields large and consistent gains. SLP adds 27.9–35.8 points on LR-Bench$^{\text{Vid}}$ and 36.9–48.7 on LR-Bench$^{\text{Img}}$, raising the 8–9B models from 32–51% to 77–80% overall, higher than all zero-shot models including Gemini-3-Flash and Qwen3-VL-235B. Appearance and Action & Pose exceed 80% on all three backbones, Spatial Relation improves from 18–29% to 62–73%, and Temporal Dynamics, which SLP does not directly target, still gains 11–24 points.

5.3 Diagnosis: Models Are Mirror-Invariant, Not Merely Inaccurate

Table 3 reports the mirror-paired analysis described in Sec. 3.4. Without SLP, MCS is only 8–12%, far *below* the 25% random reference, and the same-answer rate explains why. On 52–67% of the mirror pairs, the model gives the *same* answer to an image and its mirrored version, so the answer is decided by image-independent priors rather than by the subject's frame. Accordingly, accuracy on mirrored inputs drops to 29–35%, because mirroring reverses these priors (a watch on the left wrist now appears on the right), so the near-50% accuracy on the original inputs largely comes from priors. The remaining 25–36% of pairs are wrong on both versions, which is what viewer-centric answering produces. With SLP, MCS increases to 59–60% on all three backbones, the same-answer rate drops to about 30%, and accuracy on mirrored inputs (70–73%) becomes close to that on the original inputs, i.e., the model now changes its answer when the scene is mirrored.

5.4 Where Does the Gain Come From?

SLP is trained on 120K LR examples while the zero-shot baselines are not, so the overall gain mixes in-domain data, the structural prior, and the way the prior is injected. Two controlled experiments on Qwen3-VL-8B separate them.

Same data, no prior: SFT and LoRA plateau far below SLP (Table 5). Full-parameter SFT and LoRA (Appendix B.3) trained on the same data improve the frozen backbone by 17–21 points on images and 10–11 points on videos, reaching 59–62%, while SLP reaches 79% on both subsets with four orders of magnitude fewer trainable parameters. About *half* of the gain of SLP therefore cannot be obtained from the data alone, even when all backbone weights are updated.

Same prior, different interface: visual prompts barely help (Table 6). We render color-coded part boxes, skeletons, or the MPM itself as *visual prompts* and feed them to the frozen backbone

Table 7: Caption-based evaluation on LR-Bench^{Vid} (%; protocol 3 in Sec. 3.4). Image-subset results are in Appendix C.2.

Model	Overall		Appear.		A & P		Spatial		Temporal	
	Cov.	Prec.	Cov.	Prec.	Cov.	Prec.	Cov.	Prec.	Cov.	Prec.
Qwen3-VL-8B	67.42	70.05	30.89	65.89	**82.47**	71.76	**79.40**	68.89	63.18	69.37
+ **SLP**	**67.76**	**82.88**	**35.23**	**89.15**	80.44	**88.22**	77.70	**75.16**	**66.11**	**76.58**
InternVL3.5-8B	62.27	69.02	33.37	68.42	75.73	70.41	72.20	64.96	55.44	71.32
+ **SLP**	**68.27**	**84.86**	**41.01**	**88.92**	**81.21**	**88.66**	**73.40**	**78.47**	**65.79**	**80.76**
GLM-4.1V-9B	54.81	65.69	40.39	62.66	65.15	68.46	63.20	58.07	40.90	72.63
+ **SLP**	**63.95**	**81.00**	**44.32**	**87.41**	**75.07**	**84.82**	**64.00**	**70.94**	**62.55**	**78.43**

together with or instead of the original frames (reporting the better of the two). These baselines see the *same* L/R labels as SLP, only through pixels, yet gain at most about 7 points. Together with Table 5, this shows where the gain comes from: on LR-Bench^{Vid}, training without the prior reaches 61%, the prior without token-level injection reaches 57%, and only injecting the prior into visual tokens reaches 79%.

