ENDO4DLR: ONLINE LANGUAGE-ALIGNED 4D RECONSTRUCTION WITH PERSISTENT GAUSSIAN ANCHORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamic intraoperative scene understanding requires jointly recovering evolving 3D geometry and grounding language semantics in reconstructed scenes. Yet performing both tasks online from unposed endoscopic streams remains challenging due to camera motion, tissue deformation, and changing visibility. Existing methods either rely on per-scene optimization or focus primarily on geometric reconstruction, leaving this joint problem largely unexplored. In this work, we introduce ***Endo4DLR***, a fully *feed-forward* framework for *online language-aligned 4D reconstruction* that predicts camera parameters and language-embedded Gaussians without per-scene optimization. Our key idea is to maintain persistent Gaussian anchors that connect current local observations with reusable historical scene information. In particular, we introduce adaptive observation queries that capture fine-grained details of the current frame and are organized into local 3D Gaussian groups. In addition, we introduce spatial anchors that retrieve and aggregate historical features to predict Gaussian attributes and language embeddings. By combining current geometric estimates with retrieved historical context, *Endo4DLR* adapts to scene evolution while maintaining temporal consistency in reconstruction and language alignment. Extensive experiments on multiple endoscopic datasets demonstrate that *Endo4DLR* outperforms state-of-the-art online methods in reconstruction, camera estimation, and language grounding.

1 INTRODUCTION

Surgical navigation and embodied surgical systems depend on understanding the evolving 3D relationships between instruments and tissue (Wang et al., 2022; Zhou et al., 2025a; Long et al., 2025). Language-aligned reconstruction connects this spatial structure with descriptions of surgical activity, enabling queries such as locating an instrument performing a particular action (Qin et al., 2024; Guo et al., 2026). However, maintaining these cross-modal associations under rapid tool motion and continuous tissue deformation remains an open challenge, which motivates online language-queryable dynamic scene reconstruction from endoscopic videos.

Early neural reconstruction methods recovered surgical geometry and appearance through per-scene optimization (Wang et al., 2022). Language fields subsequently enabled natural-language queries but still required scene-specific fitting (Kerr et al., 2023; Qin et al., 2024; Li et al., 2025a), even in online mapping (Katragadda et al., 2025). Recently, feed-forward streaming models enabled reconstruction from unposed videos with no such optimization, but primarily recover geometry and appearance (Wang et al., 2025b; Lan et al., 2026; Wu et al., 2026b). SLARM (Qiu et al., 2026) further incorporates language alignment into dynamic streaming reconstruction, but assumes known camera parameters. This assumption breaks down in intraoperative settings, which require updating a language-queryable scene representation as the scene evolves without known camera parameters. Motivated by this gap, we study joint camera estimation and language-queryable reconstruction from unposed endoscopic streams, as illustrated in Fig. 1.

A natural starting point is to extend a streaming Gaussian reconstructor with language prediction (Qiu et al., 2026). However, preserving fine detail with compact Gaussian predictions while maintaining stable appearance and language features under motion and changing visibility is practi-

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.

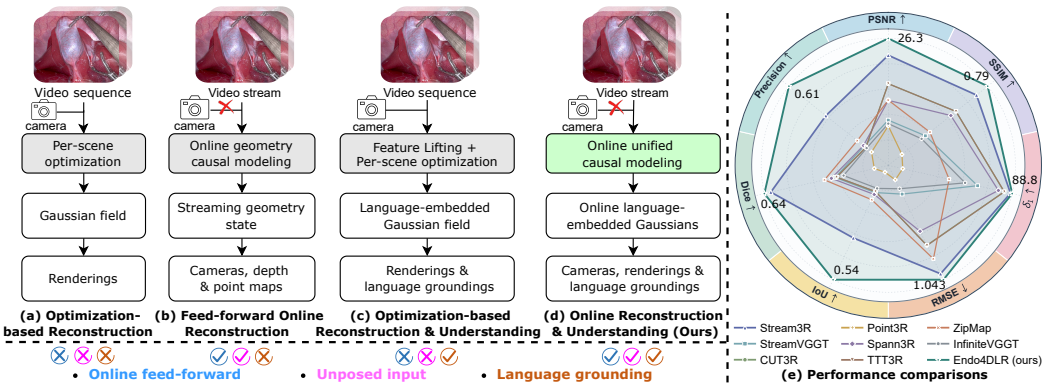


Figure 1: **Online reconstruction and language grounding.** (a–c) Existing paradigms. (d) *Endo4DLR* jointly estimates cameras and reconstructs language-queryable dynamic scenes from unposed streams without per-scene optimization. (e) Performance against online baselines.

cally difficult. Dense pixel-aligned prediction preserves local detail but introduces substantial spatial redundancy (Wang et al., 2026), while uniform subsampling may overlook thin instruments and tissue boundaries. Moreover, the resulting Gaussians also need to be associated across frames, so that previously captured detail can support predictions when current visibility is limited. Establishing such associations is difficult as motion and deformation change the spatial arrangement of newly predicted Gaussians (Wu et al., 2026b). These challenges demand an elaborate Gaussian representation that combines local detail with historical context for temporally coherent prediction.

To this end, we present ***Endo4DLR***, a fully feed-forward framework for online language-aligned 4D reconstruction that jointly predicts camera parameters and language-embedded Gaussians without per-scene optimization. Our core design connects local Gaussian construction with historical context retrieval through persistent Gaussian anchors. To construct these anchors, we refine query locations using local image features to capture fine detail, then group the lifted queries in 3D using Z-order serialization. The resulting anchors retrieve historical features through spatial proximity and feature compatibility, providing shared context for temporally coherent Gaussian attribute and language prediction. Experiments on five endoscopic datasets demonstrate superior reconstruction and language-grounding performance (Fig. 1(e)). Our key contributions are:

- We introduce *Endo4DLR*, the *first* fully feed-forward framework for online language-aligned dynamic scene reconstruction from unposed endoscopic videos, jointly estimating cameras and language-embedded Gaussians without per-scene optimization.
- We propose an anchor-based Gaussian organization with adaptive observation queries and explicit temporal context retrieval, connecting fine-grained local reconstruction with historical feature reuse for temporally consistent reconstruction and language grounding.
- *Endo4DLR* achieves superior performance over representative online feed-forward frameworks on multiple endoscopic benchmarks in both reconstruction quality and language grounding, demonstrating its effectiveness on previously unseen surgical scenes.

2 Related Work

Dynamic Scene Reconstruction. Dynamic reconstruction uses deformable neural fields (Pumarola et al., 2021; Tretschk et al., 2021) or time-varying Gaussians (Yang et al., 2024; Wu et al., 2024a; Lin et al., 2024; Lei et al., 2025). Surgical methods, including EndoNeRF (Wang et al., 2022), EndoGaussian (Liu et al., 2024), and SurgicalGaussian (Xie et al., 2024), reconstruct deforming tissue through per-scene optimization. Feed-forward methods recover geometry from unposed images (Wang et al., 2024; 2025a) or predict renderable Gaussians from multiple views (Charatan et al., 2024; Chen et al., 2024; Ye et al., 2025). CUT3R (Wang et al., 2025b) and Stream3R (Lan et al., 2026) support streaming reconstruction through recurrent states and causal attention, respectively. StreamGS (Li et al., 2025b) reconstructs Gaussians incrementally, while StreamSplat uses

pretrained depth priors for online dynamic reconstruction (Wu et al., 2026b). Our framework jointly estimates cameras and language-embedded Gaussians for reconstruction and language grounding.

Language-Embedded Scene Representations. Language-embedded representations support text-driven scene queries (Peng et al., 2023; Ding et al., 2023; Takmaz et al., 2023). LERF (Kerr et al., 2023) embeds language in radiance fields, while LangSplat (Qin et al., 2024) and Language Embedded 3D Gaussians (Shi et al., 2024) use Gaussian primitives, all requiring per-scene optimization. 4D LangSplat (Li et al., 2025a) extends querying to dynamic scenes, and ST4R-Splat (Meng et al., 2026) addresses spatio-temporal referring segmentation. Online Language Splatting (Katragadda et al., 2025) incrementally maps static scenes with language features but retains test-time optimization. LSM (Fan et al., 2024), Uni3R (Sun et al., 2026), and LangFlash (Liu et al., 2026) predict geometry and semantics from unposed images in static, non-streaming settings. SLARM (Qiu et al., 2026) supports language-aligned streaming dynamic reconstruction but relies on known camera parameters. Instead, we jointly estimate cameras and reconstruct language-aligned dynamic scenes online from unposed endoscopic streams, without per-scene optimization.

3 METHOD

We present *Endo4DLR*, a feed-forward framework for online reconstruction and language-grounded understanding of dynamic endoscopic scenes. Fig. 2 illustrates the overall framework. Given an unposed endoscopic video $\{\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^{T}$, we estimate a camera $\boldsymbol{\pi}_t$ and construct a language-embedded Gaussian representation $\mathcal{G}_t$ as each frame arrives through the online model $\mathcal{F}_\theta$:

$$(\boldsymbol{\pi}_t, \mathcal{G}_t, \mathbf{h}_t, \mathcal{M}_t) = \mathcal{F}_\theta(\mathbf{I}_t, \mathbf{h}_{t-1}, \mathcal{M}_{t-1}), \tag{1}$$

where $\mathbf{h}_t$ is the cached causal state, $\mathcal{M}_t$ stores persistent Gaussian anchors, and $\boldsymbol{\pi}_t \in \mathbb{R}^9$ contains camera intrinsics and extrinsics. An image encoder (Oquab et al., 2023) and a causal Transformer (Lan et al., 2026) produce shared features $\mathbf{Z}_t \in \mathbb{R}^{N_p\times C}$, with N_p tokens of dimension C. A camera head (Lan et al., 2026) and a Gaussian decoder use these features to predict $\boldsymbol{\pi}_t$ and $\mathcal{G}_t$ in parallel. The Gaussian decoder in our model utilizes adaptive queries to form local Gaussian groups organized by sparse 3D anchors. These anchors retrieve historical context from $\mathcal{M}_{t-1}$ to support Gaussian prediction and update the memory for subsequent frames. Each Gaussian has a center $\boldsymbol{\mu} \in \mathbb{R}^3$, scale $\boldsymbol{s} \in \mathbb{R}^3_+$, unit rotation quaternion $\boldsymbol{r} \in \mathbb{R}^4$, opacity $\alpha \in [0,1]$, and color $\boldsymbol{c} \in [0,1]^3$ (Kerbl et al., 2023), together with a language feature $\boldsymbol{\ell} \in \mathbb{R}^{d_\ell}$. The representation supports RGB-D rendering and language grounding in image and 3D space, associating geometry and semantics with shared Gaussian primitives to ensure their consistency. We detail Gaussian construction in Sec. 3.1, persistent anchors in Sec. 3.2, and training and inference in Sec. 3.3.

3.1 ADAPTIVE LANGUAGE-EMBEDDED GAUSSIAN CONSTRUCTION

Pixel-aligned Gaussian prediction provides detailed scene coverage but introduces substantial spatial redundancy (Wang et al., 2026). Structured Gaussian representations alleviate this redundancy by organizing local primitives around anchors, preserving fine scene detail with fewer Gaussians (Lu et al., 2024). For online reconstruction, this structure also offers a natural way to share information across frames. We therefore represent the current observation with adaptive Gaussian queries and organize them into local Gaussian groups. Each group shares an anchor that connects its current features with historical context for Gaussian prediction.

