OR-SpaPose: A Multi-view MLLM for Spatial Scene Understanding in Operating Rooms

Anonymous ACL submission

Abstract

Reliable spatial understanding in Operating Rooms (ORs) is essential for safety monitoring and robotic assistance, yet OR scenes are crowded, dynamic, and frequently occluded. Existing MLLMs often rely on single-view RGB observations or external structural inputs, making fine-grained spatial reasoning under viewpoint ambiguity difficult. We propose OR-SpaPose, a multi-view spatial-pose MLLM that aggregates synchronized RGB views, injects pose and estimated geometry features into the visual token space for language-based reasoning over personnel, instruments, locations, and interactions. To support training and evaluation, we build an automated pipeline that converts expert 2D instance masks and synchronized RGB views into spatially grounded QA pairs using multi-view geometric constraints, without manually annotating 3D object geometry. Experiments on knee replacement surgeries show that OR-SpaPose outperforms single-view supervised baselines and general-purpose MLLMs, with clear gains under reduced reference-view visibility.

1 Introduction

The operating room (OR) is a high-intensity environment defined by the coordination of medical staff, patients, and numerous devices. In this confined space, multiple personnel move around the patient while manipulating surgical instruments and equipment (Özsoy et al., 2025, 2022). Reliable OR understanding therefore requires models to reason about the physical layout of the scene, including which entities are present, where personnel and instruments are located, how they are positioned relative to one another, and which human-object interactions are taking place. Such spatial and interaction-centric understanding is important for safety monitoring, workflow awareness, and downstream surgical assistance (Maier-Hein et al., 2017). Figure 1 contrasts our RGB-only multi-view setting with single-view and modality-rich OR understanding paradigms.

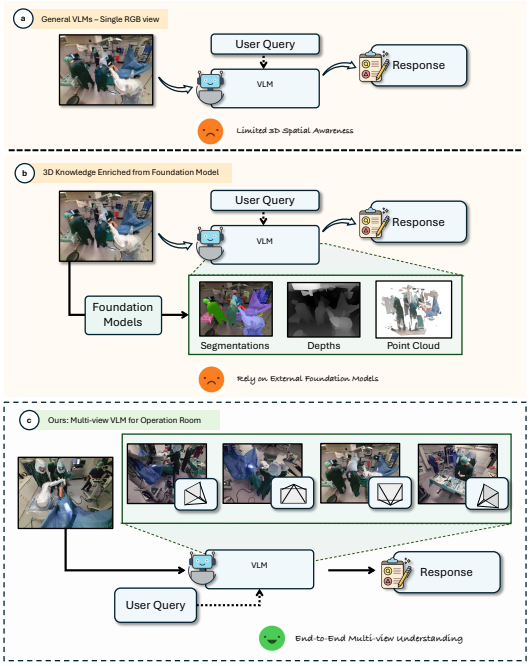

Figure 1: **Overview of different paradigms for OR scene understanding.** (a) General VLMs usually process a single RGB view and lack explicit spatial evidence, making them sensitive to occlusion and viewpoint ambiguity. (b) OR foundation models build domain-specific models for surgical scenes, often using additional modalities or structured annotations such as depth, point clouds, masks, or scene graphs. (c) Our method provides an end-to-end multi-view MLLM that integrates synchronized RGB views with token-aligned pose and geometry cues for spatial QA.

Visual spatial reasoning in this setting is difficult because the relevant evidence is often incomplete from any single view. Prior surgical vision-language systems and general-purpose MLLMs are commonly evaluated with single-view RGB observations and do not explicitly model multi-view spatial constraints (Liu et al., 2025; Wenyu et al., 2024), which limits geometric reasoning when instruments, hands, or staff are occluded. Other approaches improve 3D understanding by relying on

richer structural inputs such as depth, point clouds, 3D pose, scene graphs, or explicit spatial coordinates (He et al., 2025; Özsoy et al., 2026, 2022). These signals are valuable, but they also make comparisons sensitive to the available sensors and annotation formats. Our goal is to study language-based spatial QA over synchronized multi-view RGB streams, using the complementary views to reduce occlusion and viewpoint ambiguity while avoiding manual 3D object geometry labels.

Multimodal Large Language Models (MLLMs) provide a flexible interface for language-based scene understanding because they can answer natural-language questions about visual content. Despite this potential, recent OR studies show that general-purpose MLLMs struggle to extract clinically meaningful signals from surgical scenes, particularly for spatial reasoning, localization, and interaction-centric queries (Özsoy et al., 2026; He et al., 2025). This limitation restricts their use in safety-critical monitoring and robotic assistance within surgical environments.

We argue that robust spatial understanding in the OR requires two critical signals often missing in general MLLMs. First, ORs typically have multiple fixed cameras. Using multi-view RGB data can mitigate occlusion and provide stronger geometric cues than a single image, effectively resolving perspective ambiguities. Second, OR activities are inherently human-centric. Human pose provides a useful structural prior for interpreting where staff are located and how they interact with surrounding instruments and equipment (Li et al., 2025; Hamoud et al., 2025). However, this cue is often not explicitly integrated into general-purpose MLLMs for OR spatial reasoning. Consequently, reasoning about "who is where, and what they are interacting with" remains challenging under occlusion.