5.5 Ablation of the Two Channels

In Table 4, the semantic channel alone adds +24.5 points overall, so explicit part identity is the main source of the gain. It already raises Spatial Relation from 25.0 to 63.0, since the part map tells the model which way the subject faces (Sec. 4.1). The positional channel adds +8.2 on Spatial Relation and +4.0 overall by giving background tokens a person-anchored frame.

5.6 Robustness: Where Do the Gains Concentrate?

We further break down the caption-based results of Qwen3-VL-8B on LR-Bench^{Vid} by viewpoint, person count, and body visibility (Table 9, Appendix C.1). The gains are not limited to the easy case of a single person facing the camera: Precision improves in every group, most on frontal views (+15.6), where viewer-centric answering causes the most errors, and least on back, egocentric, and top-down views, where the baseline is already most accurate. It remains with two or more people (+9 to +10), even though the MPM does not single out the queried person, and on partially occluded subjects (+16.6). The remaining failures (Appendix C.4) mostly occur under heavy occlusion and unusual poses, where mesh recovery itself is unreliable.

5.7 Generalization to Open-Ended Captioning

Although SLP is trained only on forced-choice QA, Table 7 shows that the effect transfers to free-form captions: Precision increases by 13–16 points overall, 21–25 on Appearance, and 16–18 on Action & Pose, while Coverage stays the same or slightly increases. SLP thus does not make the model mention left and right more often, but makes the mentioned sides correct.

Efficiency. The extra inference cost, including the mesh model, is small (Appendix B.5). On one H800 with Qwen3-VL-8B, MPM extraction takes 0.30 s per image and 0.57 s per 16-frame clip, compared with 2.8 s for MLLM inference, and it is computed only once per input. SLP itself adds only 0.01–0.02 s. Qualitative examples and failure cases are in Appendix C.

6 Conclusion

We studied human subject-centered LR discrimination in MLLMs. LR-Bench, a human-labeled image-and-video benchmark with a mirror-paired protocol, shows that current models are not only inaccurate but largely mirror-invariant, answering from priors and the viewer's frame rather than the subject's. Spatial Latent Prompt injects a Mesh-Derived Part Map into frozen visual tokens with fewer than 0.1M parameters and raises three backbones from near chance to 77–80% while restoring mirror-equivariance, which neither full fine-tuning nor visual prompting with the same labels achieves. Remaining errors occur mostly where mesh recovery fails and should shrink with better mesh models.

AI USE STATEMENT

Generative AI tools were used as follows. *Data pipeline:* Qwen3-VL-235B-A22B pre-generated candidate questions (without answers) for LR-Bench, and Gemini 3.1 Pro generated LR training QA pairs. All benchmark labels were assigned and verified by human annotators (Sec. 3). *Evaluation:* Qwen3-235B-A22B served as the text-only caption judge (Sec. 3.4). *Writing:* LLM-based tools were used for language polishing. All technical content, claims, and numbers were produced and verified by the authors.

REPRODUCIBILITY STATEMENT

Evaluation prompts, the caption-judge prompt, and input preprocessing are given in Appendix A. The SLP architecture is specified in Sec. 4 and Appendices B.1–B.2, the training data and hyperparameters in Sec. 4.3 and Sec. 5.1, and the train/test source separation in Appendix A.1. We will release LR-Bench (annotations, source identifiers, and mirrored-pair evaluation scripts), the SLP code, and the trained adapter weights for all three backbones.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint*, arXiv:2303.08774, 2023.

Vaibhav Agrawal, Varghese Kuruvilla, Harsh Rangwani, and Ravi Kiran S. Partconcepts: A unified mechanism for fine-grained part localization and generation. In *NeurIPS*, 2026.

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. In *NeurIPS*, 2022.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. In *arxiv*, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint*, arXiv:2308.12966, 2023. URL `https://arxiv.org/abs/2308.12966`.

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

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

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint*, arXiv:2305.06500, 2023.

Michael Dorkenwald, Nimrod Barazani, Cees G. M. Snoek, and Yuki M. Asano. PIN: Positional insert unlocks object localisation abilities in VLMs. In *IEEE CVPR*, 2024.

