LangOR: 3D Language Field Reconstruction for Operating Room

Wei Li[1], Ruiyang Li[1(✉)], Jialun Pei[1], Haowei Sun[2], Shi Qiu[1], and Pheng-Ann Heng[1]

[1] Department of Computer Science and Engineering and the Institute of Medical Intelligence and XR, The Chinese University of Hong Kong, Hong Kong, China
{liruiruiyang}@outlook.com
[2] South China University of Technology, China

Abstract. The advancement of intelligent operating rooms (OR) necessitates a comprehensive semantic understanding of surgical environments. However, existing methods are primarily confined to 2D perception and depend heavily on large-scale annotated data, which restricts spatial reasoning and incurs prohibitive annotation costs. To address these challenges, we propose LangOR, the first annotation-free framework designed for 3D language field reconstruction of ORs. LangOR leverages domain-constrained structured prompting to enhance the robustness of Vision-Language Models (VLMs) in zero-shot surgical domain inference, enabling precise segmentation and interpretation of surgical personnel and equipment in 2D images. We then employ geometric constraints and contrastive learning to achieve instance field reconstruction in 3D Gaussian Splatting (3DGS) representation from sparse views. Furthermore, we introduce an instance-semantics association module that effectively lifts semantic information from 2D masks to 3D instances for cross-view consistency. In this manner, LangOR decomposes the given surgical scene into OR semantic instances, facilitating comprehensive spatial interpretation. Extensive evaluations demonstrate that LangOR significantly outperforms state-of-the-art methods, reaching 63.52% mIoU on 4D-OR and 56.05% mIoU on MM-OR datasets. The results reveal that our scene understanding approach holds substantial promise for next-generation intelligent ORs. Source code is available at: https://github.com/Selena1105/LangOR.

Keywords: Language Field Reconstruction · Scene Understanding · Operating Room · 3D Gaussian Splatting.

1 Introduction

Scene understanding in the operating room (OR) identifies and segments surgical entities [5, 6, 13–15, 22] to establish the contextual awareness vital for intelligent surgical systems, enabling safe robotic navigation, automated surgical transcription, and optimized resource allocation [4, 10]. These capabilities improve surgical efficiency, minimize errors, and enhance overall patient safety [12]. While existing

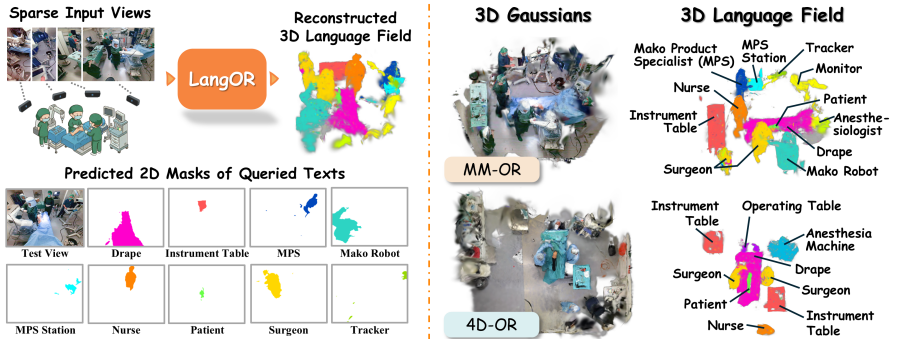

Fig. 1. We propose LangOR for 3D language field reconstruction of the operating room, achieving superior performance on semantic understanding and text retrieval.

methods have extensively explored scene understanding [2, 4, 6, 10], they rely on dense manual annotations that are labor-intensive and pose privacy risks [14]. Hence, we aim to construct a zero-shot language field of the OR, jointly reconstructing physical scene and semantics in a unified 3D representation (Fig. 1).

Existing open-vocabulary language field reconstruction methods [3, 16, 21, 25] utilize 3D Gaussian Splatting (3DGS) [8] as reprensentations, and optimize 3D semantics via two distinct approaches. Render-based methods align rendered features with 2D semantic maps [16, 17, 21] but degrade 3D semantics, whereas lifting-based methods directly project 2D features onto 3D representations [3, 7, 25]. However, both methods struggle in the OR due to: (1) *Lack of OR domain knowledge.* Relying on vision-language alignment of CLIP [18], they can only distinguish general categories (e.g., 'table'), but fail on OR-specific entities (e.g., 'anesthesia machine'). (2) *Reliance on dense views.* OR data typically presents extreme viewpoint sparsity and occlusions [14]. Under sparse observations, render-based methods generate view-inconsistent semantic maps, and lifting-based methods also fail in reliable 3D projection.