Adaptive Gaussian query generation. Instead of fixed or random sampling in image space, which can undersample fine-grained scene structures such as thin instruments and tissue boundaries, we adapt Gaussian query locations to the current image content to better preserve local detail. Specifically, for the current observation $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$, we partition the image into regular tiles $\{\Omega_b\}_{b=1}^{B}$ and initialize m uniformly spaced queries within each tile. These locations form the initial query set $\mathbf{U}^0 = \{\boldsymbol{u}_i^0\}_{i=1}^{N_q}$, where $N_q = B \cdot m$. Given the shared features $\mathbf{Z}_t$, we first use a DPT-style network (Ranftl et al., 2021) to obtain a dense pixel-aligned feature map $\mathbf{F}_t \in \mathbb{R}^{H\times W\times d}$. We then refine the initial queries using their local features: $\mathbf{U}_t = \{\boldsymbol{u}_{t,i}\}_{i=1}^{N_q} = \mathcal{Q}_\theta(\mathbf{U}^0, \mathbf{F}_t)$. Here $\mathcal{Q}_\theta$ predicts local adjustments to the initial query coordinates. To lift the refined Gaussian queries into 3D, we also predict a pixel-aligned pointmap $\mathbf{P}_t \in \mathbb{R}^{H\times W\times 3}$ in the shared world coordinate system using a DPT-style network within the Gaussian decoder. At each query location $\boldsymbol{u}_{t,i}$, we sample

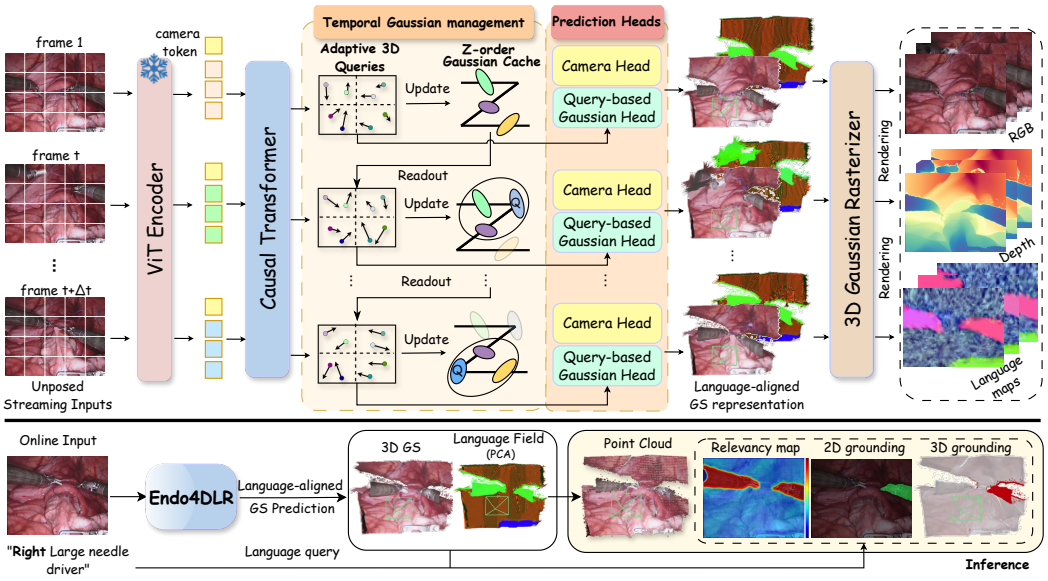


Figure 2: **Overview of *Endo4DLR*.** An image encoder and a causal Transformer extract shared spatiotemporal features from incoming frames for parallel camera and Gaussian prediction. Adaptive Gaussian queries are lifted into 3D and organized into local groups using Z-order serialization. The group anchors retrieve historical context and update the anchor memory as new frames arrive. The Gaussian decoder combines current geometric estimates with historical context to construct language-embedded Gaussians, supporting RGB-D rendering and natural-language grounding.

the corresponding 3D position and local feature: $\boldsymbol{x}_{t,i} = \mathbf{P}_t[\boldsymbol{u}_{t,i}]$, $\boldsymbol{f}_{t,i} = \mathbf{F}_t[\boldsymbol{u}_{t,i}]$, where $[\cdot]$ denotes bilinear sampling. Each Gaussian query thus carries a 3D position $\boldsymbol{x}_{t,i}$ and a feature $\boldsymbol{f}_{t,i}$ for subsequent spatial grouping and attribute decoding.

Z-order anchor construction. We then construct local anchors from the lifted Gaussian queries. Queries initialized within the same image tile may lie on an instrument and the tissue behind it, placing them at different depths in the reconstructed scene. We therefore group the queries according to their 3D positions so that each anchor aggregates spatially related queries and provides local context. A straightforward way is to construct neighborhoods through k-nearest-neighbor (KNN) search, but repeating this search for each incoming frame adds computational overhead. Following the serialized neighborhood mapping strategy (Wu et al., 2024b), we use serialization to approximate spatial neighborhoods in place of exact KNN search. Specifically, we employ Z-order serialization (Wang et al., 2026), which maps 3D positions to a one-dimensional sequence while preserving spatial locality. Sorting queries by this sequence allows us to form local groups from consecutive entries. We normalize and quantize each position to L bits per axis and compute its Morton code by interleaving the coordinate bits, $Z(\hat{x}, \hat{y}, \hat{z}) = \sum_{j=0}^{L-1} (\hat{x}_j 2^{3j} + \hat{y}_j 2^{3j+1} + \hat{z}_j 2^{3j+2})$, where $\hat{x}_j, \hat{y}_j, \hat{z}_j \in \{0, 1\}$ denote the bits of the quantized coordinates. We sort the queries by their Morton codes and partition the ordered sequence into A_t groups of K queries each, denoted by $\{\mathcal{N}_{t,a}\}_{a=1}^{A_t}$. For each group, we aggregate the query positions and features to obtain an anchor position $\boldsymbol{p}_{t,a} = \sum_{i \in \mathcal{N}_{t,a}} \omega_{t,i} \boldsymbol{x}_{t,i}$ and feature $\boldsymbol{a}_{t,a} = \sum_{i \in \mathcal{N}_{t,a}} \omega_{t,i} \boldsymbol{f}_{t,i}$, where $\boldsymbol{p}_{t,a} \in \mathbb{R}^3$, $\boldsymbol{a}_{t,a} \in \mathbb{R}^d$, and $\omega_{t,i}$ are geometric confidence scores sampled from the pointmap prediction at the query locations and normalized within each group. These anchors provide shared context for Gaussian decoding and temporal retrieval (Sec. 3.2). In practice, we retrieve context once per anchor and share it across the group, reducing query-level lookups and supporting temporally consistent language prediction.

Language-embedded Gaussian prediction. Features learned for geometric reconstruction emphasize spatial structure but may lack the fine appearance details needed for Gaussian rendering (Yu et al., 2026). In this regard, to retain the reconstructed geometry while learning attributes suited to rendering, we decouple Gaussian position from attribute prediction. Each query is aligned with the current pointmap, allowing us to directly use its position as the Gaussian center, $\boldsymbol{\mu}_{t,i} = \boldsymbol{x}_{t,i}$. We then

predict the associated attributes through a dedicated decoder conditioned on local anchor features and historical context. For anchor a, let $\widetilde{\boldsymbol{a}}_{t,a} \in \mathbb{R}^d$ denote its feature enriched with historical information through the temporal memory described in Sec. 3.2. We decode the attributes of its associated Gaussians: $\{(\boldsymbol{s}_{t,i}, \boldsymbol{r}_{t,i}, \alpha_{t,i}, \boldsymbol{c}_{t,i})\}_{i\in\mathcal{N}_{t,a}} = \mathcal{D}_g(\widetilde{\boldsymbol{a}}_{t,a})$, where $\mathcal{D}_g$ denotes the group attribute decoder. This formulation uses the current geometric estimates to locate Gaussians and the enriched anchor features to determine their rendering attributes. To make these Gaussians language-queryable, we additionally predict a language feature for each query. Since spatial grouping can bring together queries carrying different semantic cues, we combine each query's local feature with the shared anchor context in a separate semantic decoder. We first predict a pixel-aligned language map from $\mathbf{F}_t$ and sample an initial language feature $\boldsymbol{\ell}^0_{t,i} \in \mathbb{R}^{d_\ell}$ at each query location, $\boldsymbol{\ell}^0_{t,i} = \mathrm{MLP}(\mathbf{F}_t)[\boldsymbol{u}_{t,i}]$. To incorporate historical information while retaining this local prediction, we refine it with a contextual residual: $\boldsymbol{\ell}_{t,i} = \boldsymbol{\ell}^0_{t,i} + \mathcal{D}_\ell(\boldsymbol{f}_{t,i}, \widetilde{\boldsymbol{a}}_{t,a}), i \in \mathcal{N}_{t,a}$, where $\mathcal{D}_\ell$ is the semantic residual decoder. The resulting language features combine query-specific predictions from the current frame with historical context, encouraging temporal semantic consistency while preserving local semantic distinctions.

3.2 Persistent Gaussian Anchors for Temporal Context Retrieval

Historical scene information helps stabilize online reconstruction and language grounding by supplementing local cues missing from the current observation. However, the Gaussian groups in Sec. 3.1 are constructed per frame, so the same local region may be represented by different queries and anchors over time. The causal KV cache incorporates history into visual encoding (Lan et al., 2026), but leaves these associations implicit at the Gaussian level. We therefore maintain persistent Gaussian anchors that retain and update local scene features, explicitly linking current groups to relevant historical context for attribute and language prediction.

Specifically, we represent the persistent Gaussian anchors as $\mathcal{M}_{t-1} = \{(\boldsymbol{p}^m_j, \boldsymbol{a}^m_j, e^m_j)\}_{j=1}^{M_{t-1}}$, where M_{t-1} denotes the number of stored anchors. Each anchor retains a position $\boldsymbol{p}^m_j$, a local feature $\boldsymbol{a}^m_j$, and a spatial extent e^m_j. For a current anchor a, we define its spatial extent as $e_{t,a} = (\sum_{i\in\mathcal{N}_{t,a}} \omega_{t,i}\|\boldsymbol{x}_{t,i} - \boldsymbol{p}_{t,a}\|_2^2)^{1/2}$, where $\omega_{t,i}$ are the normalized confidence weights (Sec. 3.1). This extent provides a local spatial scale for aligning current anchors with historical anchors. To associate current groups with persistent anchors, we combine local feature correlations for cross-frame correspondence (Karaev et al., 2025) with 3D proximity to distinguish visually similar regions at different locations. For efficient retrieval, we reuse the Z-order indexing in Sec. 3.1. We quantize current and historical anchor positions under a common coordinate normalization and sort the historical anchors by their Morton codes. We collect $\mathcal{C}_{t,a}$ from neighboring entries around the current anchor's insertion position in this ordered sequence. For each candidate $j \in \mathcal{C}_{t,a}$, we compute:

$$s_{t,aj} = -\|\boldsymbol{p}_{t,a} - \boldsymbol{p}^m_j\|_2/(e_{t,a} + e^m_j) + \lambda_f(\boldsymbol{a}^\top_{t,a}\boldsymbol{a}^m_j)/(\|\boldsymbol{a}_{t,a}\|_2\|\boldsymbol{a}^m_j\|_2), \tag{2}$$

where larger normalized distances reduce the matching score, while greater feature similarity increases it, with λ_f weighting the feature term. The compatibility scores $s_{t,aj}$ guide the selection and weighting of historical features, allowing each current anchor to retrieve context from spatially and semantically related scene regions. We aggregate the selected features as $\boldsymbol{m}_{t,a} = \sum_{j\in\mathcal{C}_{t,a}} \beta_{t,aj}\boldsymbol{a}^m_j$, where $\beta_{t,aj} \geq 0$ and $\sum_{j\in\mathcal{C}_{t,a}} \beta_{t,aj} \leq 1$. These weights control the relative contributions of the selected anchors and the overall strength of historical context. A learned fusion network combines this context with the current anchor feature, $\widetilde{\boldsymbol{a}}_{t,a} = \phi_f(\boldsymbol{a}_{t,a}, \boldsymbol{m}_{t,a})$, to condition Gaussian attribute prediction and semantic residual decoding in Sec. 3.1. To encourage temporally consistent attribute and semantic predictions, the anchor feature consistency objective $\mathcal{L}_{\mathrm{temp}}$ in Sec. 3.3 aligns current anchor features with their matched historical features. To maintain persistent anchors as the scene evolves, we then update their states using the anchors constructed from the current frame, $\mathcal{M}_t = \mathcal{T}_{\mathrm{mem}}\Big(\mathcal{M}_{t-1}, \{(\boldsymbol{p}_{t,a}, \boldsymbol{a}_{t,a}, e_{t,a})\}_{a=1}^{A_t}\Big)$. Here $\mathcal{T}_{\mathrm{mem}}$ updates the positions, features, and extents of matched anchors, inserts unmatched current anchors, and removes obsolete anchors. These updates incorporate new observations while preserving local context for subsequent Gaussian groups, supporting temporally consistent attribute and language prediction.