Mingqi Gao, Yunqi Miao, and Jungong Han. SAM-Body4D: Training-free 4d human body mesh recovery from videos. *arXiv preprint arXiv:2512.08406*, 2025.

Google. Gemini 3 flash. https://google.com, 2026.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. MotionBench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *IEEE CVPR*, pp. 8450–8460, 2025.

Ji Hyeok Jung, Eun Tae Kim, Seoyeon Kim, Joo Ho Lee, Bumsoo Kim, and Buru Chang. Is 'right' right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning. In *IEEE CVPR*, pp. 14257–14267, 2025.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, 2023.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024.

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 *ICML*, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE CVPR*, pp. 22195–22206, 2024.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *IEEE CVPR*, 2022.

Yansheng Li, Linlin Wang, Tingzhu Wang, Xue Yang, Junwei Luo, Qi Wang, Youming Deng, Wenbin Wang, Xian Sun, Haifeng Li, Bo Dang, Yongjun Zhang, Yi Yu, and Junchi Yan. STAR: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. *IEEE TPAMI*, 47(3):1832–1849, 2025.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. In *ACL*, pp. 635–651, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023b.

Jingping Liu, Xingchen Peng, Yan Zhou, Ziyan Liu, Jie Zhai, Ronghao Chen, Huacan Wang, and Xiaofeng Jia. MirrorQA: Benchmarking multimodal LLMs on mirror-orientation reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *ACL*, pp. 40464–40476, July 2026a.

Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. Hoigen-1m: A large-scale dataset for human-object interaction video generation. *arXiv preprint arXiv:2503.23715*, 2025a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *ECCV*, 2024.

Yuansen Liu, Haiming Tang, Jinlong Peng, Jiangning Zhang, Xiaozhong Ji, Qingdong He, Donghao Luo, Zhenye Gan, Junwei Zhu, Yunhang Shen, et al. Human-MME: A holistic evaluation benchmark for human-centric multimodal large language models. In *ICLR*, 2026b.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution. In *ICLR*, 2025b.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing Du, Tianli Zhou, Wenjie Zhang, Huping Ding, Jiahe Li, Wen Li, Gui Hu, Yiliang Gu, Siran Yang, Jiamang Wang, Hailong Sun, Yibo Wang, Hui Sun, Jinlong Huang, Yuping He, Shengze Shi, Weihong Zhang, Guodong Zheng, Junpeng Jiang, Sensen Gao, Yi-Feng Wu, Sijia Chen, Yuhui Chen, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Ovis2.5 technical report. *arXiv:2508.11737*, 2025.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M. de Melo, Alan Yuille, and Jieneng Chen. 3DSR-Bench: A comprehensive 3D spatial reasoning benchmark. In *ICCV*, 2025.

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 *ICLR*, 2024.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Nazia Tasnim, Keanu Nichols, Yuting Yang, Nicholas Ikechukwu, Elva Zou, Deepti Ghadiyaram, and Bryan A Plummer. Seeing isn't orienting: A cognitively grounded benchmark reveals systematic orientation failures in mllms supplementary. *arXiv preprint arXiv:2603.11410*, 2026.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. MiMo-vl technical report, 2025a.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025b. URL https://arxiv.org/abs/2507.01006.

Maria Tsimpoukelli, Jacob L. Menick, Serkan Cabi, S. M. Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *NeurIPS*, 2021.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. CapRL: Stimulating dense image caption capabilities via reinforcement learning. *arXiv preprint arXiv:2509.22647*, 2025.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, Gege Qi, and Yunjian Zhang. Spatial-Bench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024a.

Lu Yang, Wenhe Jia, Shan Li, and Qing Song. Deep learning technique for human parsing: A survey and outlook. *International Journal of Computer Vision*, 132(8):3270–3301, 2024b.

Timing Yang, Sicheng He, Hongyi Jing, Jiawei Yang, Zhijian Liu, Chuhang Zou, and Yue Wang. Fast SAM 3D Body: Accelerating sam 3d body for real-time full-body human mesh recovery. In *ECCV*, 2026a.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, et al. SAM 3D Body: Robust full-body human mesh recovery. In *IEEE CVPR*, 2026b.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. MiniCPM-V 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2023.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *ICLR*, 2025.