To address these challenges, we propose LangOR, a 3D language field reconstruction framework of the OR. To the best of our knowledge, LangOR is the first annotation-free method for OR semantic scene understanding, which enables OR segmentation and interpretation directly in 3D space and guarantees robust language field reconstruction under sparse views (see Fig. 2). First, we propose a **Robust Vision-Language Model (VLM) based OR Perceptor** to perform zero-shot prediction of paired 2D masks and semantics. Specifically, we leverage a context-aware prompting strategy to achieve zero-shot grounding of OR scenarios using a VLM and further improve the robustness of OR prediction from multi-round responses with a semantic mask refinement module. Second, we introduce an **OR Instance Field Reconstruction** module that reconstructs 3D scene under geometric constraints and learns instance features via mask-aware contrastive learning. Third, we associate 2D semantics from the VLM with 3D instances under sparse views with a **Cross-View Pixel-level**

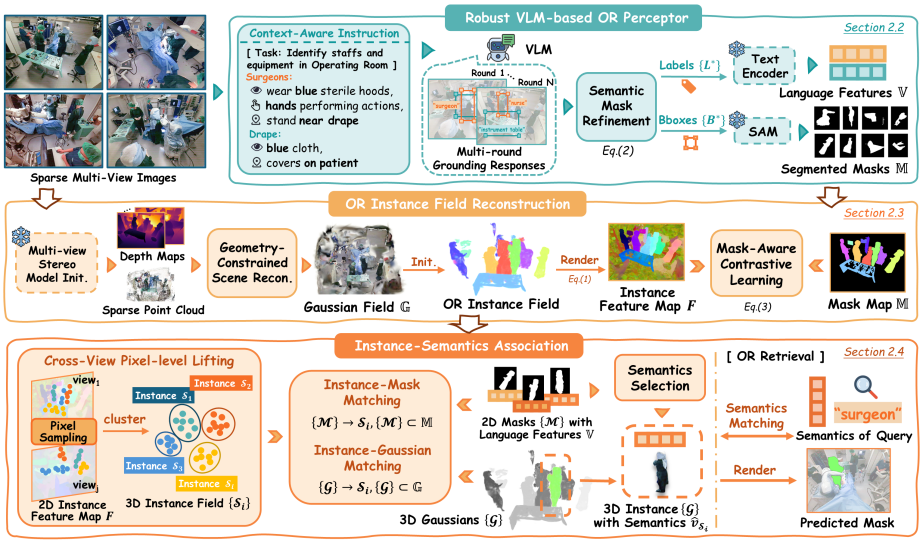

Fig. 2. The overall framework of LangOR. The proposed method achieves zero-shot 3D language field reconstruction of the operating room from sparse multi-view images.

Lifting strategy and an **Instance-Mask-Gaussian Matching** scheme. The proposed framework can ultimately reconstruct a reliable 3D language field for OR using sparse views.

Our main contributions are summarized as follows:

- We propose LangOR, the first annotation-free 3D semantic understanding framework of the OR, which integrates surgical entities and corresponding semantics in a coherent 3D space.
- We develop a Robust VLM-based OR Perceptor that utilizes domain-specific contexts to guide semantic extraction, enabling reliable zero-shot OR prediction.
- We reconstruct high-fidelity OR instance field and associate sparse 2D semantics with constructed 3D instances, achieving superior 3D semantic segmentation on two OR datasets.