In this work, we present a Multi-view Spatial-Pose MLLM designed for spatial question answering in the OR. Our model keeps a pretrained Qwen3-VL backbone as the semantic reasoning core and augments its visual representation with two residual branches: a VGGT-based spatial encoder that provides geometry-aware features, and a pose encoder that supplies structured human cues. Both auxiliary signals are aligned to the base visual token grid and integrated via gated token-level residuals. This design avoids additional sensing hardware and preserves the language model interface, while making the added computational cost explicit in our analysis.

To support training and evaluation under OR constraints, we develop an automated pipeline that converts synchronized RGB views and expert 2D instance masks from MM-OR into spatially grounded question–answer pairs using multi-view geometric constraints. The 2D masks provide auditable object support during dataset construction, but the pipeline does not require manually written QA pairs or manually annotated 3D object geometry such as centroids, boxes, or spatial relation labels. At inference time, OR-SpaPose takes synchronized RGB views as input and does not use depth maps or ground-truth masks.

Our contributions are summarized as follows:

- We propose an RGB-only multi-view spatial-pose MLLM for OR scene understanding, designed to handle occlusion and dense interactions in cluttered medical environments.
- We enhance a Qwen3-VL MLLM by injecting geometry cues from a VGGT spatial encoder and human cues from a pose encoder via token-level residual adapters to preserve pre-trained semantic knowledge.
- We introduce an automated pipeline that generates multi-view spatial QA pairs from expert 2D masks and RGB views without requiring manually written QA pairs or 3D geometry labels.

Experiments on held-out surgeries demonstrate gains over single-view supervised baselines and general-purpose MLLMs, with diagnostic analyses for occlusion, component contributions, and inference cost.

2 Related Work

2.1 Spatial Understanding in MLLMs

Spatial reasoning remains difficult for MLLMs when questions require object-relative positions, metric cues, or viewpoint-dependent interpretation rather than semantic recognition alone (Chen et al., 2024; Cheng et al., 2024). Recent work improves this ability through spatial supervision, as in SpatialVLM (Chen et al., 2024) and SpatialRGPT (Cheng et al., 2024), or through auxiliary geometric inputs such as depth, point clouds, video-derived 3D representations, and geometry encoders (Wu et al.; Zhu et al., 2024; Zheng et al., 2025). These directions are complementary to ours:

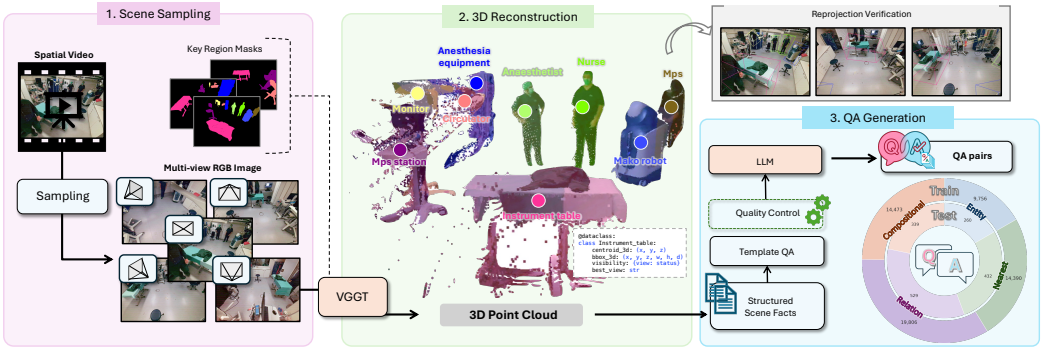

Figure 2: **Dataset construction pipeline.** Left: synchronized OR frames are sampled from MM-OR together with expert masks. Middle: multi-view geometry reconstruction produces camera parameters, point maps, object-level 3D boxes, centroids, visibility metadata, and relation facts; the `instrument_table` example illustrates the key stored fields. Top right: reprojection verification checks the reconstructed 3D geometry by projecting it back to image views and comparing the projected boxes with 2D mask boxes. Bottom: template QA pairs are generated from structured facts, filtered by automatic quality checks, and then paraphrased by an LLM. Bottom right: train/test QA distributions are reported by question type.

they often target general 3D grounding or video representations with explicit spatial coordinates, while we study language-based spatial QA over synchronized OR RGB views under occlusion.

2.2 Multi-view Perception and Pose Priors

Multi-view perception reduces single-view ambiguity by aggregating complementary evidence across viewpoints (Srivastav et al., 2024; Bragagnolo et al., 2024), and has been used for reconstruction, tracking, detection, and increasingly multimodal reasoning (Li et al., 2024b; Zheng et al., 2025). For open-ended QA, however, naive view concatenation can obscure view-specific details, while heavy cross-view attention is costly and brittle in cluttered scenes. Human pose is also useful for human-object interaction understanding, but it is rarely injected as a structured signal that directly participates in MLLM reasoning under occlusion. We therefore align multi-view pose and dense geometry as lightweight auxiliary token streams.