3.3 Training and Online Inference

Training objectives. Following (Wang et al., 2024; Lan et al., 2026), we supervise the camera head and the pointmap prediction within the Gaussian decoder with $\mathcal{L}_{\mathrm{geo}}$. We combine rendering

supervision with direct supervision of per-query color and language predictions,

$$\mathcal{L}_{\text{rend}} = \lambda_{\text{rgb}}\mathcal{L}_{\text{photo}}(\widetilde{\mathbf{I}}_t, \mathbf{I}_t) + \lambda_{\text{lang}}\mathcal{L}_{\text{lang}}(\widetilde{\mathbf{L}}_t, \mathbf{L}_t^*; \mathbf{B}_t),$$
$$\mathcal{L}_{\text{query}} = \frac{1}{|\mathcal{V}_t|}\sum_{i\in\mathcal{V}_t}\left[\lambda_c\|\boldsymbol{c}_{t,i} - \boldsymbol{c}_{t,i}^*\|_1 + \lambda_\ell\mathcal{L}_{\text{lang}}(\boldsymbol{\ell}_{t,i}, \boldsymbol{\ell}_{t,i}^*; b_{t,i})\right]. \tag{3}$$

Here $\widetilde{\mathbf{I}}_t$ and $\widetilde{\mathbf{L}}_t$ are rendered RGB and language maps, and $\mathbf{L}_t^* \in \mathbb{R}^{H\times W\times d_\ell}$ is the reference language map. The binary instrument mask $\mathbf{B}_t \in \{0,1\}^{H\times W}$ distinguishes instrument and background regions, and $b_{t,i}$ denotes its label at query i. The set $\mathcal{V}_t$ indexes queries within the valid image region, with targets $\boldsymbol{c}_{t,i}^*$ and $\boldsymbol{\ell}_{t,i}^*$ sampled at $\boldsymbol{u}_{t,i}$ from $\mathbf{I}_t$ and $\mathbf{L}_t^*$, respectively. We combine L_1 and SSIM losses in $\mathcal{L}_{\text{photo}}$. Language supervision combines L_1 and cosine losses over valid locations. We further encourage temporal consistency through anchor feature alignment, $\mathcal{L}_{\text{temp}} = \sum_a \nu_{t,a}[1-\cos(\boldsymbol{a}_{t,a}, \overline{\boldsymbol{a}}_{t,a}^m)]/(\sum_a \nu_{t,a}+\epsilon)$, where $\overline{\boldsymbol{a}}_{t,a}^m$ is the feature of the highest-ranked matched historical anchor from Sec. 3.2, and $\nu_{t,a} \geq 0$ is the corresponding matching weight, set to zero for unmatched anchors, and $\epsilon > 0$ is a small constant. The main training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{query}} + \lambda_r\mathcal{L}_{\text{rend}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}. \tag{4}$$

Online inference. As shown in Fig. 2, *Endo4DLR* predicts cameras and language-embedded Gaussians from incoming frames, updating the causal state and anchor memory without per-scene optimization. These Gaussians support RGB-D rendering and language grounding. Given a unit text embedding $\boldsymbol{q} \in \mathbb{R}^{d_\ell}$ in the shared language space, $r_{t,i}(\boldsymbol{q}) = \cos(\boldsymbol{\ell}_{t,i}, \boldsymbol{q})$ measures Gaussian relevance in 3D. We rasterize unit-normalized Gaussian language features and divide by the accumulated opacity to obtain $\widetilde{\mathbf{L}}_t$. We then threshold the relevance map $S_t(\boldsymbol{u}; \boldsymbol{q}) = \widetilde{\mathbf{L}}_t[\boldsymbol{u}]^\top \boldsymbol{q}$ for 2D grounding.

4 Experiments

Datasets and Training. We use five endoscopic datasets, including EndoNeRF (Wang et al., 2022), EndoVis17 (Allan et al., 2019), EndoVis18 (Allan et al., 2020), StereoMIS (Hayoz et al., 2023), and StereoLung (Zhou et al., 2025b), comprising 65 distinct scenes and over 21,000 frames with camera motion and tissue deformation. We use an $8:2$ scene-level split for training and testing. Camera poses are evaluated on a StereoMIS test sequence segment and all test sequences from SCARED (Allan et al., 2021). Our model is trained for 30 epochs with 8 frames per iteration from a random scene. Training images are cropped to different resolutions to improve generality. More details about datasets and training are provided in the supplementary material.

Baselines. Since no existing method addresses fully feed-forward online language-aligned 4D reconstruction from unposed endoscopic videos, we compare against natural-scene methods with differing support for unposed inputs, online reconstruction, and language grounding. SLARM (Qiu et al., 2026) assumes known cameras, while DGD (Labe et al., 2024) and 4DLangSplat (Li et al., 2025a) require per-scene optimization. We therefore compare primarily with online reconstructors for unposed image streams, including StreamVGGT (Zhuo et al., 2026), InfiniteVGGT (Yuan et al., 2026), Spann3R (Wang & Agapito, 2025), CUT3R (Wang et al., 2025b), Point3R (Wu et al., 2026a), TTT3R (Chen et al., 2026), ZipMap (Jin et al., 2026), and Stream3R (Lan et al., 2026), while DGD and 4DLangSplat are included as additional language-grounding baselines. Since the online reconstructors do not natively predict language-embedded Gaussians, we adapt them using our basic decoding design, *i.e.*, we retain their camera predictions and obtain Gaussian centers from their point maps, with DPT-style attribute decoding. The baselines predict Gaussians densely per pixel, while *Endo4DLR* incorporates adaptive queries and anchor-based context retrieval. All methods use the same endoscopic data, rasterizer, and language embedding space for a fair comparison.

Metrics. Following (Lan et al., 2026; Katragadda et al., 2025), we evaluate RGB rendering with PSNR, SSIM, and LPIPS, depth prediction with AbsRel, RMSE, and $\delta < 1.25$ (δ_1), and 3D reconstruction with accuracy (Acc) and completion (Comp). Camera poses are assessed with ATE and translational/rotational RPE, and language grounding with IoU, Dice, precision, and recall.

4.1 Comparison with the State of the Art

Appearance and Depth Estimation. We evaluate RGB-D renderings on test scenes from all five datasets, using per-frame median scaling for depth evaluation following Lan et al. (2026). As

Table 1: **Online reconstruction and camera pose estimation results.** RMSE, ATE, RPE_t, Acc., and Comp. are in cm, RPE_r in degrees. Colored cells mark the first, second, and third ranks.

Method	Appearance			Depth			Camera Pose			3D Recon.	
	PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	RMSE ↓	δ_1(%) ↑	ATE ↓	RPE_t ↓	RPE_r ↓	Acc. ↓	Comp. ↓
S-VGGT (Zhuo et al., 2026)	9.366	0.295	0.817	0.322	4.148	64.003	1.401	0.584	0.219	1.107	0.892
Inf-VGGT (Yuan et al., 2026)	8.643	0.264	0.801	0.420	5.097	55.040	1.464	0.499	0.250	1.066	0.927
Spann3R (Wang & Agapito, 2025)	13.466	0.496	0.744	0.149	1.809	78.887	1.431	0.449	0.460	0.929	0.975
CUT3R (Wang et al., 2025b)	16.916	0.537	0.707	0.135	1.493	83.041	1.603	0.464	0.181	0.826	0.964
Point3R (Wu et al., 2026a)	8.171	0.108	0.928	0.852	8.224	10.362	1.707	0.465	0.178	5.061	2.445
TTT3R (Chen et al., 2026)	16.917	0.537	0.707	0.134	1.487	83.166	1.683	0.444	0.178	0.814	0.963
ZipMap (Jin et al., 2026)	13.515	0.331	0.779	0.066	1.275	83.601	1.382	0.380	0.242	0.931	0.967
Stream3R (Lan et al., 2026)	22.163	0.683	0.499	0.104	1.099	87.577	1.417	0.534	0.172	0.582	0.764
Endo4DLR (ours)	26.294	0.789	0.185	0.097	1.043	88.804	1.268	0.410	0.168	0.564	0.755

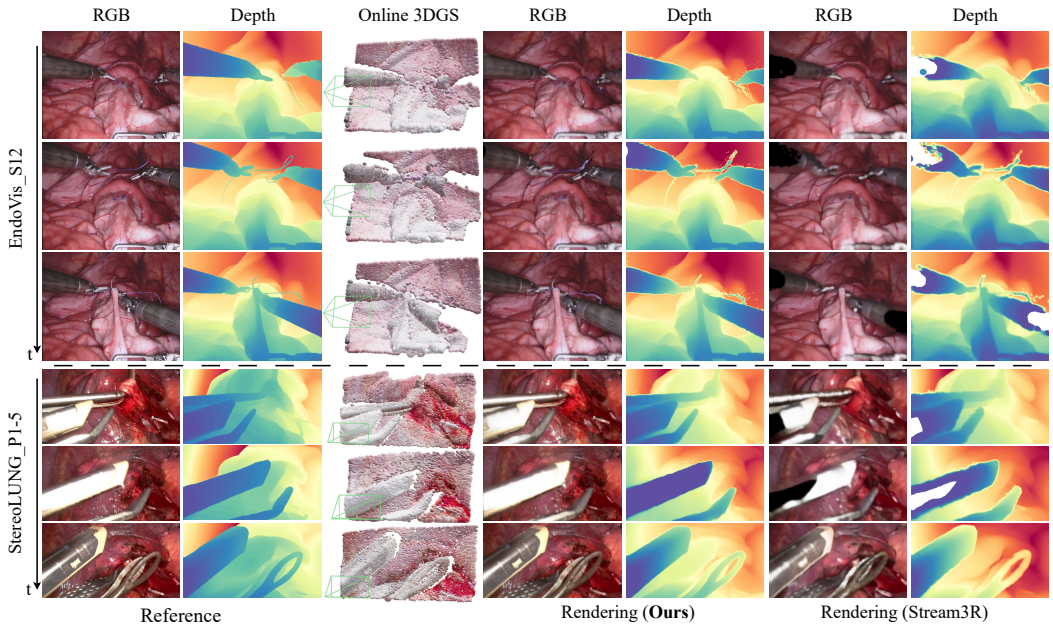


Figure 3: **Qualitative comparisons of online reconstruction.** Reference RGB-D images are shown alongside our Gaussian reconstructions and RGB-D renderings from *Endo4DLR* and Stream3R.

shown in Tab. 1, *Endo4DLR* outperforms the compared baselines across all appearance metrics. Compared to Stream3R, our model improves PSNR by 4.13 dB and SSIM by 0.106, while reducing LPIPS by 62.9%. Depth estimates also improve over Stream3R across all three metrics, with *Endo4DLR* achieving the lowest RMSE and highest δ_1 among the compared methods. In Fig. 3, our RGB renderings preserve clearer instrument contours and tissue details, and the rendered depth also better follows the reference geometry structure, whereas Stream3R exhibits noticeable missing regions in the EndoVis scene. These results demonstrate the effectiveness of our predicted Gaussians in preserving scene appearance and local geometry during online reconstruction.