2 Method

2.1 Preliminaries of Gaussian Splatting

3DGS [8] represents a 3D scene using a set of 3D Gaussians, each defined by a center $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$: $\mathcal{G}(\boldsymbol{x}) = e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$. These Gaussians are then projected to the image plane via tile-based rasterization. The pixel color $\boldsymbol{C}$ is computed by α-blending the overlapping Gaussians. To incorporate semantic information, recent methods augment each Gaussian with a

high-dimensional feature $\boldsymbol{f}_i$ [16,21,27]. The pixel color $\boldsymbol{C}$ and rendered semantic feature map $\boldsymbol{F}$ are calculated as below:

$$\boldsymbol{C} = \sum_{i \in \mathcal{N}} \boldsymbol{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \; \boldsymbol{F} = \sum_{i \in \mathcal{N}} \boldsymbol{f}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{1}$$

where $\boldsymbol{c}_i$ is the color of the i-th Gaussian, α_i denotes its opacity scaled by the 2D projected density, and $\mathcal{N}$ is the set of ordered gaussians overlapping the pixel.

2.2 Robust VLM-based OR Perceptor

Current VLMs struggle with fine-grained OR semantics due to high visual similarity within the categories of surgical staff (e.g., 'surgeon' vs. 'nurse') and equipment, as well as inevitable hallucinations and inconsistent predictions. To ensure robust OR perception, our Context-Aware Prompting strategy models clinicians' behavior and spatial layout as discriminative contexts for visually similar entities. We further use a Semantic Mask Refinement Module to obtain robust masks and semantics from zero-shot predictions.

Context-Aware Prompting (CAP) Strategy. CAP strategy guides VLM grounding in the spatial layout and clinicians' behaviors in the OR, following strict clinical protocols. We model OR entities (personnel and equipment) using three contextual descriptors: (1) *Spatial Context* (C_{spat}) describes relative positions (e.g., 'standing near the instrument table'); (2) *Behavior Context* (C_{beh}) represents the behaviors of surgical personnel (e.g., 'holding a surgical tool'); (3) *Visual Appearance Context* (C_{vis}) captures distinguishable physical features (e.g., 'wearing a blue sterile hood, covered by green cloth'). Personnel are defined as $C_p = \{C_{spat}, C_{beh}, C_{vis}\}$, while equipment is described by $C_e = \{C_{spat}, C_{vis}\}$. These contexts form a structured VLM instruction to generate preliminary bounding boxes (B) and semantic labels (L) of OR entities.

Semantic Mask Refinement Module. We further improve the semantic prediction reliability through multi-round inference. Specifically, we collect t candidates $\{B_i, L_i\}_{i=1}^{t}$ for each image from N rounds of responses. Bounding boxes $\{B_i\}$ with Intersection over Union (IoU) above a threshold are clustered. The representative box B_k^* of the k-th cluster $\mathcal{U}_k$ is defined as:

$$B_k^* = \underset{B_i \in \mathcal{U}_k}{\arg\max} \left[\frac{1}{t-1} \sum_{j=1, j \neq i}^{t} \mathrm{IoU}(B_i, B_j) \right]. \tag{2}$$

The boxes $\{B^*\}$ are then used as visual prompts for SAM [9] to generate masks $\mathbb{M}$. The semantic labels $\{L^*\}$ are then obtained by majority voting within each cluster and encoded into semantic features $\mathbb{V}$ using a text encoder [1].

2.3 OR Instance Field Reconstruction

We reconstruct the OR instance field through a two-stage process: first, we establish an accurate 3D scene reconstruction via geometric priors; then, we

embed instance features into the 3D scene with mask-aware contrastive learning. Unlike language-aligned semantic features, these instance features are specifically designed to partition the reconstructed scene into discrete 3D instances.

Geometry-Constrained Scene Reconstruction. Our OR Instance Field utilizes explicit 3DGS [8] as representation. Standard Structure-from-Motion (SfM) [19] or Multi-View Stereo (MVS) [20] initialization often fails in OR due to extreme view sparsity (4–5 viewpoints) and textureless surfaces (e.g. drapes), making it difficult for previous photometric-based methods [8] to recover accurate geometry. To tackle this, we utilize a feed-forward MVS model [24] to obtain robust initial camera poses, sparse point clouds, and corresponding depth maps. Following [26], we apply geometric loss including depth, normal, and scale consistency to supervise the Gaussian field $\mathbb{G}$. This shifts the learning objective from minimizing photometric discrepancy to enforcing geometric consistency, ensuring high-fidelity scene reconstruction for subsequent language field reconstruction.