2.3 Visual Reasoning in Surgical Environments

Operating rooms are difficult for spatial reasoning because of clutter, reflective instruments, frequent occlusion, and dynamic workflows (Özsoy et al., 2022, 2025). Datasets such as 4D-OR (Özsoy et al., 2022) and MM-OR (Özsoy et al., 2025), together with OR-specific MLLMs such as Spatial-ORMLLM (He et al., 2025), have advanced structured OR understanding. Yet surgical VQA remains hard to compare across systems because available sensors and annotations differ: recent methods may rely on RGB-D streams, depth estimates, point clouds, masks, or scene graphs (Özsoy et al., 2022, 2025; He et al., 2025; Özsoy et al., 2026). We focus on the common RGB-camera setting and evaluate whether estimated geometry and pose can improve language-based OR spatial QA under occlusion.

3 Spatial QA Construction

We construct a multi-view operating-room spatial QA dataset from MM-OR (Özsoy et al., 2025), which records knee-replacement surgeries with synchronized RGB cameras, expert 2D instance masks, and semantic annotations. We sample one candidate every 10 annotated frames, retain scenes with the required views and masks, and pair each QA item with five RGB views. Directional questions use a *reference view* to define the coordinate frame, while the remaining *supporting views* provide evidence for resolving occlusion and viewpoint ambiguity. The pipeline uses RGB streams and masks during construction, but requires no depth input, manual 3D labels, or manually written QA pairs.

Figure 2 summarizes the pipeline: sample synchronized frames, estimate multi-view geometry with VGGT (Wang et al., 2025), lift masked pixels into 3D, derive scene facts, and generate QA pairs with deterministic templates plus constrained paraphrasing. Answers are computed from structured scene facts rather than from a language model.

Scene facts. For each scene, VGGT predicts camera parameters, dense depth maps, and point maps from synchronized RGB views. Expert masks aggregate object-level 3D points across views; after outlier removal, we fit object boxes and re-

Table 1: **QA types.** Counts are computed over the generated QA dataset.

Type	Subtype	Count	Example question
Entity	role	5,005	What roles can be recognized in the scene?
Entity	equipment	5,011	What equipment or objects can be recognized in the scene?
Nearest	role–role	6,632	Which two people are closest to `mps`?
Nearest	role–equipment	8,190	Which two equipment items are closest to `nurse`?
Relation	spatial	18,584	Where is `nurse` located with respect to `mps_station`?
Relation	semantic	1,751	What is the relationship between `nurse` and `instrument_table`?
Compositional	nearest+spatial	5,505	Where is the equipment closest to `anesthesia_equipment` relative to `mako_robot`?
Compositional	semantic+spatial	1,538	Where is the person close to `patient` located relative to `saw`?
Compositional	spatial chain	7,769	If `instrument_table` is behind `mps_station`, and `mps_station` is also behind `anesthesia_equipment`, what direction can be inferred?

project them to the camera views to compare projected boxes with the original 2D mask boxes (Figure 2). This check helps identify gross geometry failures before centroids are used for proximity and directional-relation facts. The resulting fact graph stores entity classes, object IDs, 3D centroids and boxes, nearest-neighbor rankings, reference-view spatial relations, MM-OR semantic interaction triples, and per-view visibility metadata.

Spatial labels. For directional questions, object centroids are transformed into the reference camera frame. We normalize the relative offset between subject and object by the average 3D object extent. A horizontal relation is emitted only when the normalized left–right or depth offset exceeds 0.25. If both axes exceed 0.25, we assign a diagonal label such as `leftFront` or `rightBehind`; otherwise, the dominant valid axis gives `leftOf`/`rightOf` or `inFrontOf`/`behind`.

Occlusion calculation. We use a visibility score to describe how clearly each object appears in each view, based on mask area, inter-object overlap, and image-boundary truncation:

$$s_{\text{vis}} = 0.50 s_{\text{area}} + 0.35(1 - r_{\text{overlap}}) + 0.15(1 - r_{\text{edge}}).$$

where $s_{\text{area}} = \min(1, r_{\text{area}}/0.08)$ and larger scores indicate clearer visibility. Mask area receives the largest weight because very small objects are hard to inspect; overlap directly captures occlusion by nearby entities; boundary truncation captures partial cropping at the image edge. Scores are mapped to *clear* (≥ 0.65), *partial* ($0.35 \leq s_{\text{vis}} < 0.65$), and *heavy* ($< 0.35$); non-visible objects in annotated views are marked *absent*, and non-annotated views are marked *unknown*. These metadata support filtering and the occlusion analysis in Section 5.

QA types and balancing. The dataset contains four question groups: *Entity* questions ask which roles or equipment are present; *Nearest* questions ask for the top-2 closest roles or equipment items; *Relation* questions cover reference-view spatial directions and annotated semantic interactions; and *Compositional* questions combine facts through nearest+spatial, semantic+spatial, or spatial-chain reasoning. We generate candidates with deterministic templates and constrained paraphrasing that preserves entities and answers. To limit scene and entity-pair imbalance, each scene contributes at most 12 QA pairs, with caps of 2 entity, 3 nearest, 4 relation, and 3 compositional questions; spatial pairs keep only the strongest evidence instance. The final set contains 10,016 entity (16.7%), 14,822 nearest (24.7%), 20,335 relation (33.9%), and 14,812 compositional (24.7%) questions. Table 1 reports the subtype distribution.