Camera Pose Estimation. As in Lan et al. (2026), predicted trajectories are aligned to the ground truth using the Sim(3) transformation per sequence. Tab. 1 shows that *Endo4DLR* achieves the lowest ATE and rotational RPE, and the second-lowest translational RPE. Compared to Stream3R, it reduces ATE by 10.5% and translational RPE by 23.2%, substantiating the effectiveness of our model in joint camera estimation and dynamic scene reconstruction from unposed surgical streams.

3D Reconstruction. We evaluate surface geometry by back-projecting rendered and reference depth maps into point clouds in each frame's camera coordinate system. A single scale factor, computed as the median of the per-frame median-depth ratios, is applied to the entire predicted sequence. As shown in Tab. 1, *Endo4DLR* achieves the lowest Acc. and Comp. errors among the compared methods, improving both reconstruction accuracy and completeness. Together with the qualitative

Table 2: **Quantitative comparisons of language grounding.** Level-wise columns report IoU. Mean metrics are averaged over the four query levels. † denotes per-scene optimization methods.

Method	Tool	Tool+Act.	Tool+Pos.	Full	mIoU ↑	mDice ↑	mPrec. ↑	mRec. ↑
DGD† (Labe et al., 2024)	0.4816	0.5182	0.5481	0.6010	0.5372	0.6072	0.7063	0.6951
4DL-Splat† (Li et al., 2025a)	0.4091	0.3863	0.4884	0.4568	0.4351	0.5084	0.6541	0.6009
S-VGGT (Zhuo et al., 2026)	0.1842	0.1811	0.1734	0.1693	0.1770	0.2680	0.2085	0.5332
Inf-VGGT (Yuan et al., 2026)	0.1227	0.1293	0.1077	0.1094	0.1173	0.1848	0.1532	0.3307
Spann3R (Wang & Agapito, 2025)	0.1738	0.1699	0.1775	0.1664	0.1719	0.2661	0.1810	0.6618
CUT3R (Wang et al., 2025b)	0.1804	0.1761	0.1751	0.1723	0.1760	0.2689	0.1888	0.6697
Point3R (Wu et al., 2026a)	0.0271	0.0249	0.0254	0.0248	0.0256	0.0435	0.0852	0.1090
TTT3R (Chen et al., 2026)	0.1800	0.1753	0.1746	0.1729	0.1757	0.2687	0.1884	0.6701
ZipMap (Jin et al., 2026)	0.2815	0.2924	0.2655	0.2841	0.2809	0.3714	0.3021	0.5652
Stream3R (Lan et al., 2026)	0.4762	0.4876	0.4817	0.4908	0.4841	0.5890	0.5331	0.7677
Endo4DLR (ours)	0.5467	0.5509	0.5415	0.5379	0.5442	0.6393	0.6134	0.7527

reconstructions in Fig. 3, these results support the effectiveness of our Gaussian representation in recovering surgical scene geometry under a shared sequence-level scale.

Language Grounding. We evaluate open-vocabulary grounding on unseen test scenes using four query levels: tool type, tool with action, tool with position, and the full description combining all three. These queries assess whether the reconstructed language field can identify instruments and distinguish them by their actions and locations. As shown in Tab. 2, *Endo4DLR* outperforms the online baselines in IoU across all four levels. Compared to Stream3R, it improves mean IoU and mean Dice by 0.0601 and 0.0503, respectively. It also achieves higher mean IoU than DGD and 4DLangSplat without per-scene optimization, demonstrating effective language grounding in previously unseen surgical scenes. Fig. 4 presents results across successive frames containing multiple instruments. For descriptions specifying the left or right forceps, *Endo4DLR* localizes the queried instrument more accurately, whereas Stream3R also activates unrelated instruments. The predictions remain focused on the target across the illustrated frames as the scene evolves. The corresponding activated Gaussians successfully locate the queried instrument in 3D, demonstrating the effectiveness of our representation in associating language descriptions with surgical scene regions. Additional demonstrations are available on our anonymous project page.

4.2 Ablation Studies

Effect of adaptive queries. To assess whether adaptive query sampling improves local scene reconstruction, we replace adaptive queries with fixed queries uniformly distributed across image tiles. Tab. 3 shows that PSNR decreases from 26.29 to 25.64 dB and mIoU from 0.544 to 0.495. Fig. 5(a) also shows less distinct instrument boundaries in the RGB renderings. These results demonstrate that adaptive queries preserve local detail and improve both reconstruction and language grounding.

Effect of temporal context. To examine the contribution of explicit historical context beyond the causal Transformer, we disable Gaussian anchor memory while retaining the backbone's temporal features. Tab. 3 shows that PSNR decreases to 22.72 dB and mIoU to 0.503, accompanied by blurred instrument details in Fig. 5(b). In Fig. 4, this variant also activates unrelated instruments across successive frames. These findings support explicit context retrieval for improving reconstruction and maintaining accurate language grounding as the scene evolves.

Effect of Z-order organization. To evaluate Z-order organization for local Gaussian construction and historical association, we replace it with raster-order grouping and Euclidean nearest-neighbor retrieval. Adaptive queries and temporal memory remain enabled. As shown in Tab. 3, PSNR decreases to 24.89 dB, LPIPS increases to 0.257, and mIoU drops to 0.514. Combined with the RGB-D comparison in Fig. 5(c), these results support Z-order organization for constructing local Gaussian groups and retrieving their historical context.

Effect of geometry-attribute decoupling. To assess the benefit of separating geometric placement from attribute prediction, we jointly decode Gaussian positions and attributes from anchor features while retaining the separate semantic decoder. Tab. 3 shows that PSNR decreases from 26.29 to 20.06 dB and depth RMSE increases from 1.043 to 1.166 cm. Fig. 5(d) also shows no-

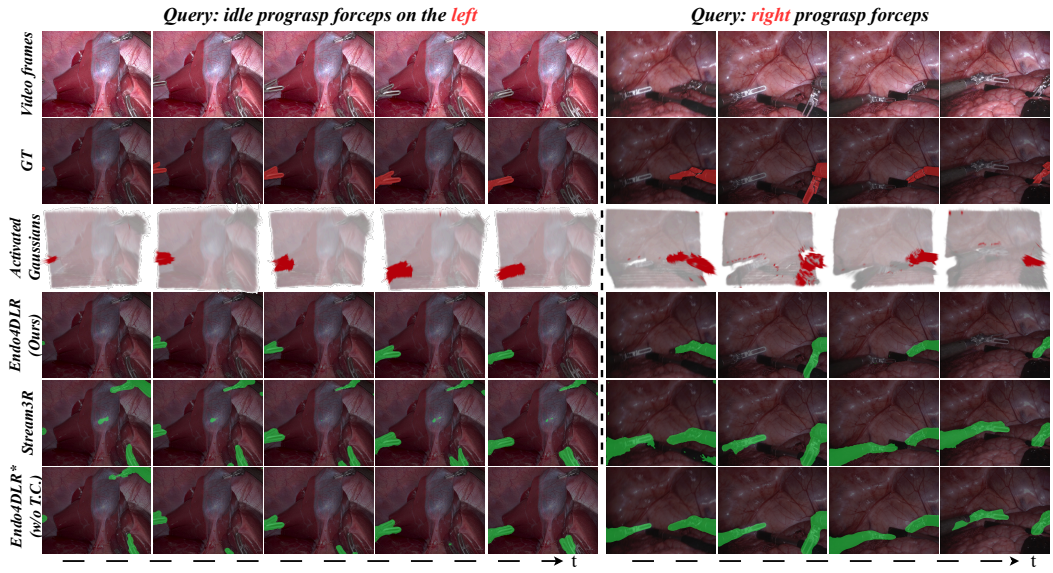
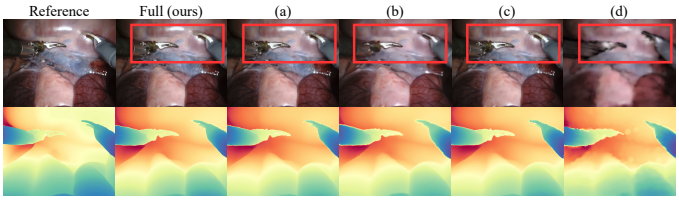


Figure 4: **Qualitative comparisons of language grounding** in image and 3D space. Ground-truth masks and activated Gaussians are red, and predicted masks are green. *Endo4DLR** is the ablation version that removes explicit temporal context retrieval, *i.e.*, variant (b) in Fig. 5 and Tab. 3.



Variant	PSNR ↑	LPIPS ↓	RMSE ↓	mIoU ↑
(a)	25.64	0.224	1.077	0.495
(b)	22.72	0.385	1.127	0.503
(c)	24.89	0.257	1.081	0.514
(d)	20.06	0.616	1.166	0.493
Full model	**26.29**	**0.185**	**1.043**	**0.544**

Figure 5: **Ablations.** Left: RGB (top) and depth (bottom) renderings. Table 3 (right): quantitative results, with RMSE in cm. Variants: (a) w/o adaptive query refinement, (b) w/o temporal context, (c) w/o Z-order organization, and (d) w/o geometry–attribute decoupling. Zoom in for details.

ticeable blurring and distorted depth around instruments. These results highlight the benefit of our geometry–attribute decoupling, which preserves current-frame geometric structure while using contextual features to improve appearance reconstruction.

Effect of a shared language space. Methods following the LangSplat framework (Qin et al., 2024; Li et al., 2025a) use scene-specific autoencoders to compress language features for efficient Gaussian rendering. However, independently learned spaces yield inconsistent supervision coordinates across scenes. We instead use a single autoencoder fitted on training scenes to map language features into a common latent space. Tab. 4 shows substantial improvements in both mIoU and mDice, supporting a shared supervision space for learning language grounding across scenes. Supplementary material provides further details.

Table 4: Effect of language feature compression.

Encoding	mIoU ↑	mDice ↑
Per-scene	0.162	0.193
Shared (ours)	**0.544**	**0.639**

5 Conclusion

We presented *Endo4DLR*, a fully feed-forward framework for online language-aligned 4D reconstruction from unposed endoscopic streams, jointly estimating cameras and predicting language-embedded Gaussians without per-scene optimization. Persistent Gaussian anchors connect adaptive local construction with historical context reuse, supporting temporally consistent appearance and language prediction. Experiments on five endoscopic datasets demonstrate improved reconstruction and language grounding, with ablations substantiating the effectiveness of the proposed design.