Xiaoyun Zheng, Liwei Liao, Xufeng Li, Jianbo Jiao, Rongjie Wang, Feng Gao, Shiqi Wang, and Ronggang Wang. Pku-dymvhumans: A multi-view video benchmark for high-fidelity dynamic human modeling. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

A Benchmark and Evaluation Details

A.1 Training and Benchmark Data Separation

To rule out any data leakage between training and evaluation, LR-Bench and the training data of SLP are kept disjoint at the *source* level, rather than merely at the level of question text. The underlying images and videos are drawn from separate source pools, so that no source item, and therefore no mother video, temporally adjacent clip, or repeated subject, is shared between the two splits. Because the separation is enforced on the raw visual sources before any question generation, near-duplicate risks such as different clips of the same mother video or neighboring frames of the same shot cannot cross the training/evaluation boundary. Strong performance on LR-Bench therefore cannot be attributed to memorizing benchmark-adjacent visual content.

A.2 Input Preprocessing

For all evaluations, images and video frames are resized so that the longer side is 512 pixels while preserving the aspect ratio. For videos we uniformly sample 16 frames. The same preprocessing is used for binary QA, mirror-consistency, and caption-based evaluation. For the mirrored version of an item, every frame is flipped horizontally and the ground-truth answer is swapped; the question text is unchanged.

A.3 Evaluation Prompts

LR binary QA prompt. We append a concise output instruction after each question:

```
[Question] Output your option directly.
```

This discourages long chain-of-thought outputs and makes parsing reliable. The script then extracts the predicted option and compares it with the ground truth.

Caption generation prompt.

```
Give a detailed description of the people in this video, covering their
    appearance, actions & pose, and spatial relationships with egocentric
     left-right details.
For appearance, describe clothing, accessories, and distinctive features,
     with left-right details -- for example, which ear has an earring, or
     which arm has a tattoo/pattern.
For actions and poses, specify which side of body part is involved --
    such as which hand, arm, leg, or foot performs each action, and which
     direction each person turns, faces.
For spatial relationships, describe where objects or other people are
    positioned relative to each person -- specifically whether they are
    to the person's left or right.
Be explicit about left and right throughout.
```

Although the prompt uses the phrase "egocentric left-right", all judgments follow the subject-centered FoR defined in the main paper.

Caption judge prompt. The judge (Qwen3-235B-A22B, text-only) receives the caption, the question, and the ground-truth answer, and outputs one of *Correct*, *Incorrect*, or *Not Mentioned*. System prompt:

```
You are a strict and precise evaluator for video caption quality
    assessment.
Your task is to determine whether a generated caption contains the
    correct left/right spatial information to answer a given question
    about a video.
Always output valid JSON only, with no markdown, no extra text.
```

User prompt template:

```
You are given:
- [Generated Caption]: A caption describing a video, generated by a
    vision-language model.
```

```
- [Question]: A specific question about left/right spatial relationships
    in the video.
- [Ground Truth Answer]: The correct answer to the question.

Your task is to classify how well the [Generated Caption] covers the
    information needed to answer the [Question].

## Classification Categories
Choose exactly ONE of the following:
- "Correct": The caption contains information that correctly reflects or
    indirectly reasons the ground truth answer to the question, even if
    phrased differently.
- "Incorrect": The caption mentions the relevant subject but gives the
    wrong left/right direction, contradicting the ground truth.
- "Not Mentioned": The caption does not mention the subject of the
    question at all, making it impossible to answer from the caption.

## Input
[Generated Caption]: {caption}
[Question]: {question}
[Ground Truth Answer]: {answer}

## Output Format
Output a JSON object with exactly two keys:
- "reasoning": A brief step-by-step analysis (2-3 sentences) comparing
    the caption against the question and ground truth, focusing on left/
    right correctness.
- "category": One of exactly: "Correct", "Incorrect", "Not Mentioned"

Output the JSON object only. No markdown fences, no extra text.
```

Coverage and Precision are computed from the label counts as defined in Sec. 3.4.