QA Filtering and Validation. After candidate generation, we remove invalid types, malformed answers, invalid visibility records, duplicate entity pairs, missing source facts, and other noisy candidates. For spatial questions, a reference view is required; for compositional questions, the supporting facts must be present and traceable. This keeps each QA item grounded in structured evidence before balancing. The dataset contains 5,011 scenes and 59,985 QA pairs from 20 surgeries. We split by surgery to reduce leakage: 18 surgeries for training (4,881 scenes, 58,425 QA pairs) and 2 held-out surgeries for testing (130 scenes, 1,560 QA pairs).

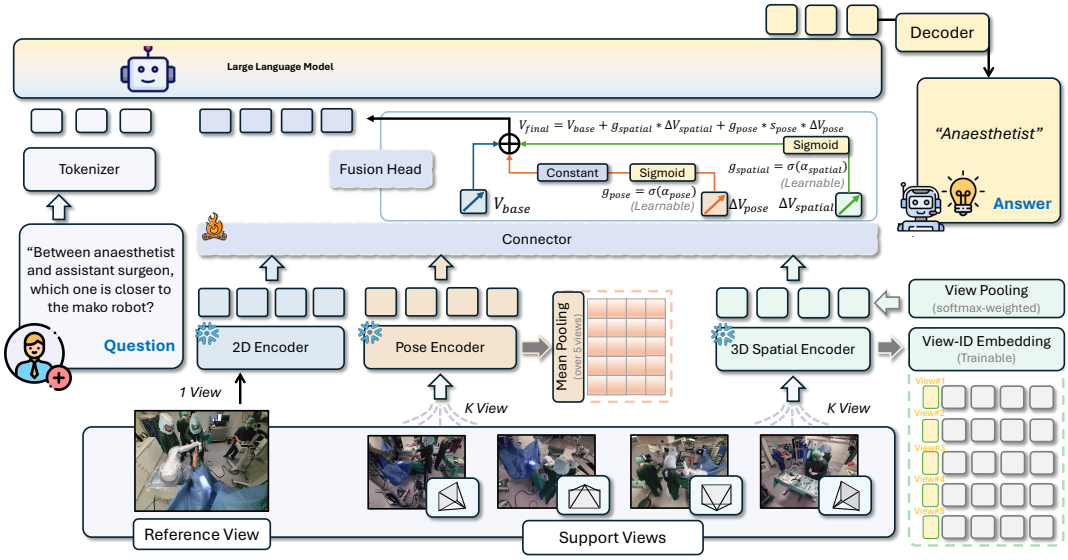

Figure 3: **Architecture of our Multi-view Spatial-Pose MLLM framework.** The reference view is encoded by Qwen3-VL to produce base visual tokens, while K synchronized views ($K = 5$ in our experiments) are processed by two auxiliary streams: a pose encoder for human-structure cues and VGGT for multi-view geometry. Trainable adapters align both streams to the Qwen3-VL token grid, and gated residual fusion injects the resulting pose and spatial tokens before the language model answers the query.

Table 2: **QA quality validation by question group.** We audit 250 stratified test QA pairs. Gold, Ans., and Valid denote gold-answer correctness, image answerability, and overall validity rates (%).

Group	#	Gold	Ans.	Valid
Entity	51	100.0	100.0	100.0
Nearest	51	96.1	100.0	96.1
Rel-spatial	53	100.0	100.0	100.0
Rel-semantic	45	97.8	97.8	97.8
Compositional	50	100.0	100.0	100.0
Overall	250	98.8	99.6	98.8

To estimate label quality after paraphrasing, two annotators independently audit 250 balanced test QA pairs by checking the synchronized views, question, and answer; a third reviewer verifies flagged cases and resolves disagreements. As shown in Table 2, validity is high after filtering because answers are generated from structured scene facts: entity questions use the visible object inventory, spatial and compositional questions use deterministic geometric facts, and semantic-relation questions use MM-OR annotations. The few errors mainly appear in nearest-neighbor and semantic-relation questions, where close distances, visibility, or interaction ambiguity can affect the label.

4 Method

Given a multimodal query q, a reference-view image $I^{(r)}$, and a synchronized set of K RGB views $\mathcal{I} = \{I^{(k)}\}_{k=1}^{K}$, our goal is to generate the answer y for an operating-room spatial reasoning question. The reference view defines the coordinate frame for view-dependent relations, while all synchronized views provide auxiliary evidence for resolving occlusion and ambiguity. We build on Qwen3-VL and keep the original language model and visual encoder as the semantic backbone. OR-SpaPose augments the backbone visual tokens with two structured residual sources: multi-view geometry from VGGT and articulated human pose from a frozen pose ViT. Both branches are projected to the Qwen3-VL visual-token space and injected additively, so the language model receives a sequence with the same length as the original backbone.

4.1 Architecture Overview

Figure 3 illustrates the overall architecture. The backbone first encodes the reference-view image with the Qwen3-VL visual encoder, producing visual tokens $\mathbf{V}_{\text{base}} \in \mathbb{R}^{N \times d}$. In parallel, the auxiliary inputs keep the synchronized views in a stacked tensor format ($K = 5$ in our experiments). VGGT processes this multi-view stack to produce dense geometry-aware patch embeddings, while the pose branch estimates human-structure features from all available views. The final visual tokens are

$$\mathbf{V}_{\text{final}} = \mathbf{V}_{\text{base}} + \sigma(\alpha_{\text{spatial}})\Delta\mathbf{V}_{\text{spatial}} + \sigma(\alpha_{\text{pose}})s_{\text{pose}}\Delta\mathbf{V}_{\text{pose}}. \quad (1)$$

where α_{spatial} and α_{pose} are learnable scalar gates initialized to negative values, s_{pose} is a fixed pose

scale, and $\sigma(\cdot)$ denotes the sigmoid function. This near-zero initialization makes the model start close to the original Qwen3-VL distribution and learn how much structural evidence to use during fine-tuning.