AI USE STATEMENT

In this work, we used generative AI tools to assist manuscript writing and editing, and to retrieve and review related literature. The authors reviewed and revised the AI-assisted manuscript content and take full responsibility for the final manuscript, including its text, methodology, and reported results.

ETHICS STATEMENT

This study adheres to the ICLR Code of Ethics and uses publicly available endoscopic datasets to investigate surgical scene reconstruction and language grounding. The work focuses on computational method development and evaluation using existing surgical recordings. Potential clinical applications would require further validation of accuracy, robustness, and safety.

REPRODUCIBILITY STATEMENT

The model architecture and training objectives are described in Sec. 3. Appendix B provides the model configuration, anchor retrieval and update algorithms, data preparation, training settings, and online inference procedure. The experimental setup and evaluation metrics are described in Sec. 4, with additional evaluation protocols and result aggregation details provided in Appendix C.1. We provide the core implementation and model configuration through an anonymous code repository linked from our anonymous project page (`https://endo4dlr.pages.dev/`) to support reproducibility. The page also includes video demonstrations and interactive visualizations of reconstruction and language grounding.

REFERENCES

Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19457–19467, 2024.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. In *International Conference on Learning Representations*, volume 2026, pp. 50694–50718, 2026.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pp. 370–386, 2024. doi: 10.1007/978-3-031-72664-4_21.

Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. PLA: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7010–7019, 2023.

Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, Boris Ivanovic, Marco Pavone, and Yue Wang. Large spatial model: End-to-end unposed images to semantic 3d. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Diandian Guo, Xikai Yang, Ruiyang Li, Jialun Pei, and Pheng-Ann Heng. Surglqa: Scalable long-horizon surgical video question answering. In *MICCAI*, pp. 1–10, 2026.

Michel Hayoz, Christopher Hahne, Mathias Gallardo, Daniel Candinas, Thomas Kurmann, Maximilian Allan, and Raphael Sznitman. Learning how to robustly estimate camera pose in endoscopic videos. *International journal of computer assisted radiology and surgery*, 18(7):1185–1192, 2023.

Haian Jin, Rundi Wu, Tianyuan Zhang, Ruiqi Gao, Jonathan T Barron, Noah Snavely, and Aleksander Hołyński. Zipmap: Linear-time stateful 3d reconstruction via test-time training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21748–21759, 2026.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labeling real videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Saimouli Katragadda, Cho-Ying Wu, Yuliang Guo, Xinyu Huang, Guoquan Huang, and Liu Ren. Online language splatting. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25882–25892. IEEE, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592433.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19729–19739, 2023.

Isaac Labe, Noam Issachar, Itai Lang, and Sagie Benaim. Dgd: Dynamic 3d gaussians distillation. In *European Conference on Computer Vision*, pp. 361–378. Springer, 2024.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Bo Dai, Shuai Yang, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. In *International Conference on Learning Representations*, volume 2026, pp. 42746–42768, 2026.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Jiahui Lei, Yijia Weng, Adam W. Harley, Leonidas Guibas, and Kostas Daniilidis. MoSca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6165–6177, 2025.

Wanhua Li, Renping Zhou, Jiawei Zhou, Yingwei Song, Johannes Herter, Minghan Qin, Gao Huang, and Hanspeter Pfister. 4d langsplat: 4d language gaussian splatting via multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22001–22011. IEEE, 2025a.

Yang Li, Jinglu Wang, Lei Chu, Xiao Li, Shiu-Hong Kao, Ying-Cong Chen, and Yan Lu. StreamGS: Online generalizable gaussian splatting reconstruction for unposed image streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25841–25850, 2025b.

Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21136–21145, 2024.

Yifan Liu, Chenxin Li, Chen Yang, and Yixuan Yuan. EndoGaussian: Real-time gaussian splatting for dynamic endoscopic scene reconstruction. *arXiv preprint arXiv:2401.12561*, 2024.

Yilong Liu, Wanhua Li, Chen Zhu-Tian, and Hanspeter Pfister. Langflash: Feed-forward 3d language gaussian splatting from sparse unposed images. *arXiv preprint arXiv:2605.23287*, 2026.

Yonghao Long, Anran Lin, Derek Hang Chun Kwok, Lin Zhang, Zhenya Yang, Kejian Shi, Lei Song, Jiawei Fu, Hongbin Lin, Wang Wei, et al. Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. *Science Robotics*, 10(104):eadt3093, 2025.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20654–20664, 2024.

Yuming Meng, Dong Wu, and Hongbin Zha. ST4R-Splat: Spatio-temporal referring segmentation in 4d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17598–17608, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 815–824, 2023.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10318–10327, 2021.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20051–20060, 2024.

Zhicheng Qiu, Jiarui Meng, Tong-an Luo, Yican Huang, Xuan Feng, Xuanfu Li, and Zhan Xu. SLARM: Streaming and language-aligned reconstruction model for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29023–29034, 2026.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 12159–12168. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01196. URL https://doi.org/10.1109/ICCV48922.2021.01196.

Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. 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.

Xiangyu Sun, Haoyi Jiang, Liu Liu, Seungtae Nam, Gyeongjin Kang, Xinjie Wang, Wei Sui, Zhizhong Su, Wenyu Liu, and Eunbyung Park. Uni3R: Unified 3d reconstruction and semantic understanding via generalizable gaussian splatting from unposed multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33280–33290, 2026.

Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3d instance segmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12959–12970, 2021.

Can Wang, Lei Liu, Wei Jiang, and Dong Xu. Z-order transformer for feed-forward gaussian splatting. *arXiv preprint arXiv:2605.13465*, 2026.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

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

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10510–10522, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *Medical Image Computing and Computer Assisted Intervention*, pp. 431–441, 2022. doi: 10.1007/978-3-031-16449-1_41.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5249–5260. IEEE, 2025.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20310–20320, 2024a.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4840–4851. IEEE, 2024b.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *Advances in Neural Information Processing Systems*, 38:69675–69699, 2026a.

Zike Wu, Qi Yan, Xuanyu Yi, Lele Wang, and Renjie Liao. StreamSplat: Towards online dynamic 3d reconstruction from uncalibrated video streams. In *International Conference on Learning Representations*, 2026b.

Weixing Xie, Junfeng Yao, Xianpeng Cao, Qiqin Lin, Zerui Tang, Xiao Dong, and Xiaohu Guo. SurgicalGaussian: Deformable 3d gaussians for high-fidelity surgical scene reconstruction. In *Medical Image Computing and Computer Assisted Intervention*, pp. 617–627, 2024.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20331–20341, 2024.

Botao Ye, Boqi Chen, Haofei Xu, Daniel Barath, and Marc Pollefeys. YoNoSplat: You only need one model for feedforward 3d gaussian splatting. *arXiv preprint arXiv:2511.07321*, 2025.

Haibao Yu, Kuntao Xiao, Jiahang Wang, Ruiyang Hao, Yuxin Huang, Guoran Hu, Haifang Qin, Bowen Jing, Yuntian Bo, and Ping Luo. Recondrive: Fast feed-forward 4d gaussian splatting for autonomous driving scene reconstruction. *arXiv preprint arXiv:2603.07552*, 2026.

Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. Infinitevggt: Visual geometry grounded transformer for endless streams. *arXiv preprint arXiv:2601.02281*, 2026.

Jun Zhou, Bingchen Gao, Kai Wang, Jialun Pei, Pheng-Ann Heng, and Jing Qin. Landmark-free preoperative-to-intraoperative registration in laparoscopic liver resection. *IEEE Transactions on Medical Imaging*, 44(11):4350–4362, 2025a.

Jun Zhou, Jialun Pei, Yan Li, Shi Qiu, Yanhui Wan, Weixin Si, Pheng-Ann Heng, and Jing Qin. Surgre4dgs: Open-vocabulary vision-language gaussian splatting for retrievable 4d surgical scene reconstruction. *OpenReview preprint*, 2025b. URL https://openreview.net/forum?id=xgdPfzy9Ww.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations*, volume 2026, pp. 88055–88072, 2026.

APPENDICES

The appendix is organized as follows:

- **Section A: Preliminaries.** Gaussian representation and rendering.
- **Section B: Implementation Details.** Model configuration, persistent anchor retrieval and update, training, and online inference.
- **Section C: Additional Experimental Results.** Per-dataset comparisons, additional ablations, and qualitative results.
- **Section D: Limitations, future work, and broader impact.** Current limitations, future directions, and broader impact.

A PRELIMINARIES

3D Gaussian representation. 3D Gaussian Splatting (Kerbl et al., 2023) represents a scene with anisotropic Gaussian primitives. For a fixed frame, we omit the time index and write the representation as $\mathcal{G} = \{g_i\}$. Each primitive has a center $\boldsymbol{\mu}_i$, scale $\boldsymbol{s}_i$, rotation quaternion $\boldsymbol{r}_i$, opacity α_i, and color $\boldsymbol{c}_i$. Its spatial influence is given by:

$$G_i(\boldsymbol{x}) = \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right], \qquad \boldsymbol{\Sigma}_i = \mathbf{R}_i \,\mathrm{diag}(\boldsymbol{s}_i)^2 \mathbf{R}_i^\top, \tag{5}$$

where $\mathbf{R}_i$ is the rotation matrix corresponding to $\boldsymbol{r}_i$.

Differentiable rendering. Given camera intrinsics $\mathbf{K}$ and world-to-camera extrinsics $(\mathbf{R}_c, \boldsymbol{t}_c)$ corresponding to $\boldsymbol{\pi}_t$, the Gaussian center in camera coordinates is $\boldsymbol{y}_i = \mathbf{R}_c\boldsymbol{\mu}_i + \boldsymbol{t}_c$. Perspective projection gives its image center $\overline{\boldsymbol{\mu}}_i = \mathrm{proj}_{\mathbf{K}}(\boldsymbol{y}_i)$. A first-order approximation of this projection yields the image-space covariance:

$$\overline{\boldsymbol{\Sigma}}_i = \mathbf{J}_i \mathbf{R}_c \boldsymbol{\Sigma}_i \mathbf{R}_c^\top \mathbf{J}_i^\top, \tag{6}$$

where $\mathbf{J}_i \in \mathbb{R}^{2\times 3}$ is the Jacobian of the perspective projection evaluated at $\boldsymbol{y}_i$. The projected opacity at pixel $\boldsymbol{u}$ is:

$$\widehat{\alpha}_i(\boldsymbol{u}) = \alpha_i \exp\left[-\frac{1}{2}(\boldsymbol{u}-\overline{\boldsymbol{\mu}}_i)^\top \overline{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{u}-\overline{\boldsymbol{\mu}}_i)\right]. \tag{7}$$

For each pixel, the contributing Gaussians are indexed in front-to-back order. Their compositing weights account for both opacity and occlusion:

$$w_i(\boldsymbol{u}) = \widehat{\alpha}_i(\boldsymbol{u}) \prod_{j<i} \big(1 - \widehat{\alpha}_j(\boldsymbol{u})\big). \tag{8}$$

RGB and depth are rendered as:

$$\widetilde{\mathbf{I}}(\boldsymbol{u}) = \sum_i w_i(\boldsymbol{u}) \boldsymbol{c}_i, \qquad \widetilde{\mathbf{D}}(\boldsymbol{u}) = \frac{\sum_i w_i(\boldsymbol{u}) z_i}{\sum_i w_i(\boldsymbol{u})}, \tag{9}$$

where z_i is the camera-space depth of $\boldsymbol{y}_i$. Depth is normalized by the accumulated opacity at valid pixels to obtain the expected depth of the visible surface.