Mask-Aware Contrastive Learning of Instance Features. Inspired by previous 3D cross-view semantic learning methods [3, 25], we adopt contrastive learning to optimize the instance features. Specifically, we render the 3D Gaussian instance features $\boldsymbol{f}_g$ into a 2D feature map $\boldsymbol{F}$ following Eq. 1 for an arbitrary training view. Using the masks $\mathbb{M}$ obtained in Section 2.2, we compute the centroid feature $\hat{\boldsymbol{f}}_m$ for a target instance mask $\mathcal{M}$ as $\hat{\boldsymbol{f}}_m = \frac{1}{|\mathcal{M}|}\sum_{\boldsymbol{p}\in\mathcal{M}} \boldsymbol{F}(\boldsymbol{p})$, where $\boldsymbol{p}$ denotes the image pixel. Then, we apply a contrastive loss to the rendered feature map to ensure intra-instance consistency and inter-instance separation:

$$\mathcal{L}_{con} = \sum_{\mathcal{M}\in\mathbb{M}} \left(\sum_{\boldsymbol{p}\in\mathcal{M}} \mathcal{L}_{pos}(\boldsymbol{p}) + \lambda \sum_{\boldsymbol{p}\notin\mathcal{M}} \mathcal{L}_{neg}(\boldsymbol{p}) \right), \tag{3}$$

where $\mathcal{L}_{pos}(\boldsymbol{p}) = \max(0, 1-\cos(\hat{\boldsymbol{f}}_m, \boldsymbol{F}(\boldsymbol{p})))$ encourages features within the same mask be close to their centroid, while $\mathcal{L}_{neg}(\boldsymbol{p}) = \max(0, \cos(\hat{\boldsymbol{f}}_m, \boldsymbol{F}(\boldsymbol{p})))$ penalizes cosine similarity between the target centroid and features of other instances.

2.4 Instance-Semantics Association

Cross-View Pixel-level Lifting (CVP-Lifting). We introduce a pixel-level cross-view lifting strategy to robustly associate 2D masks with 3D instances under sparse views and ensure cross-view consistency.

To mitigate boundary inaccuracies caused by occlusions in sparse views, we employ weighted random sampling [23] on the rendered feature map. We assign higher weights to pixels near the mask center using a Gaussian distribution, scaled by the inverse mask area to ensure small objects are sufficiently sampled. This location-based weight of a pixel $\boldsymbol{p}$ belonging to a 2D mask $\mathcal{M}$ is defined as:

$$w(\boldsymbol{p}) = \frac{1}{|\mathcal{M}|} \exp\left(-\frac{\|\boldsymbol{p} - \boldsymbol{c}_{\mathcal{M}}\|^2}{2\sigma^2} \right), \tag{4}$$

where $\|\boldsymbol{p}-\boldsymbol{c}_{\mathcal{M}}\|$ is the Euclidean distance between $\boldsymbol{p}$ and the mask center $\boldsymbol{c}_{\mathcal{M}}$, and σ is a hyperparameter controlling the spatial fall-off of the Gaussian distribution.

After sampling, we lift sampled pixels across views into 3D instances $\{\mathcal{S}_i\}$ by hierarchical density-based clustering [11] of pixel-level instance features. For instance $\mathcal{S}_i = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_s\}$, we compute an instance prototype $\hat{\boldsymbol{f}}_{\mathcal{S}_i}$ as below:

$$\hat{\boldsymbol{f}}_{\mathcal{S}_i} = \frac{1}{|\mathcal{S}_i|} \sum_{\boldsymbol{p} \in \mathcal{S}_i} \boldsymbol{F}(\boldsymbol{p}), \tag{5}$$

where $\boldsymbol{F}(\boldsymbol{p})$ denotes the rendered instance feature of the pixel $\boldsymbol{p}$. This process lifts 2D pixels to 3D instances, bridging the gap for subsequent matching.