B Method Details

B.1 Background Relative Position Encoding

Given the MPM, we merge all non-background categories into a foreground mask and compute its normalized centroid $(\bar{r}, \bar{c})$ on the visual-token grid. For a background token at normalized coordinate (r_i, c_i) the relative coordinate is $\boldsymbol{\rho}_i = (\Delta r_i, \Delta c_i) = (r_i - \bar{r},\ c_i - \bar{c})$. Each scalar $x \in \{\Delta r_i, \Delta c_i\}$ is encoded with L frequency bands,

$$\gamma(x) = [\sin(2^0 \pi x), \ldots, \sin(2^{L-1} \pi x),\ \cos(2^0 \pi x), \ldots, \cos(2^{L-1} \pi x)], \tag{4}$$

and the two axes are concatenated, $\gamma(\boldsymbol{\rho}_i) = [\gamma(\Delta r_i), \gamma(\Delta c_i)] \in \mathbb{R}^{4L}$, with $L = 4$. A learnable linear layer maps this to the token dimension, $\boldsymbol{\delta}_i^{\text{pos}} = \mathbf{W}_{\text{pos}} \gamma(\boldsymbol{\rho}_i) + \mathbf{b}_{\text{pos}}$, with $\mathbf{W}_{\text{pos}}$ and $\mathbf{b}_{\text{pos}}$ zero-initialized so that the visual representation is unchanged at the start of training.

Multi-person scenes. The centroid is computed over *all* foreground human regions rather than over a single target person. A target-specific centroid would require an explicit grounding step that identifies which person each question refers to, i.e., exactly the kind of privileged information we avoid. The global centroid provides a human-centric anchor for the scene, and the MLLM combines it with the per-token part embeddings, the visual content, and the question to resolve which person is meant. Empirically this suffices to improve the Spatial-Relation subset substantially (Sec. 5.5).

B.2 Video MPM Pipeline

Deriving the MPM for videos builds directly on off-the-shelf mesh recovery (Gao et al., 2025; Yang et al., 2026b;a). Because SLP only needs the anatomical left and right regions *within each frame*, it requires neither cross-frame identity tracking nor mask propagation.

Per-frame mesh recovery and part projection. SAM3D-Body (Yang et al., 2026b), accelerated with Fast SAM3D-Body (Yang et al., 2026a), is run independently on each frame. For every detected person the mesh is projected onto the image plane and converted into the six-category part map plus background, exactly as for images.

Frame alignment. The MLLM consumes 16 uniformly sampled frames per clip, and the MPM is built only for these 16 frames, so the injected prior is exactly aligned with the model input and no dense processing of the full video is needed.

Multi-person handling. When several people are present, their part maps are composited into one frame-level MPM in order of decreasing mask area, so larger (typically closer) subjects are drawn first and smaller ones overlaid on top. No "target" person is selected. The merged MPM describes the human-structural layout of the whole scene, and the MLLM resolves the queried subject from the question and visual context, exactly as the baseline must.

B.3 Fine-Tuning Baselines

The full-parameter SFT and LoRA baselines in Table 5 are trained on exactly the same 120K LR QA pairs as SLP, for the same number of optimization steps and with the same batch size of 192, but without the MPM. Both use AdamW with a cosine decay schedule and linear warm-up, and the LoRA adapters are attached to all attention and MLP projections of the LLM. Evaluation follows the protocol of Sec. 5.1.

B.4 Robustness of the Mesh-Derived Part Map

Figure 4 shows MPMs produced by SAM3D-Body across common and moderately challenging scenes, where the meshes provide reliable part-level left/right cues in the large majority of cases. Failures concentrate under extreme visibility degradation: heavy occlusion, severe truncation, or egocentric views where only a single hand is visible. SLP is tolerant to small local errors because patch-level pooling aggregates evidence over a token's receptive field and lateral sharpening (Eq. 2) emphasizes confident lateral regions, so a globally correct layout with pixel-level noise yields an essentially unchanged prompt.