Language-embedded rendering. Language features can be attached to Gaussian primitives and rendered through the same compositing process (Qin et al., 2024; Shi et al., 2024). For language grounding, we normalize each Gaussian feature as $\widehat{\boldsymbol{\ell}}_i = \boldsymbol{\ell}_i / \|\boldsymbol{\ell}_i\|_2$ and render the language map at valid pixels as:

$$\widetilde{\mathbf{L}}(\boldsymbol{u}) = \frac{\sum_i w_i(\boldsymbol{u}) \widehat{\boldsymbol{\ell}}_i}{\sum_i w_i(\boldsymbol{u})}. \tag{10}$$

Given a unit text embedding $\boldsymbol{q}$ in the same language space, Gaussian relevance is $r_i(\boldsymbol{q}) = \cos(\boldsymbol{\ell}_i, \boldsymbol{q})$. Image-space relevance is computed as $S(\boldsymbol{u}; \boldsymbol{q}) = \widetilde{\mathbf{L}}(\boldsymbol{u})^\top \boldsymbol{q}$, without further normalizing the rendered feature vectors. Thresholding this relevance map produces a grounding mask Li et al. (2025a).

B Implementation Details

B.1 Model Configuration

We initialize the reconstruction backbone from Stream3R (Lan et al., 2026), retaining its DINO image encoder and causal Transformer. The Gaussian decoder uses 64-dimensional pixel-aligned features and anchor features. Each group contains 14 lifted queries and produces 14 Gaussians, with each query providing one Gaussian center. Tab. 5 summarizes the sampling, representation, and temporal retrieval settings.

Table 5: **Model configuration** for Gaussian construction and temporal context retrieval.

Parameter	Value
Gaussian construction	
Image tile size	14×14 pixels
Initial queries per tile m	98
Query offset per image axis	$[-3, 3]$ pixels
Queries per group K	14
Gaussians per anchor	14
Pixel and anchor feature dimension d	64
Language feature dimension d_ℓ	6
Temporal context retrieval	
Morton quantization bits per axis L	10
Morton retrieval window	128 anchors
Maximum retrieved anchors per current anchor	4
Feature compatibility coefficient λ_f	0.5

B.2 Persistent Gaussian Anchor Retrieval and Update

To retrieve historical context for current Gaussian groups, we jointly normalize current and stored anchor positions before Morton encoding. Historical anchors are sorted once per frame, and candidates are collected around each current anchor's insertion position. A small number of Euclidean neighbors supplement the serialized neighborhood to capture nearby anchors separated in Morton order. The spatial and feature compatibility in Eq. 2 guides candidate selection and weighting, using the retrieval configuration in Tab. 5.

For each current anchor, the selected historical features are aggregated with weights $\beta_{t,aj}$ and fused with its current feature through ϕ_f. These weights reflect both relative candidate compatibility and the reliability of the association, reducing the historical contribution for uncertain matches. An unmatched anchor receives zero historical context. The fused feature conditions Gaussian decoding, while the current anchor feature supplies new information for updating the persistent state. Each accepted association links a current anchor to its highest-ranked historical anchor. When several current anchors share the same historical match, their estimates are pooled using geometric confidence and matching reliability. For each stored component $v_j^m \in \{\boldsymbol{p}_j^m, \boldsymbol{a}_j^m, e_j^m\}$, the pooled estimate $\overline{v}_j$ is incorporated through:

$$v_j^m \leftarrow \eta v_j^m + (1-\eta)\overline{v}_j, \tag{11}$$

where $\eta = 0.8$. Feature updates use accepted associations, whereas position and extent updates require reliable spatial agreement. Components without an accepted update remain unchanged. Unmatched current anchors initialize new entries, and memory retention prioritizes newly observed and recently updated anchors within the allocated capacity. This process incorporates current scene information while preserving local context for subsequent Gaussian prediction.

Algorithm 1 summarizes retrieval, fusion, and state update. Here, $\mathrm{Associate}$ returns the selected candidates, their aggregation weights, and the accepted historical match, or $\varnothing$ if no match is accepted. Pool computes the confidence-weighted estimates described above, and EMA applies Eq. 11 to eligible components. For an empty memory, candidate retrieval returns no entries, $\boldsymbol{m}_{t,a} = \mathbf{0}$, and all current anchors initialize the persistent state.

Algorithm 1 Persistent Gaussian anchor retrieval and update

Require: Current anchors $\mathcal{A}_t = \{\mathcal{A}_{t,a}\}_{a=1}^{A_t}$, where $\mathcal{A}_{t,a} = (\boldsymbol{p}_{t,a}, \boldsymbol{a}_{t,a}, e_{t,a})$, with geometric confidences; previous memory $\mathcal{M}_{t-1}$
Ensure: Fused features $\{\widetilde{\boldsymbol{a}}_{t,a}\}_{a=1}^{A_t}$ and updated memory $\mathcal{M}_t$

$\mathcal{M}' \leftarrow \mathcal{M}_{t-1}$ ▷ Initialize the state to update
$(Z_t, Z^m) \leftarrow \mathrm{Morton}(\mathrm{JointNormalize}(\{\boldsymbol{p}_{t,a}\}, \{\boldsymbol{p}_j^m\}))$
$\mathcal{I} \leftarrow \mathrm{argsort}(Z^m)$ ▷ Sort historical anchors
for $a = 1, \ldots, A_t$ **do**
  $\mathcal{C}_{t,a} \leftarrow \mathrm{Candidates}(Z_{t,a}, \boldsymbol{p}_{t,a}, Z^m, \mathcal{I}, \mathcal{M}_{t-1})$ ▷ Serialized and spatial neighbors
  $(\mathcal{K}_{t,a}, \{\beta_{t,aj}\}, j_a^\star) \leftarrow \mathrm{Associate}(\mathcal{A}_{t,a}, \mathcal{C}_{t,a})$ ▷ Spatial and feature compatibility
  $\boldsymbol{m}_{t,a} \leftarrow \sum_{j \in \mathcal{K}_{t,a}} \beta_{t,aj} \boldsymbol{a}_j^m$ ▷ Aggregate historical context
  $\widetilde{\boldsymbol{a}}_{t,a} \leftarrow \phi_f(\boldsymbol{a}_{t,a}, \boldsymbol{m}_{t,a})$ ▷ Fuse current and historical features
end for
for each historical anchor j **do**
  $\mathcal{J}_j \leftarrow \{a : j_a^\star = j\}$ ▷ Current anchors associated with j
  if $\mathcal{J}_j \neq \varnothing$ **then**
    $(\overline{\boldsymbol{p}}_j, \overline{\boldsymbol{a}}_j, \overline{e}_j) \leftarrow \mathrm{Pool}(\{\mathcal{A}_{t,a}\}_{a \in \mathcal{J}_j})$ ▷ Confidence-weighted current estimates
    $(\boldsymbol{p}_j^m, \boldsymbol{a}_j^m, e_j^m) \leftarrow \mathrm{EMA}\big((\boldsymbol{p}_j^m, \boldsymbol{a}_j^m, e_j^m), (\overline{\boldsymbol{p}}_j, \overline{\boldsymbol{a}}_j, \overline{e}_j)\big)$ ▷ Update eligible components in $\mathcal{M}'$
  end if
end for
$\mathcal{B}_t \leftarrow \{\mathcal{A}_{t,a} : j_a^\star = \varnothing\}$ ▷ Initialize unmatched anchors
$\mathcal{M}_t \leftarrow \mathrm{Retain}(\mathcal{M}' \cup \mathcal{B}_t)$ ▷ Prioritize new and recently updated anchors
return $\{\widetilde{\boldsymbol{a}}_{t,a}\}_{a=1}^{A_t}, \mathcal{M}_t$

B.3 Training Details

Datasets and training details. Tab. 6 details the scene-level split used in our experiments. We use the existing instance masks and enriched language descriptions (Zhou et al., 2025b) to construct pixel-aligned language supervision, with the same annotations used for all compared methods. Reference depth maps are estimated from rectified stereo pairs using FoundationStereo (Wen et al., 2025). RGB images, depth maps, and language maps share the same cropping and spatial augmentation strategies, preserving their pixel correspondence following Lan et al. (2026). Each training clip contains eight frames sampled from a randomly selected training scene and arranged in temporal order. The frames are cropped to either 518×406 or 518×280, chosen randomly per clip.

We train *Endo4DLR* for 30 epochs on four NVIDIA RTX A6000 GPUs using AdamW with a learning rate of 5×10^{-5}, weight decay of 0.05, and $(\beta_1, \beta_2) = (0.9, 0.95)$. A two-epoch linear warmup is followed by cosine decay to 10^{-6}. Each GPU processes one eight-frame clip per iteration, with four gradient accumulation steps, yielding an effective batch size of 16 clips. Training uses mixed precision, with the pretrained image encoder frozen. For the objectives in Eqs. 3 and 4, we set $\lambda_{\mathrm{rgb}} = \lambda_{\mathrm{lang}} = \lambda_\ell = 1$, $\lambda_c = 0.5$, and $\lambda_{\mathrm{temp}} = 5 \times 10^{-3}$. The camera loss within $\mathcal{L}_{\mathrm{geo}}$ has a weight of 2. After an initial warmup of 2,000 steps, the rendering weight λ_r gradually increases to 0.5 over the next 3,000 steps. Following Wang et al. (2025b), camera supervision is applied only to sequences with reference trajectories during training.

Language feature compression. Following LangSplat (Qin et al., 2024), we compress language features to reduce the storage and rendering costs of language-embedded Gaussians. We use a single autoencoder trained across training scenes to establish a shared latent space. Instance descriptions are encoded into 768-dimensional BioBERT-based sentence embeddings (Lee et al., 2020) and compressed into 6-dimensional features. The autoencoder uses an encoder with dimensions $768 \to 512 \to 256 \to 64 \to 6$ and a decoder with the reverse dimensions. Its latent outputs are L_2-normalized. We train the autoencoder on descriptions from training scenes for 300 epochs using AdamW, with a learning rate of 10^{-3}, weight decay of 10^{-4}, and batch size of 4,096. The objective combines mean squared reconstruction error with a cosine loss weighted by 0.01. The trained encoder is frozen and shared across methods to encode both reference descriptions and text queries.

Table 6: **Dataset statistics** for model training and evaluation.

Category	Split	EndoNeRF (Wang et al., 2022)	StereoMIS (Hayoz et al., 2023)	StereoLung (Zhou et al., 2025b)	EndoVis17 (Allan et al., 2019)	EndoVis18 (Allan et al., 2020)	**Total**
Scenes	Train	4	3	21	13	11	52
	Test (unseen)	1	1	5	3	3	13
Frames	Train	426	744	13,500	2,025	1,639	18,334
	Test (unseen)	121	425	2,287	375	447	3,655
	Total	547	1,169	15,787	2,400	2,086	**21,989**

B.4 Online Inference

At inference, *Endo4DLR* processes unposed frames sequentially, carrying the causal state $\mathbf{h}_t$ and persistent anchor memory $\mathcal{M}_t$ across frames. Both states are initialized at the beginning of each sequence. Input frames undergo the same cropping as during training, followed by resizing to a 518-pixel long edge and cropping to dimensions divisible by the encoder's patch size. For each incoming frame, the model predicts camera parameters and language-embedded Gaussians through a feed-forward pass, as summarized in Algorithm 2. Rendering uses the current frame's Gaussians, whose attributes incorporate the retrieved historical context. We report the forward inference speed on a single NVIDIA RTX A6000, excluding rendering time.