4.2 Pose Encoder: Multi-view Human Structure

The pose branch uses a frozen pretrained pose ViT backbone. For each image item, we collect the synchronized RGB views, resize them to the pose encoder input resolution, and apply ImageNet normalization. The frozen pose encoder outputs feature maps $\mathbf{P}^{(k)} \in \mathbb{R}^{C_p \times H_p \times W_p}$ for each view. We aggregate the multi-view evidence by averaging feature maps across views:

$$\bar{\mathbf{P}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{P}^{(k)}. \tag{2}$$

This design lets the pose branch recover human configuration cues even when the reference view is partially occluded.

The aggregated pose map is converted to Qwen3-VL visual tokens using a lightweight pose adapter. The adapter resizes the pose feature grid to the Qwen3-VL visual grid, applies the same spatial merge factor as the backbone, and projects the merged pose features to the language hidden dimension:

$$\Delta \mathbf{V}_{\text{pose}} = f_{\text{pose}}\Big(\text{Merge}\Big(\text{Resize}(\bar{\mathbf{P}})\Big)\Big). \tag{3}$$

The projected pose tokens are added to the backbone visual tokens through the learnable pose gate.

4.3 Spatial Encoder: VGGT Geometry Injection

The spatial branch uses a frozen VGGT-1B encoder to extract dense geometry-aware features from the synchronized views. In the image setting, the views are passed as a stacked temporal dimension, allowing VGGT to jointly model cross-view geometry without requiring depth sensors or manual camera calibration at inference time.

To preserve view identity inside the multi-view stack, the VGGT spatial encoder adds a trainable view-index embedding to the VGGT aggregator outputs. Let $\mathbf{S}^{(k)}$ be the VGGT patch sequence for view k and $\mathbf{e}^{(k)}$ the corresponding view embedding. The spatial tokens are augmented as

$$\tilde{\mathbf{S}}^{(k)} = \mathbf{S}^{(k)} + \mathbf{e}^{(k)}. \tag{4}$$

The resulting VGGT embeddings are aligned to the Qwen3-VL token grid by a lightweight MLP connector. The connector resizes VGGT patch maps to the target grid specified by the Qwen3-VL visual metadata, applies the backbone's temporal and spatial merge factors, temporally pools stacked views when needed, normalizes the merged representation, and projects it to the hidden dimension:

$$\Delta \mathbf{V}_{\text{spatial}} = f_{\text{spatial}}\Big(\text{Merge}\Big(\text{Resize}(\tilde{\mathbf{S}})\Big)\Big). \tag{5}$$

The connector's final projection is zero-initialized and the spatial residual is additionally controlled by the scalar gate α_{spatial}, which stabilizes training while allowing the model to learn dense spatial cues.

4.4 Training Objective

We use supervised fine-tuning with the standard autoregressive cross-entropy objective:

$$\mathcal{L} = -\sum_{t} \log P_\theta\left(y_t \mid y_{<t}, \mathbf{V}_{\text{final}}, q\right). \tag{6}$$

This objective trains the answer generator to condition on the fused reference-view, pose, and geometry tokens. The freezing strategy and optimization details are described in Section 5.

5 Experiments

5.1 Experimental Setup

Data and tasks. We evaluate on the surgery-level split described in Section 3: 18 surgeries with 58,425 QA pairs for training and 2 held-out surgeries with 1,560 QA pairs for testing. Each QA item is paired with one reference view and five synchronized RGB views. The benchmark covers entity recognition, 3D nearest-neighbor reasoning, spatial and semantic relations, and compositional questions that combine multiple facts. All supervised variants are trained and evaluated on the same split and answer inventory.

Baselines. We compare OR-SpaPose with representative RGB/image-input MLLMs, including open-source MLLMs, closed-source MLLMs, Spatial-MLLM, and single-view supervised Qwen3-VL baselines. We do not directly compare with 3D/video models that require depth, point clouds, ground-truth masks, or scene graphs at inference time, since these inputs are outside our RGB-only setting. The supervised Qwen3-VL

Table 3: **Main results.** Scores are reported in percentage (%). Closed-source models are evaluated on a stratified 245-sample subset; other models use the full test set.