Instance-Mask-Gaussian Matching. For each unsampled pixel $\boldsymbol{p}_u$, we assign it to an instance $\mathcal{S}_i$ with the maximal cosine similarity between its feature $\boldsymbol{F}(\boldsymbol{p}_u)$ and the instance prototype $\hat{\boldsymbol{f}}_{\mathcal{S}_i}$. Then each 2D mask $\mathcal{M}$ is associated with the most frequent instance $\mathcal{S}_i$ among its pixels $\{\boldsymbol{p}\}$. Similarly, we match each 3D Gaussian $\mathcal{G}$ with the instance whose prototype $\hat{\boldsymbol{f}}_{\mathcal{S}_i}$ exhibits the highest cosine similarity to the Gaussian's instance feature $\boldsymbol{f}_g$. We summarize these matching relationships as:

$\{\mathcal{M}\} \to \mathcal{S}_i, \{\mathcal{M}\} \subset \mathbb{M}; \{\mathcal{G}\} \to \mathcal{S}_i, \{\mathcal{G}\} \subset \mathbb{G}$.

We determine the instance's semantics $\hat{\boldsymbol{v}}_{\mathcal{S}_i}$ by selecting a semantic label from the semantic features $\{V\}$ of its associated masks $\{\mathcal{M}\}$. The selected semantics has the highest cosine similarity to others. When the mask count is ≤ 2, we use the average semantic feature instead. In this way, we achieve 3D language field reconstruction for OR scenes. For OR retrieval, we compute the similarity between the query text embedding and each instance's semantics $\hat{\boldsymbol{v}}_{\mathcal{S}_i}$ to retrieve the corresponding 3D Gaussians and render the mask onto 2D views.

3 Experiments

3.1 Experiment Settings

Datasets and Evaluation Metrics. We build our dataset from two public OR benchmarks: MM-OR [14] and 4D-OR [13]. Both datasets provide 2048×1536 5-view videos of surgeries, including 2D segmentations for 3 specific views. MM-OR and 4D-OR include 12 and 8 object classes, respectively. We construct 10 scenes per dataset by extracting 5 viewpoints at 10 timepoints across various surgical phases. For each scene, we use a 4:1 train-test viewpoint random split.

We evaluate performance using three metrics: mIoU evaluates the average IoU between predicted ($\mathcal{M}_{pred}$) and ground-truth (GT) masks ($\mathcal{M}_{gt}$); mLocAcc measures the percentage of predicted masks whose pixel with the highest semantic relevancy correctly falls inside the GT mask [16]; mAcc calculates the pixel-wise accuracy as: $\text{mAcc} = \frac{1}{K} \sum_{i=1}^{K} (|\mathcal{M}_{gt}^i \cap \mathcal{M}_{pred}^i|)/(|\mathcal{M}_{gt}^i|)$, where K denotes the number of categories.

Implementation Details. We use Qwen3-VL-235B-A22B-Instruct as the backbone VLM (Sec. 2.2) and conduct 5 independent trials for each scene. We set the optimization hyperparameters $\lambda = 0.2$ in Eq. 3 and $\sigma = 0.5$ in Eq. 4. The training process consists of two stages: 500 iterations to build a high-fidelity 3D

Table 1. Quantitative comparison on 4D-OR [13] and MM-OR [14] datasets. **Bold** and underline represent the best and second-best performances, respectively. '†' denotes methods enhanced with OR Perceptor.