For language grounding, text queries are encoded into unit embeddings $\boldsymbol{q}$ using the BioBERT-based sentence encoder and frozen shared autoencoder described in Sec. B.3. Gaussian relevance in 3D is measured by $r_{t,i}(\boldsymbol{q}) = \cos(\boldsymbol{\ell}_{t,i}, \boldsymbol{q})$. For image-space grounding, we render normalized Gaussian language features following Eq. 10 and compute $S_t(\boldsymbol{u}; \boldsymbol{q}) = \widetilde{\mathbf{L}}_t[\boldsymbol{u}]^\top \boldsymbol{q}$, without further normalizing the rendered features. We use a shared relevance threshold $\tau = 0.8$ across the compared methods.

Algorithm 2 Online language-aligned 4D reconstruction

Require: Unposed image stream $\{\mathbf{I}_t\}_{t=1}^{T}$
Ensure: Camera parameters $\boldsymbol{\pi}_t$ and language-embedded Gaussians $\mathcal{G}_t$ for each frame

$\mathbf{h}_0 \leftarrow \mathrm{InitState}(), \quad \mathcal{M}_0 \leftarrow \varnothing$
for $t = 1, \ldots, T$ **do**
  $(\mathbf{Z}_t, \mathbf{h}_t) \leftarrow \mathrm{Encode}(\mathbf{I}_t, \mathbf{h}_{t-1})$ ▷ Causal visual encoding
  $\boldsymbol{\pi}_t \leftarrow \mathrm{CameraHead}(\mathbf{Z}_t)$
  $(\mathbf{P}_t, \mathbf{F}_t) \leftarrow \mathrm{DenseDecode}(\mathbf{Z}_t)$ ▷ Within the Gaussian decoder
  $\mathbf{U}^0 \leftarrow \mathrm{TileQueries}(\mathbf{I}_t)$
  $\mathbf{U}_t \leftarrow \mathcal{Q}_\theta(\mathbf{U}^0, \mathbf{F}_t)$ ▷ Adaptive query refinement
  $\boldsymbol{x}_{t,i} \leftarrow \mathbf{P}_t[\boldsymbol{u}_{t,i}], \quad \boldsymbol{f}_{t,i} \leftarrow \mathbf{F}_t[\boldsymbol{u}_{t,i}] \quad \forall i$ ▷ Lift queries into 3D
  $\boldsymbol{\ell}^0_{t,i} \leftarrow \mathrm{MLP}(\mathbf{F}_t)[\boldsymbol{u}_{t,i}] \quad \forall i$ ▷ Sample local language features
  $(\{\mathcal{N}_{t,a}\}, \mathcal{A}_t) \leftarrow \mathrm{AnchorConstruction}(\mathbf{P}_t, \mathbf{F}_t, \mathbf{U}_t)$ ▷ Sec. 3.1
  $(\{\widetilde{\boldsymbol{a}}_{t,a}\}, \mathcal{M}_t) \leftarrow \mathrm{RetrieveUpdate}(\mathcal{A}_t, \mathcal{M}_{t-1})$ ▷ Algorithm 1
  for $a = 1, \ldots, A_t$ **do**
    $\{(\boldsymbol{s}_{t,i}, \boldsymbol{r}_{t,i}, \alpha_{t,i}, \boldsymbol{c}_{t,i})\}_{i \in \mathcal{N}_{t,a}} \leftarrow \mathcal{D}_g(\widetilde{\boldsymbol{a}}_{t,a})$ ▷ Group attribute decoding
    $\boldsymbol{\mu}_{t,i} \leftarrow \boldsymbol{x}_{t,i} \quad (i \in \mathcal{N}_{t,a})$ ▷ Current geometric estimates
    $\boldsymbol{\ell}_{t,i} \leftarrow \boldsymbol{\ell}^0_{t,i} + \mathcal{D}_\ell(\boldsymbol{f}_{t,i}, \widetilde{\boldsymbol{a}}_{t,a}) \quad (i \in \mathcal{N}_{t,a})$ ▷ Contextual language refinement
  end for
  $\mathcal{G}_t \leftarrow \{(\boldsymbol{\mu}_{t,i}, \boldsymbol{s}_{t,i}, \boldsymbol{r}_{t,i}, \alpha_{t,i}, \boldsymbol{c}_{t,i}, \boldsymbol{\ell}_{t,i})\}_{i=1}^{N_q}$
  output $(\boldsymbol{\pi}_t, \mathcal{G}_t)$
end for

C Additional Experimental Results

C.1 Per-Dataset Results

We further examine reconstruction and language grounding across the five endoscopic datasets through per-dataset comparisons with Stream3R (Lan et al., 2026). Separate camera-pose results on StereoMIS and SCARED complement this analysis with an evaluation of cross-dataset pose

generalization. For the overall results in Tabs. 1 and 2, appearance metrics are averaged over test frames, depth metrics use valid-pixel weighting, and 3D reconstruction metrics are averaged over test sequences. Language-grounding metrics are averaged over query entries within each query level, followed by averaging across the four levels for the mean metrics. The per-dataset results apply the same aggregation within each dataset, so their contributions to the overall results depend on the corresponding evaluation counts. Camera-pose results are averaged equally across StereoMIS and SCARED.

Online reconstruction. Tab. 7 shows that *Endo4DLR* improves PSNR, SSIM, and LPIPS over Stream3R on every dataset, demonstrating effective appearance reconstruction across different surgical scenes. The improvements extend to all reported reconstruction metrics on EndoVis17 and StereoLung, and all three depth metrics on EndoVis18. These results support the ability of the predicted Gaussian representation to preserve both appearance and local geometry across unseen surgical scenes.

Table 7: **Per-dataset reconstruction comparison with Stream3R (Lan et al., 2026).** RMSE, Acc., and Comp. are in cm, and δ_1 is in percent. Colored cells indicate the better result within each dataset.

Dataset	Method	Appearance			Depth			3D Recon.	
		PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	RMSE ↓	δ_1 ↑	Acc. ↓	Comp. ↓
EndoNeRF	Stream3R	21.388	0.770	0.420	0.050	2.881	99.650	2.518	2.436
	Endo4DLR	29.679	0.853	0.147	0.049	2.899	99.257	2.484	2.477
EndoVis17	Stream3R	23.115	0.763	0.396	0.095	1.382	92.210	0.490	0.673
	Endo4DLR	26.177	0.826	0.153	0.071	1.106	94.842	0.458	0.641
EndoVis18	Stream3R	20.392	0.751	0.443	0.142	0.938	82.660	0.484	0.714
	Endo4DLR	24.428	0.817	0.202	0.125	0.883	84.061	0.493	0.703
StereoMIS	Stream3R	21.382	0.537	0.557	0.224	2.054	56.530	0.873	1.575
	Endo4DLR	25.511	0.685	0.249	0.229	2.111	56.105	0.775	1.603
StereoLung	Stream3R	22.540	0.679	0.520	0.068	0.678	96.580	0.250	0.351
	Endo4DLR	26.645	0.793	0.178	0.064	0.643	96.721	0.243	0.342

Camera pose estimation. We evaluate camera trajectories on a held-out StereoMIS sequence segment and additionally on SCARED (Allan et al., 2021) without fine-tuning. SCARED provides a complementary test in static endoscopic scenes, assessing pose generalization beyond the training datasets. Predicted trajectories undergo per-sequence $\mathrm{Sim}(3)$ alignment before computing ATE and RPE, following Lan et al. (2026). As shown in Tab. 8, *Endo4DLR* reduces ATE and translational RPE relative to Stream3R on both datasets. On SCARED, ATE decreases from 1.502 to 1.229 cm, while rotational RPE remains comparable. These results demonstrate that the camera estimation learned in our framework transfers to an unseen endoscopic dataset without additional adaptation.

Table 8: **Camera pose estimation on StereoMIS (Hayoz et al., 2023) and SCARED (Allan et al., 2021).** SCARED is evaluated without fine-tuning. ATE and $\mathrm{RPE_t}$ are in cm, and $\mathrm{RPE_r}$ is in degrees.

Dataset	Method	ATE ↓	$\mathrm{RPE_t}$ ↓	$\mathrm{RPE_r}$ ↓
StereoMIS	Stream3R	1.332	0.496	0.114
	Endo4DLR	1.306	0.419	0.105
SCARED	Stream3R	1.502	0.571	0.229
	Endo4DLR	1.229	0.401	0.230

Language grounding. Tab. 9 shows that *Endo4DLR* achieves higher IoU than Stream3R at every query level on each dataset, together with improvements in mean Dice and precision. These gains extend from tool-category queries to descriptions incorporating actions and spatial relations, demonstrating effective language grounding across unseen surgical scenes.

Table 9: **Per-dataset language-grounding comparison with Stream3R (Lan et al., 2026).** Level-wise columns report IoU. Mean metrics are averaged over the four query levels.

Dataset	Method	Tool	Tool+Act.	Tool+Pos.	Full	mIoU ↑	mDice ↑	mPrec. ↑	mRec. ↑
EndoNeRF	Stream3R	0.6164	0.5891	0.5913	0.5866	0.5959	0.6815	0.6358	0.7777
	Endo4DLR	0.7211	0.7036	0.7315	0.7106	0.7167	0.8111	0.8403	0.8554
EndoVis17	Stream3R	0.5795	0.5588	0.4995	0.5008	0.5347	0.6351	0.5612	0.8102
	Endo4DLR	0.6277	0.5967	0.5396	0.5392	0.5758	0.6747	0.6043	0.8502
EndoVis18	Stream3R	0.5086	0.5176	0.4986	0.5005	0.5063	0.6007	0.5760	0.7157
	Endo4DLR	0.5713	0.5796	0.5406	0.5462	0.5594	0.6358	0.6307	0.6997
StereoMIS	Stream3R	0.2917	0.2269	0.3330	0.2497	0.2753	0.3990	0.2935	0.7073
	Endo4DLR	0.2949	0.2586	0.3708	0.2673	0.2979	0.4090	0.3465	0.5782
StereoLung	Stream3R	0.4745	0.5211	0.5013	0.5411	0.5095	0.6157	0.5671	0.7877
	Endo4DLR	0.5667	0.5949	0.5705	0.5919	0.5810	0.6765	0.6607	0.7799

C.2 Additional Ablations and Analysis

Effect of query group size. The query group size K determines how many queries share an anchor's context, controlling the granularity of local Gaussian groups and the number of anchors used for temporal retrieval. To assess its effect on prediction quality and efficiency, we compare $K \in \{7, 14, 28, 49\}$ at $m = 98$, keeping the total query and Gaussian counts fixed. Fig. 6 shows that smaller groups generally improve reconstruction and grounding quality, while Fig. 7 shows that larger groups increase inference speed. Increasing K from 7 to 14 raises speed from approximately 5.4 to 6.1 FPS, with PSNR and mIoU decreasing by 0.44 dB and 0.0047, respectively. Further increasing K to 28 provides only a small speed gain, while PSNR drops to 25.14 dB and mIoU to 0.523. At $K = 49$, inference reaches approximately 6.8 FPS, but quality declines further across the plotted metrics. We therefore adopt $K = 14$ to balance reconstruction and grounding quality with inference efficiency.