Model	Nearest	Rel-spatial	Rel-semantic	Entity	Compositional	Overall
Closed-source comparison (245-sample subset)						
Gemini-2.5-Flash	18.00	32.00	40.00	12.13	35.00	27.17
GPT-4o	26.00	20.00	37.78	65.51	28.00	35.41
Claude-4.5-Haiku	25.00	30.00	46.67	70.47	34.69	41.28
GPT-5.1	30.00	7.00	44.44	74.32	24.49	35.93
Claude-Sonnet-4.6	32.00	17.00	60.00	80.03	37.00	44.90
OR-SpaPose-4B	**69.00**	**82.00**	57.78	87.40	68.00	73.14
OR-SpaPose-8B	67.00	81.00	**66.67**	**91.42**	**78.00**	**77.02**
Open-source and supervised models (full test set)						
LLaVA-NeXT-Interleave-7B (Li et al., 2024b)	24.54	20.45	4.44	4.09	59.88	26.96
LLaVA-OneVision-7B (Li et al., 2024a)	5.44	12.19	17.78	5.63	57.08	19.14
Spatial-MLLM (Wu et al.)	38.89	70.16	37.50	54.78	66.67	56.50
Qwen3-VL-4B (Bai et al., 2025)	61.34	80.37	51.11	80.04	58.41	69.43
Qwen3-VL-8B (Bai et al., 2025)	63.43	78.10	55.56	86.25	61.36	71.11
OR-SpaPose-4B	62.15	**83.47**	57.78	85.20	62.39	72.53
OR-SpaPose-8B	**65.74**	82.44	**66.67**	**90.37**	**71.09**	**76.22**

baselines receive only the reference-view image, whereas OR-SpaPose additionally injects pose and geometry cues extracted from the synchronized RGB views. All Qwen3-VL variants use the same image resolution constraints, decoding protocol, canonical answer space, and scoring script, so the comparison isolates the effect of multi-view structural evidence.

Implementation. All supervised models are fine-tuned for one epoch. OR-SpaPose freezes the Qwen3-VL backbone and auxiliary encoders, training only lightweight projectors, adapters, and gates. The Qwen3-VL stream uses the reference view, while pose/VGGT branches use all five synchronized RGB views. Hyperparameters, decoding settings, model size, and runtime estimates are in Appendix B.

Evaluation. All models are scored in the same canonical answer space. Entity-list and nearest-neighbor questions use recall over canonical class names; semantic relations, yes/no answers, and other closed-set answers use normalized exact match. Directional answers in both relation and compositional questions use a spatial partial-credit scorer: single-axis labels require an exact match, while diagonal labels receive component-level credit. Closed-source MLLMs are evaluated on a stratified 245-sample subset with the same output normalizer; details are in Appendix C.

5.2 Main Results

Table 3 summarizes the main results. On the full test set, OR-SpaPose-8B improves Qwen3-VL-8B from 71.11 to 76.22 overall, with the clearest gains on compositional reasoning, semantic relations, and entity recognition; OR-SpaPose-4B shows the same overall trend. Gains are smaller on nearest-neighbor and spatial-relation questions because the supervised single-view baseline is already strong on many visible entities and simple pairwise relations, but multi-view structure still improves the overall score. General MLLMs are much weaker on spatial grounding, and Spatial-MLLM, despite being designed for spatial understanding, still lags behind OR-SpaPose-8B. On the closed-source subset, strong API models do relatively well on entity-like recognition but remain weak on reference-view directions and nearest-neighbor reasoning, indicating that OR spatial QA requires explicit scene structure rather than prompting alone.

5.3 Ablation Study

Table 4 shows that geometry is the main contributor, raising the overall score by +2.96 on 4B and +4.80 on 8B. Its gains are most visible on nearest-neighbor, entity, and compositional questions, where multi-view layout directly supports object presence, proximity, and cross-object reasoning. Pose gives smaller overall gains (+1.67 on 4B and +1.55 on 8B) but is useful for relation-oriented questions, especially when human configuration helps disambiguate interactions. The full model obtains the best overall score for both backbones, although gains are not fully additive: geometry-only remains stronger on some nearest-neighbor cases, while pose+geometry is strongest on relation and entity scores. This suggests that the cues are complementary but occasionally interfere when pose or reconstruction evidence is noisy or

Table 4: **Ablation study.** Δ denotes the overall gain over the corresponding Qwen3-VL baseline.

Backbone	Variant	Near.	Rel-sp.	Rel-sem.	Entity	Comp.	Overall	Δ
4B	Base	61.34	80.37	51.11	80.04	58.41	69.43	–
	+Pose	61.11	81.61	44.44	83.71	**62.68**	71.10	+1.67
	+Geometry	**62.85**	83.37	51.11	**86.24**	61.06	72.39	+2.96
	+Pose+Geometry	62.15	**83.47**	**57.78**	85.20	62.39	**72.53**	+3.10
8B	Base	63.43	78.10	55.56	86.25	61.36	71.11	–
	+Pose	63.77	80.79	64.44	86.56	62.83	72.66	+1.55
	+Geometry	**67.94**	80.27	62.22	89.72	**71.09**	75.91	+4.80
	+Pose+Geometry	65.74	**82.44**	**66.67**	**90.37**	**71.09**	**76.22**	+5.11

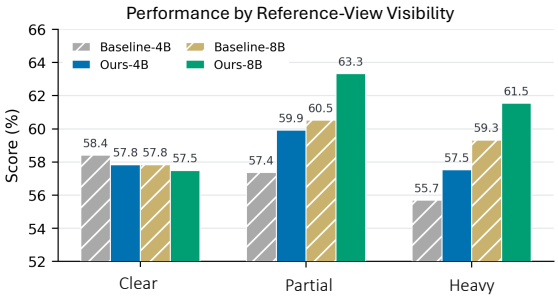

Figure 4: **Visibility-stratified performance.** Test samples are grouped into *clear*, *partial*, and *heavy* bins using the reference-view visibility score defined in Section 3.

irrelevant to the question.