Methods	4D-OR			MM-OR		
	mIoU$_{(\%)}$ ↑	mLocAcc$_{(\%)}$ ↑	mAcc$_{(\%)}$ ↑	mIoU$_{(\%)}$ ↑	mLocAcc$_{(\%)}$ ↑	mAcc$_{(\%)}$ ↑
LEGaussians [21]	$5.36_{\pm 2.83}$	$17.86_{\pm 14.21}$	$32.38_{\pm 8.97}$	$2.71_{\pm 0.93}$	$14.87_{\pm 9.62}$	$16.66_{\pm 8.51}$
LangSplat [16]	$5.18_{\pm 3.20}$	$3.33_{\pm 7.03}$	$17.96_{\pm 13.20}$	$3.23_{\pm 2.27}$	$7.56_{\pm 7.17}$	$10.17_{\pm 6.28}$
OpenGaussian [25]	$14.78_{\pm 9.38}$	/	$33.16_{\pm 18.38}$	$5.57_{\pm 3.12}$	/	$13.85_{\pm 8.63}$
LaGa [3]	$19.44_{\pm 9.11}$	$48.95_{\pm 49.63}$	$50.72_{\pm 14.22}$	$10.55_{\pm 3.06}$	$40.23_{\pm 15.66}$	$22.57_{\pm 8.12}$
LangSplat†	$18.32_{\pm 4.70}$	$54.52_{\pm 21.69}$	$47.37_{\pm 14.02}$	$8.98_{\pm 4.96}$	$26.74_{\pm 13.41}$	$19.38_{\pm 7.9}$
OpenGaussian†	$20.34_{\pm 10.57}$	/	$36.61_{\pm 13.47}$	$15.67_{\pm 4.65}$	/	$36.09_{\pm 13.55}$
LaGa†	$\underline{41.48}_{\pm 13.61}$	$\underline{88.05}_{\pm 14.48}$	$\mathbf{80.43}_{\pm 8.08}$	$\underline{29.25}_{\pm 7.21}$	$\underline{55.61}_{\pm 10.77}$	$\mathbf{74.47}_{\pm 4.53}$
LangOR(ours)	$\mathbf{63.52}_{\pm 8.05}$	$\mathbf{92.57}_{\pm 9.73}$	$\underline{74.95}_{\pm 7.12}$	$\mathbf{56.05}_{\pm 5.87}$	$\mathbf{90.62}_{\pm 7.69}$	$\underline{68.93}_{\pm 7.24}$

scene representation (Sec. 2.3) and 300 iterations for instance field reconstruction (Sec. 2.3). All experiments are conducted on a single NVIDIA RTX 3090 GPU. During inference, LangOR achieves a speed of 47.1 FPS, and consumes 5 GB of VRAM and 2.6 GB of RAM.

3.2 Quantitative and Qualitative Results

We compare LangOR against four state-of-the-art (SOTA) methods: render-based (LangSplat [16] and LEGaussians [21]) and lifting-based (OpenGaussian [25] and LaGa [3]), all initialized with the same sparse pointclouds generated by VGGT [24]. mLocAcc is inapplicable to OpenGaussian due to the absence of pixel-level relevancy.

In Tab. 1, LangOR achieves an mIoU of 63.52%, tripling that of the SOTA method LaGa (19.44%) on 4D-OR dataset. LangOR also reaches a 92.57% of mLocAcc, successfully localizing almost all target objects. In the complex MM-OR dataset, LangOR maintains high performance with 56.05% mIoU and 90.62% mLocAcc, highlighting LangOR's superior semantic-spatial alignment and cross-view consistency for precise localization and segmentation in OR environments. Moreover, we integrate OR-Perceptor into other methods and yield substantial gains, which demonstrates the OR-Perceptor's effectiveness and versatility in OR scene understanding, though SOTA methods enhanced with OR-Perceptor remain inferior to our method. The Friedman test and Bonferroni-corrected post-hoc Wilcoxon signed-rank tests confirm that LangOR significantly outperforms all SOTA methods with $p < 0.001$ on mIoU.

Since render-based approaches fail to generate effective masks under sparse views, we focus our qualitative comparison on lifting-based methods (Fig. 3). While original lifting methods suffer from inaccurate semantics due to a lack of domain-specific knowledge, they show significant improvement with OR-Perceptor. Notably, our instance field reconstruction and instance-semantic association strategies successfully distinguish visually similar categories (e.g., 'surgeon' vs. 'nurse')