Figure 4: Examples of Mesh-Derived Part Maps from SAM3D-Body. Reconstructions are reliable in common scenarios; failures arise mainly under extreme occlusion, truncation, or minimal visible body area.

B.5 Efficiency and Latency

Table 8 reports the end-to-end cost discussed in Sec. 5.7. MPM extraction is off the interactive path: it is computed once per image or clip and reused across all questions about that input, and because laterality is defined per frame the 16 frames are processed independently and can be batched. SLP adds a negligible overhead to MLLM inference since it adds a prompt to existing visual tokens without changing the token count. The MPM cost can be reduced further with faster mesh estimators or bypassed when body pose is available from depth sensors.

Table 8: End-to-end latency on one H800 GPU with Qwen3-VL-8B. MLLM inference uses a fixed budget of 64 generated tokens.

	Single image		16-frame video	
Stage	Baseline	+ SLP	Baseline	+ SLP
MPM extraction (det. + mesh + render; once per input)	–	0.30 s	–	0.57 s
MLLM inference (prefill + decode)	2.75 s	2.76 s	2.81 s	2.83 s

C Additional Experimental Results

C.1 Stratified Caption Results

Table 9 breaks down the caption-based results of Qwen3-VL-8B on LR-Bench$^{\text{Vid}}$, with and without SLP, by the source attributes annotated in LR-Bench (Sec. 3.3). Precision improves in every group. The gain is largest on frontal views (+15.6), where the viewer's frame is the mirror image of the subject's frame and viewer-centric answering causes the most errors. It is smallest on the remaining viewpoints (+2.0; back, egocentric, top-down, and changing viewpoints), where the baseline is already the most accurate, likely because for back views the two frames coincide. The gain remains with two people (+9.3) and with three or more people (+10.0), even though the MPM does not single out the queried person, and is also clear on partially occluded subjects (+16.6), where the 3D body prior helps fill in missing visual evidence. Coverage, which SLP does not target, stays almost the same.

Table 9: Caption-based evaluation of Qwen3-VL-8B on LR-Bench$^{\text{Vid}}$ (Coverage / Precision, %) stratified by camera viewpoint, person count, and body visibility.

	Camera viewpoint			Person count			Body visibility	
	Front	Side	Other	1	2	≥3	Full	Partial
Qwen3-VL-8B	67.21 / 68.63	67.92 / 70.81	65.94 / 79.34	66.79 / 70.38	69.75 / 70.68	61.93 / 68.89	67.75 / 70.50	40.35 / 52.17
+ **SLP**	66.74 / **84.24**	67.76 / **81.27**	67.30 / **81.38**	68.73 / **85.23**	67.44 / **79.96**	57.57 / **78.88**	67.33 / **82.98**	56.14 / **68.75**

C.2 Caption-Based Evaluation on LR-Bench$^{\text{IMG}}$

Table 10 complements the video caption results in Sec. 5.7. SLP improves caption Precision by 14.4, 19.2, and 24.3 points overall on Qwen3-VL-8B, InternVL3.5-8B, and GLM-4.1V-9B respectively, while Coverage fluctuates within a few points: open-ended captioning depends on whether the model chooses to mention an LR detail at all, which SLP does not target, whereas the correctness of the mentioned details, which SLP does target, improves consistently. GLM-4.1V-9B's low absolute Coverage reflects a weaker base captioning tendency toward LR detail.

C.3 Qualitative Results

Figure 5 shows typical cases where the baseline answers in screen coordinates while SLP answers in the subject's frame. Figure 6 shows further QA cases across the four categories; the baseline typically follows viewer-centric cues, whereas SLP grounds its answer in the subject's frame. Figure 7 shows caption examples, where SLP corrects the side of mentioned body parts, actions, and surrounding objects while leaving the rest of the caption intact.

Table 10: Caption-based evaluation on LR-Bench^{Img} (%). Coverage and Precision as defined in Sec. 3.4.