5.4 Reference-view Visibility Analysis

Figure 4 shows that OR-SpaPose improves most consistently in the *partial* and *heavy* groups, where the reference view is missing or weakening relevant evidence. Gains are smaller in the *clear* group, which is expected because the single-view baseline already sees the key entities. Figure 5 illustrates three common failure modes. Single-view baselines may misread view-dependent directions under occlusion or truncation; they may reduce human interaction to generic proximity, e.g., predicting `closeTo` when two staff members are near each other but missing the clinical relation such as `assisting`; and they often fail to compose relations across multiple objects. OR-SpaPose uses auxiliary views to recover missing spatial evidence, pose/geometry cues to ground human-centered interaction labels, and explicit structural evidence to connect nearest, directional, and semantic facts in compositional questions.

6 Conclusion

We presented OR-SpaPose, an RGB-only multi-view MLLM for operating-room spatial QA. The model injects pose and VGGT-based geometry into a frozen Qwen3-VL backbone, improving relation, entity, and compositional reasoning on our MM-OR-derived benchmark. Ablations and visibility analysis show that geometry and pose provide complementary evidence, especially when the reference view is occluded or ambiguous. Future work will test broader procedures and camera layouts.

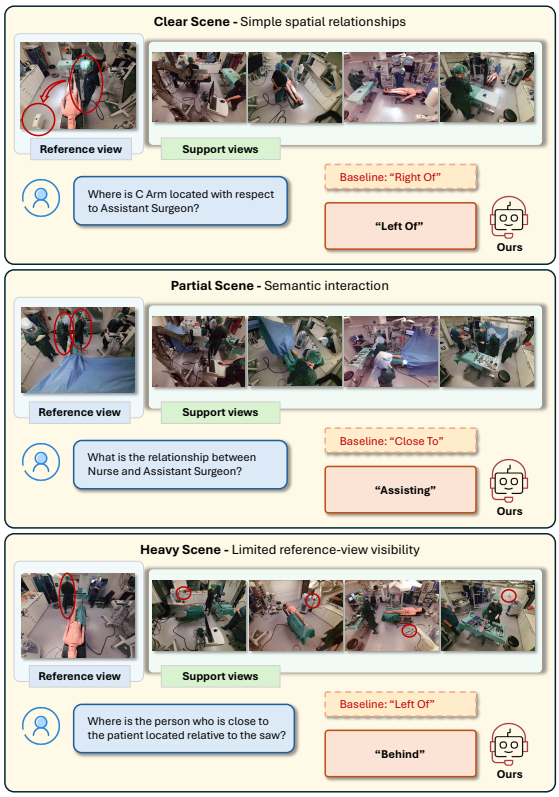

Figure 5: **Qualitative reference-view occlusion cases.** Examples show questions where answer-relevant entities or relations are occluded, truncated, or ambiguous in the reference view, and where synchronized auxiliary views help OR-SpaPose recover the correct answer.

Limitations

Evaluation on medical spatial QA is constrained by the availability and heterogeneity of public surgical datasets: existing resources differ in camera layout, annotation schema, sensor modality, and release format. We therefore conduct a controlled study on MM-OR, where synchronized RGB views and expert masks support consistent QA construction and scoring. Extending the benchmark to other procedures, hospitals, and annotation formats remains important future work.

References

Shuai Bai and 1 others. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Laura Bragagnolo, Matteo Terreran, Davide Allegro, and Stefano Ghidoni. 2024. Multi-view pose fusion for occlusion-aware 3d human pose estimation. In *ECCV*, pages 117–133.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, pages 14455–14465.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *NeurIPS*, 37:135062–135093.

Idris Hamoud, Vinkle Srivastav, Muhammad Abdullah Jamal, Didier Mutter, Omid Mohareri, and Nicolas Padoy. 2025. Multi-view video-pose pretraining for operating room surgical activity recognition. *arXiv preprint arXiv:2502.13883*.

Peiqi He, Zhenhao Zhang, Yixiang Zhang, Xiongjun Zhao, and Shaoliang Peng. 2025. Spatial-ormllm: Improve spatial relation understanding in the operating room with multimodal large language model. *arXiv preprint arXiv:2508.08199*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Yiheng Li, Ruibing Hou, Hong Chang, Shiguang Shan, and Xilin Chen. 2025. Unipose: A unified multimodal framework for human pose comprehension, generation and editing. In *CVPR*, pages 27805–27815.

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. 2025. Can multimodal large language models understand spatial relations? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 620–632.

Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, and 1 others. 2017. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696.

Ege Özsoy, Evin Pınar Örnek, Ulrich Eck, Tobias Czempiel, Federico Tombari, and Nassir Navab. 2022. 4d-or: Semantic scene graphs for or domain modeling. In *International conference on medical image computing and computer-assisted intervention*, pages 475–485. Springer.

Ege Özsoy, Chantal Pellegrini, David Bani-Harouni, Kun Yuan, Matthias Keicher, and Nassir Navab. 2026. Specialized foundation models for intelligent operating rooms. *npj Digital Medicine*.

Ege Özsoy, Chantal Pellegrini, Tobias Czempiel, Felix Tristram, Kun Yuan, David Bani-Harouni, Ulrich Eck, Benjamin Busam, Matthias Keicher, and Nassir Navab. 2025. Mm-or: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments. In *CVPR*, pages 19378–19389.