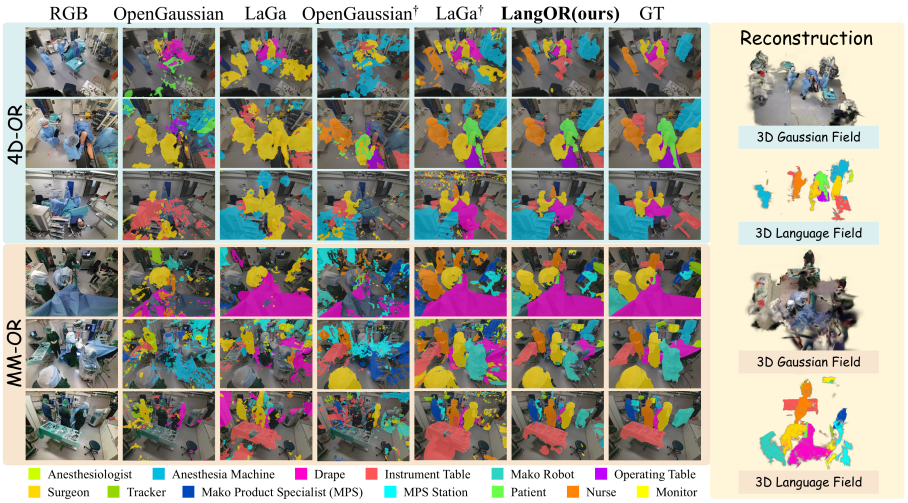

Fig. 3. Qualitative comparison. We visualize the segmentations of partial text queries of 6 distinct scenes and test views for 4D-OR [13] (top 3 rows) and MM-OR [14] (bottom 3 rows). LangOR achieves the most accurate segmentations across diverse text queries.

Table 2. Ablation study on 4D-OR [13] and MM-OR [14] datasets.

Methods	4D-OR			MM-OR		
	mIoU(%) ↑	mLocAcc(%) ↑	mAcc(%) ↑	mIoU(%) ↑	mLocAcc(%) ↑	mAcc(%) ↑
w/o OR Perceptor	$8.69_{\pm 5.61}$	$24.86_{\pm 13.85}$	$11.17_{\pm 5.81}$	$10.03_{\pm 5.62}$	$17.01_{\pm 9.75}$	$10.69_{\pm 6.88}$
w/o Geometric Loss	$59.64_{\pm 10.89}$	$89.48_{\pm 11.48}$	$69.46_{\pm 9.73}$	$51.06_{\pm 7.74}$	$\underline{84.88}_{\pm 12.01}$	$60.80_{\pm 9.36}$
w/o CVP-L.	$58.71_{\pm 14.39}$	$84.95_{\pm 19.18}$	$67.67_{\pm 14.74}$	$44.73_{\pm 7.85}$	$65.98_{\pm 12.93}$	$55.88_{\pm 8.75}$
w/ CVP-L. ($\sigma = 0.7$)	$\underline{63.35}_{\pm 8.67}$	$\mathbf{92.57}_{\pm 9.73}$	$\underline{74.83}_{\pm 7.17}$	$\underline{55.72}_{\pm 5.88}$	$\mathbf{90.62}_{\pm 7.69}$	$\underline{68.44}_{\pm 6.74}$
LangOR(ours)	$\mathbf{63.52}_{\pm 8.05}$	$\mathbf{92.57}_{\pm 9.73}$	$\mathbf{74.95}_{\pm 7.12}$	$\mathbf{56.05}_{\pm 5.87}$	$\mathbf{90.62}_{\pm 7.69}$	$\mathbf{68.93}_{\pm 7.24}$

and identify complex equipment like anesthesia machines. Even under sparse views, our method generates consistent masks with sharp boundaries.

3.3 Ablation Study

We conduct an ablation study to evaluate OR Perceptor, Geometric Loss, and CVP-Lifting strategies on both datasets (see Tab. 2).

(1) *OR-Perceptor.* We replace OR-Perceptor with the SAM-CLIP strategy [16], causing a drastic mIoU drop from 63.52% to 8.69%.

(2) *Geometric Loss.* Scene reconstruction without geometry constraints decreases the mIoU by 6.1%/8.9%, demonstrating the effectiveness of geometric-constrained scene reconstruction on instance field representation. (3) *CVP-Lifting (CVP-L.).* We compare our CVP-Lifting against a mask-level strategy [3]. Our full model outperforms baselines by 8.2%/25.3% in mIoU and 9.0%/37.3% in mLocAcc, suggesting that finer-grained lifting better associates 2D-3D infor-

mation under sparse views. Furthermore, changing σ to 0.7 results in only a negligible mIoU decrease, proving the robustness of CVP-Lifting strategy.

4 Conclusion

In this paper, we propose LangOR, the first annotation-free method for 3d language field reconstruction of the OR. Specifically, we propose a robust VLM-based OR perceptor that leverages context-aware prompting to extract accurate OR scene semantics from multi-round VLM responses. Then, we introduce an OR instance field reconstruction module that explicitly distinguishes OR entities in 3D space. Finally, we utilize a cross-view pixel-level lifting strategy to consistently match 3D instances and 2D semantics. Experiments on OR datasets demonstrate LangOR's state-of-the-art performance in OR semantic scene understanding, offering a vital step toward future intelligent operating rooms.