Model	Overall		Appear.		A & P		Spatial	
	Cov.	Prec.	Cov.	Prec.	Cov.	Prec.	Cov.	Prec.
Qwen3-VL-8B	68.20	64.81	60.00	61.94	66.34	65.20	82.00	67.74
+ **SLP**	67.14	**79.19**	56.40	**82.51**	68.21	**73.76**	82.40	**79.37**
InternVL3.5-8B	66.38	63.01	58.87	65.35	62.45	64.93	80.80	59.28
+ **SLP**	68.81	**82.16**	63.87	**84.86**	71.46	**82.49**	74.10	**78.41**
GLM-4.1V-9B	47.69	56.00	45.31	60.73	41.81	60.45	55.80	48.51
+ **SLP**	47.35	**80.28**	47.13	**84.58**	50.94	**80.84**	44.80	**72.99**

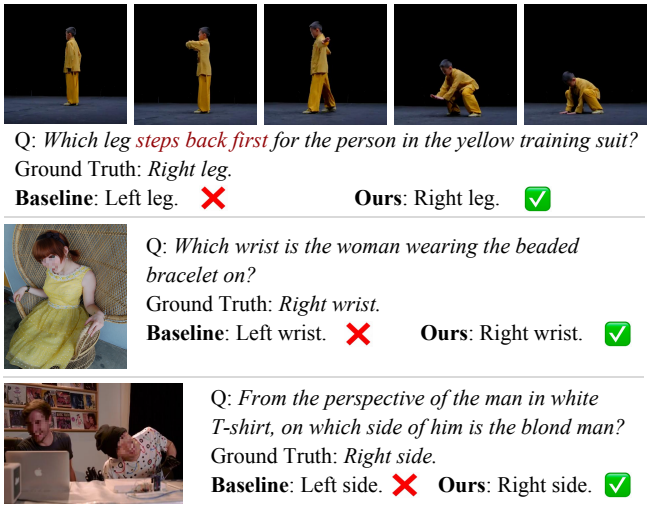


Figure 5: Qualitative results on LR-Bench: the baseline follows the viewer's frame, whereas SLP grounds its answer in the subject's frame across appearance, action tracking, and person-relative spatial relations.

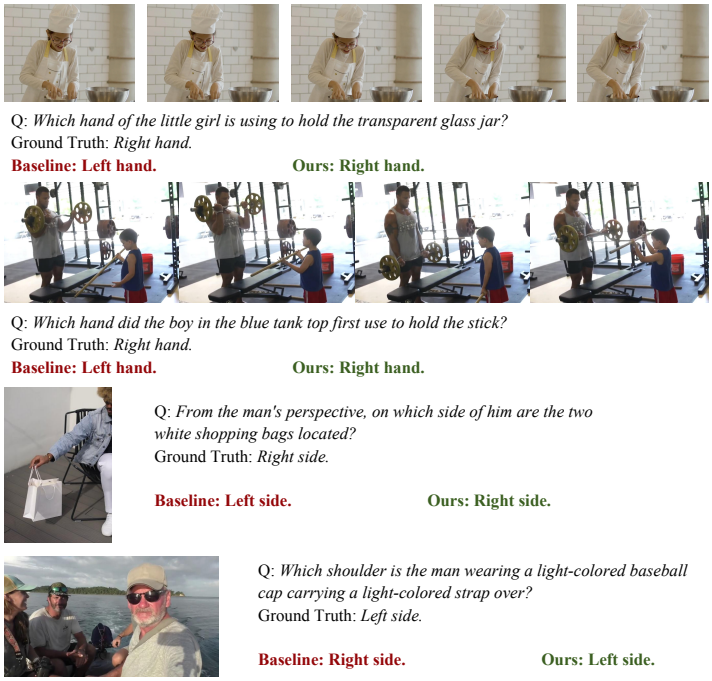


Figure 6: Additional qualitative results for LR binary QA on LR-Bench.