Vinkle Srivastav, Keqi Chen, and Nicolas Padoy. 2024. Selfpose3d: self-supervised multi-person multi-view 3d pose estimation. In *CVPR*, pages 2502–2512.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306.

Zhang Wenyu, En Ng Wei, Ma Lixin, Wang Yuwen, Zhao Jungqi, Koenecke Allison, Li Boyang, and Wang Lu. 2024. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. *arXiv preprint arXiv:2412.12693*, 25.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Duo Zheng, Shijia Huang, and Liwei Wang. 2025. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *CVPR*, pages 8995–9006.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. 2024. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*.

A Dataset Construction Details

Sampling and split. We sample one candidate frame every 10 frames from the MM-OR surgeries and retain frames with the required synchronized camera views and expert masks. Each final QA item is paired with five RGB views. Table 5 reports the surgery-level split.

Geometry and facts. VGGT is used to estimate camera parameters, dense depth maps, and point maps from synchronized RGB views. Expert masks identify object pixels, which are lifted

Table 5: **Dataset split.**

Split	Surgeries	Scenes	QA pairs	QA/scene
Train	18	4,881	58,425	11.97
Test	2	130	1,560	12.00
Total	20	5,011	59,985	11.97

Table 6: **Model size and inference cost.**

Model	Views	Params	Trainable	Time
Qwen3-VL-4B	1	6.00B	27.27M	0.225
OR-SpaPose-4B	5	6.65B	350.87M	0.530
Qwen3-VL-8B	1	10.36B	40.12M	0.293
OR-SpaPose-8B	5	11.01B	396.75M	0.598

into 3D and merged across views. For each object instance, outliers are removed before fitting an oriented 3D box. We use the resulting centroids for nearest-neighbor facts and transform centroids into the camera coordinate frame for reference-view directional labels. Semantic interaction facts are taken from the available MM-OR relation annotations.

Visibility score. Using the definition in Section 3, QA-associated entity-view records are 29.7% *clear*, 18.7% *partial*, 6.7% *heavy*, 4.9% *absent*, and 40.0% *unknown*.

Generation and filtering. Candidate QA pairs are generated with deterministic templates over the fact graph. Each candidate stores its supporting facts, involved entities, reference view, support views when needed, and visibility metadata. We remove candidates with invalid types, empty questions, duplicate entities, missing source facts, invalid answers, or invalid visibility records. For spatial questions, a reference view is required. For compositional questions, we require the relevant supporting facts and at least one support view when the question is cross-view.

Paraphrasing. To reduce template artifacts, Qwen3-14B paraphrases questions under strict constraints: preserve all required entities, keep the answer unchanged, avoid camera or view identifiers, avoid answer leakage, and output exactly one JSON field. Rewrites that fail JSON parsing, drop required entities, mention view metadata, contain meta text, or reveal the answer are rejected and fall back to the original template. This step changes only the surface form; answers and supporting facts remain fixed.

B Implementation Details

Training. All Qwen3-VL variants are trained with supervised fine-tuning for one epoch using AdamW, cosine learning-rate scheduling, bf16 precision, gradient checkpointing, and DeepSpeed ZeRO-2. The per-device batch size is 1 with 16 gradient accumulation steps. Images are constrained to 784–50,176 pixels and the maximum sequence length is 4,096. Base Qwen3-VL SFT baselines use a learning rate of 2×10^{-5}; pose/VGGT adapter variants use 5×10^{-6} with gradient clipping at 0.3.

Trainable components. The Qwen3-VL language model, Qwen3-VL visual encoder, VGGT backbone, and pose encoder are frozen. We train the Qwen visual merger/projector, VGGT connector and view embeddings, pose adapter, and residual gates. This setup isolates the contribution of lightweight structural adapters while keeping the backbone fixed.

Efficiency. Table 6 reports model size and inference cost. Runtime is measured on an NVIDIA A40 with batch size 1 over 100 test samples after 10 warm-up samples. The measurement includes preprocessing, optional pose/VGGT preparation, auxiliary encoders, and decoding, but excludes disk I/O and model loading.

C Evaluation Protocol

Scoring. Entity and nearest-neighbor questions are evaluated with recall over canonical class labels. Semantic relation, yes/no, and other closed-set questions use normalized exact match after parsing predictions into the canonical answer space. Directional answers use the same spatial partial-credit scorer for both `relation/spatial` and directional `complex/*` questions. Single-axis labels such as `leftOf` or `behind` require an exact canonical match. Diagonal labels such as `leftFront` are decomposed into two components, e.g., `leftOf` and `inFrontOf`; a prediction matching both components receives 1.0, a prediction matching one non-contradictory component receives 0.5, and a prediction containing an opposite component receives 0.0.

Closed-source models. Closed-source MLLMs are evaluated on a stratified subset sampled from the test set with seed 42. The sampler targets 50 items from each major task group and uses all available semantic-relation items, resulting in 245 examples. Each example is enriched with paths to the five synchronized views. The prompt asks the model to use all views, follow the reference-view

direction convention, and return only the final answer from the allowed label space. API decoding uses temperature 0 and a maximum of 128 tokens. Outputs are normalized with the same canonical parsers used for open-source model scoring.