Acknowledgments. This work was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project T45-401/22-N and by the Hong Kong Innovation and Technology Fund, under Project GHP/252/23SZ.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. all-MiniLM-L6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, accessed 15-January-2026
2. Bastian, L., Derkacz-Bogner, D., Wang, T.D., Busam, B., Navab, N.: Segmentor: Obtaining efficient operating room semantics through temporal propagation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 57–67. Springer (2023)
3. Cen, J., Zhou, X., Fang, J., Wen, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Tackling view-dependent semantics in 3d language gaussian splatting. arXiv preprint arXiv:2505.24746 (2025)
4. Gao, C., Rabindran, D., Mohareri, O.: Rgb-d semantic slam for surgical robot navigation in the operating room. arXiv preprint arXiv:2204.05467 (2022)
5. Guo, D., Lin, M., Pei, J., Tang, H., Jin, Y., Heng, P.A.: Tri-modal confluence with temporal dynamics for scene graph generation in operating rooms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 714–724. Springer (2024)
6. Hamoud, I., Karargyris, A., Sharghi, A., Mohareri, O., Padoy, N.: Self-supervised learning via cluster distance prediction for operating room context awareness. International Journal of Computer Assisted Radiology and Surgery **17**(8), 1469–1476 (2022)
7. Jun-Seong, K., Kim, G., Yu-Ji, K., Wang, Y.C.F., Choe, J., Oh, T.H.: Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14137–14146 (2025)

8. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
10. Li, Z., Shaban, A., Simard, J.G., Rabindran, D., DiMaio, S., Mohareri, O.: A robotic 3d perception system for operating room environment awareness. arXiv preprint arXiv:2003.09487 (2020)
11. McInnes, L., Healy, J., Astels, S., et al.: hdbscan: Hierarchical density based clustering. J. Open Source Softw. **2**(11), 205 (2017)
12. Mohan, R., Arce, J., Mokhtar, S., Cattaneo, D., Valada, A.: Syn-mediverse: a multimodal synthetic dataset for intelligent scene understanding of healthcare facilities. arXiv preprint arXiv:2308.03193 (2023)
13. Özsoy, E., Örnek, E.P., Eck, U., Czempiel, T., Tombari, F., Navab, N.: 4d-or: Semantic scene graphs for or domain modeling. In: International conference on medical image computing and computer-assisted intervention. pp. 475–485. Springer (2022)
14. Özsoy, E., Pellegrini, C., Czempiel, T., Tristram, F., Yuan, K., Bani-Harouni, D., Eck, U., Busam, B., Keicher, M., Navab, N.: Mm-or: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19378–19389 (2025)
15. Pei, J., Guo, D., Zhang, J., Lin, M., Jin, Y., Heng, P.A.: S^2former-or: Single-stage bi-modal transformer for scene graph generation in or. IEEE Transactions on Medical Imaging **44**(1), 361–372 (2024)
16. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20051–20060 (2024)
17. Qu, Y., Dai, S., Li, X., Lin, J., Cao, L., Zhang, S., Ji, R.: Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 5328–5337 (2024)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
19. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
20. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: European conference on computer vision. pp. 501–518. Springer (2016)
21. Shi, J.C., Wang, M., Duan, H.B., Guan, S.H.: Language embedded 3d gaussians for open-vocabulary scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5333–5343 (2024)
22. Tu, M., Jung, H., Kim, J., Kyme, A.: Head-mounted displays in context-aware systems for open surgery: a state-of-the-art review. IEEE Journal of Biomedical and Health Informatics **29**(2), 1165–1175 (2024)
23. Walker, A.J.: New fast method for generating discrete random numbers with arbitrary frequency distributions. Electronics Letters **10**(8), 127–128 (1974)

24. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
25. Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., Cheng, X., Zhao, C., Feng, H., Ding, E., Wang, J., et al.: Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. Advances in Neural Information Processing Systems **37**, 19114–19138 (2024)
26. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depthsplat: Connecting gaussian splatting and depth. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16453–16463 (2025)
27. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21676–21685 (2024)