C.4 Failure Cases

Figure 8 shows representative remaining failures. In the first case both hands are close to the manipulated object, so a small MPM error around the hands flips the judgment. In the second (spatial-relation) case the pose is highly non-canonical, requiring additional reasoning to map the object into the subject's frame. In the third case severe occlusion of both hands produces MPM errors that mislead the model. These cases point to more accurate mesh recovery under occlusion and explicit uncertainty handling as the natural next steps.



***Baseline Caption*:**
In the video, there are two individuals sitting closely together on a gray couch. The person on the left is wearing a light pink shirt and has curly hair. They are holding a laptop on their lap with their **left** hand, and their **right** arm is wrapped around the person next to them. The person on the right is a young child with blonde hair, who is shirtless and resting their head on the shoulder of the person on the left. The child's **right** hand is near their mouth, and they are looking at the laptop screen. Both individuals are facing slightly to the left, focusing on the laptop screen. The background is a plain white wall, and there is a white pillow behind them on the couch.

***Ours Caption*:**
In the video, there are two people sitting closely together on a gray couch. The person on the left is a woman with curly blonde hair, wearing a light pink shirt. She is holding a laptop on her lap with her **right** hand, and her **left** arm is around the child next to her. The child, who is shirtless, is sitting in front of her, leaning against her **left** side. The child has light blonde hair and is resting their **right** hand on their chest, looking at the laptop screen with a thoughtful expression. The woman's **right** hand is on the laptop, and her **left** arm is wrapped around the child, providing comfort and support. The couch has a white pillow behind them, and the background is a plain white wall.



***Baseline Caption*:**
In the image, there is a man standing on a wooden pier by the water. He is facing the camera directly, with his body slightly angled to his left. Here is a detailed description:
Appearance:
He is wearing a light-colored polo shirt and light-colored shorts. He has a cap on his head, which appears to have a logo on the front. He is wearing sunglasses, which cover both eyes. He has a short beard and mustache. His facial features are not detailed enough to specify any other distinctive marks.
Actions and Pose:
His **right** hand is resting on the wooden railing of the pier, while his **left** hand is in his pocket. He is standing upright, facing forward, with a relaxed posture.
Spatial Relationships:
Behind him, there is a body of water extending to the horizon. A bridge is visible in the distance, along with some structures on the far side of the water. To his **left** (viewer's right), there is another person sitting on a boat. This person is wearing a green shirt and is facing away from the camera, looking out towards the water. The wooden railing runs horizontally across the image, with the man's **right** hand resting on it. The pier extends into the water on both sides of the man.

***Ours Caption*:**
In the image, a man is standing by a wooden railing overlooking a body of water with a bridge in the distance. Here's a detailed description:
Appearance:
He is wearing a light gray polo shirt with short sleeves. The shirt is plain without any visible logos or patterns. He is wearing a beige baseball cap with a logo on the front, and dark sunglasses. The cap is worn slightly tilted to his **right** side. He has a short beard and mustache. His **left** hand is resting on the railing, and his **right** hand is in his pocket.
Actions and Pose:
His **left** hand is placed on the wooden railing, suggesting he is leaning slightly on it for support. His **right** hand is tucked into his pocket, indicating a relaxed posture. He is facing slightly to his **right**, with his body turned in that direction. His gaze is directed towards the camera, giving a neutral expression.
Spatial Relationships:
Behind him, to his **right**, there is another person wearing a green shirt, standing near a small vehicle (possibly a golf cart or utility vehicle). This person is facing away from the camera, looking towards the water or the bridge. The railing he is leaning on runs parallel to the water's edge. The bridge is visible in the distance, spanning across the water.

Figure 7: Qualitative examples for caption-based evaluation. Wrong LR descriptions are shown in red and correct ones in green.



Q: *After removing the green component from the roof, which hand does the lady use to place it on the table?*
Ground Truth: ***Right hand.***
Prediction: Left hand.



Q: *From the painter's perspective, on which side of their body is the glass containing paint and brushes located while painting?*

Ground Truth: ***Right side.***
Prediction: Left side.



Q: *Which hand is the elderly man in the red shirt using to press the violin strings?*

Ground Truth: ***Right hand.***
Prediction: Left hand.

Figure 8: Failure cases for subject-centered LR understanding.