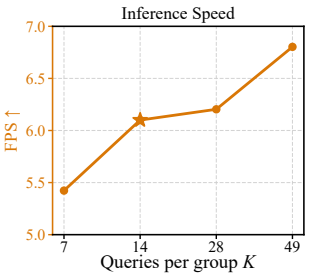


Figure 7: **Inference speed.**

Analysis of the shared language space. To examine the semantic organization of compressed language features across scenes, we visualize the same descriptions under per-scene and shared encoding. The analysis covers 543 descriptions from 62 scenes across five datasets, including six tool categories occurring in multiple datasets. Fig. 8 presents t-SNE projections of their 6D features. Under per-scene encoding, descriptions of the same tool are dispersed and intermingled with other categories. Shared encoding exhibits a clearer semantic structure, while retaining distinctions within each tool category. Inspection of representative aspirator subclusters reveals descriptions of different actions, such as being idle and retracting tissue. This structure is consistent with language grounding that distinguishes both instruments and their described activities. Together with the grounding results in Tab. 4, this qualitative analysis supports using a common encoding function to provide consistent language supervision across surgical scenes.

Effect of Gaussian density. The number of queries per tile controls how densely the current observation is sampled for Gaussian construction. To examine its effect on prediction quality and efficiency, we vary $m \in \{56, 98, 196\}$ while fixing the tile size at 14×14 and query group size at $K = 14$. These settings correspond to approximately 29%, 50%, and 100% of the Gaussian count of dense pixel-aligned prediction. As shown in Fig. 9, increasing density improves PSNR but does not consistently improve depth estimation or language grounding. Compared with $m = 196$, the default $m = 98$ halves the Gaussian count and increases inference speed from approximately 5.4 to 6.1 FPS, with PSNR and mIoU decreasing by only 0.09 dB and 0.0032, respectively. AbsRel also improves from 0.1009 to 0.0972. Further reducing m to 56 achieves lower depth error and higher mIoU, but reduces PSNR to 25.54 dB with only a small additional speed gain. These results support constructing an effective language-embedded representation from sparsely sampled queries organized through local anchors. We adopt $m = 98$ to balance reconstruction quality, language grounding, and inference efficiency.

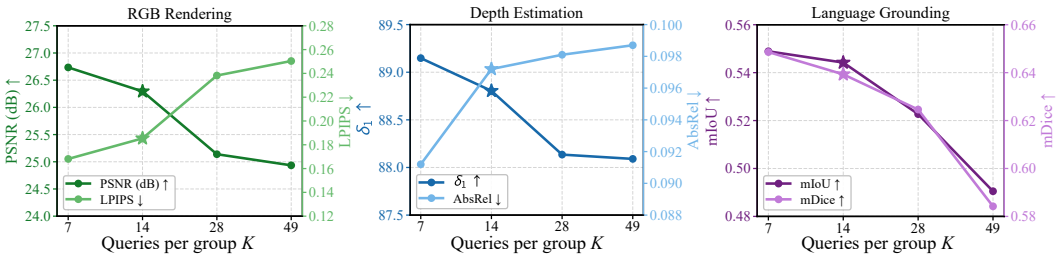


Figure 6: **Effect of query group size on prediction quality.** RGB rendering, depth estimation, and language grounding under different query group sizes. Stars indicate the default setting ($K = 14$).

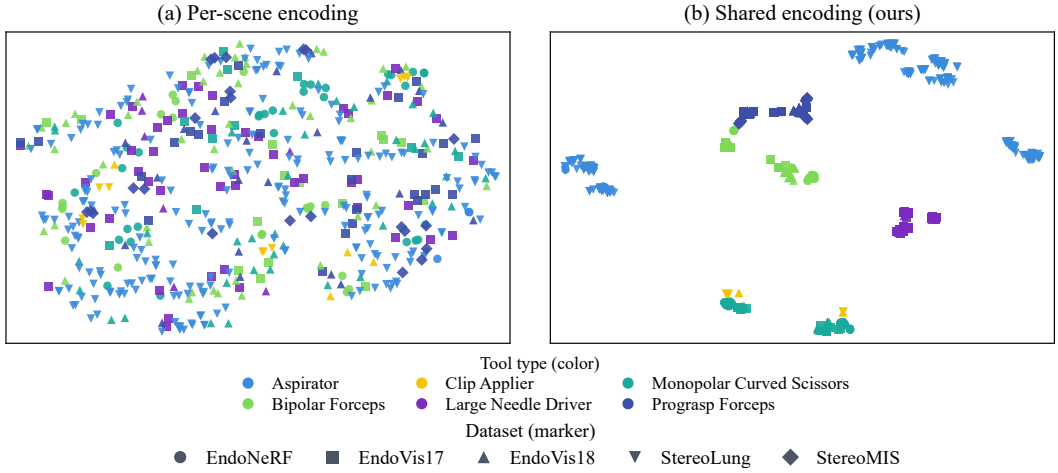


Figure 8: **Visualization of compressed language features.** The same descriptions are encoded using (a) per-scene autoencoders and (b) the shared autoencoder, and separately projected with t-SNE. Colors denote tool categories and marker shapes denote datasets.

C.3 Additional Qualitative Results

Online reconstruction. Fig. 10 presents additional qualitative comparisons with Stream3R on four unseen surgical scenes, each shown at three time steps. We compare rendered RGB-D images and the reconstructed point clouds with the references. Across these scenes, *Endo4DLR* preserves sharper instrument boundaries and finer tissue textures, while Stream3R tends to blur local appearance details. The depth renderings also recover more complete instrument shapes in regions where Stream3R produces missing predictions. The corresponding point clouds reveal these differences in the reconstructed local geometry. Across these time steps, *Endo4DLR* captures changes in instrument configuration and tissue structure, supporting the effectiveness of its Gaussian representation for fine-grained online reconstruction of dynamic surgical scenes.

Language grounding. We provide additional comparisons of language grounding across unseen surgical scenes and diverse language queries. Grounding results are obtained from rendered language features in image space, allowing direct comparison with reference segmentation masks. Fig. 11 presents frame-wise IoU on unseen surgical sequences for queries describing instrument identities, actions, and spatial locations. Compared with Stream3R (Lan et al., 2026), ZipMap (Jin et al., 2026), and CUT3R (Wang et al., 2025b), *Endo4DLR* achieves higher IoU across most annotated intervals. Its performance across these sequences supports robust language grounding as instrument configurations and scene visibility change. These results also demonstrate its ability to localize targets under varied language descriptions in unseen scenes.

Fig. 12 further illustrates the accuracy and selectivity of our grounding predictions. Across the displayed examples, *Endo4DLR* more accurately covers the queried instruments and preserves their contours, while reducing false activations on surrounding tissue and unrelated instruments. The

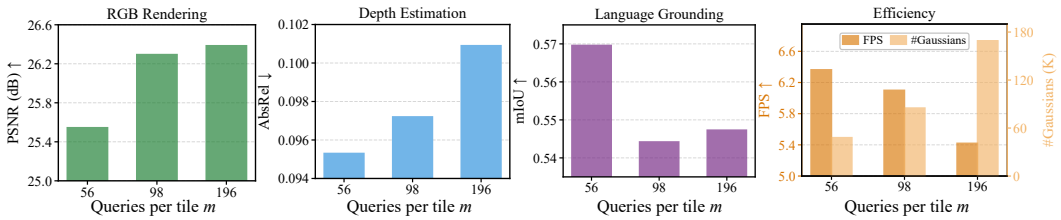
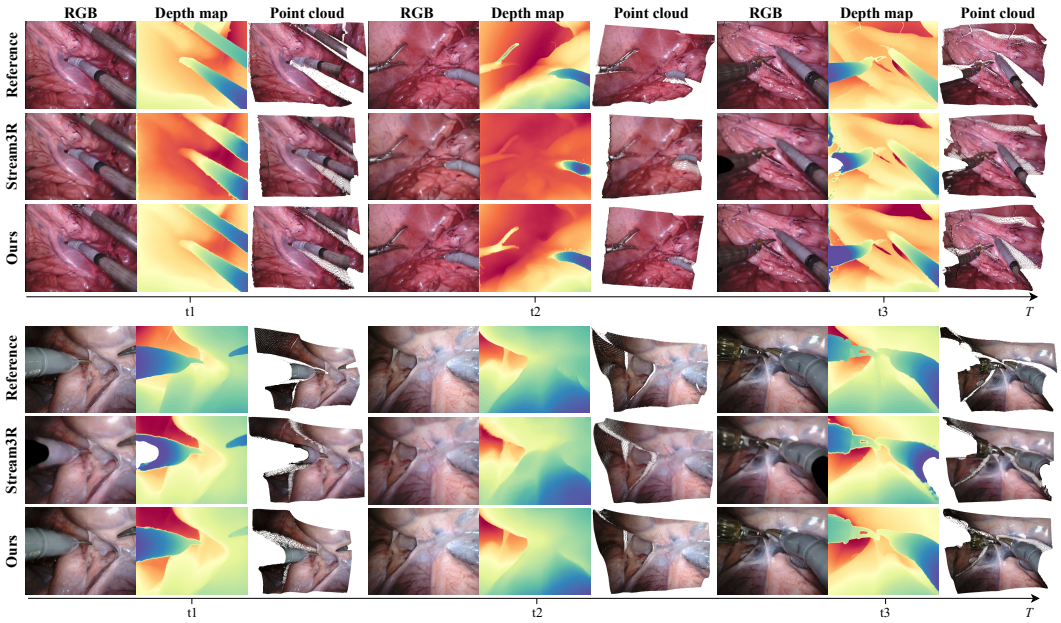
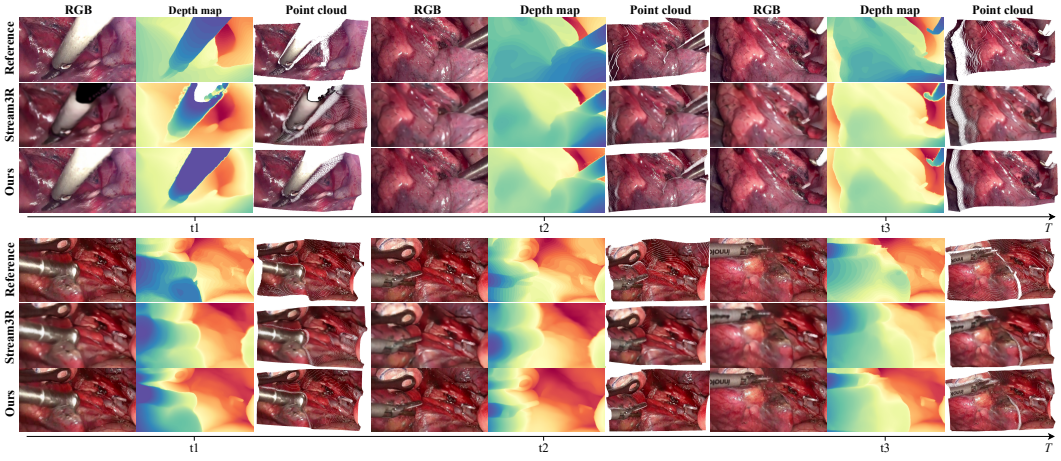


Figure 9: **Effect of Gaussian density.** Prediction quality and efficiency as the number of queries per tile m varies. The default setting is $m = 98$. Gaussian counts are averaged over evaluated frames and reported in thousands.

relevance maps concentrate responses on the described targets, supporting their separation from nearby structures. For the query describing a stapler retracting tissue on the left, *Endo4DLR* localizes the target more selectively, whereas the baselines also activate adjacent regions. The bipolar forceps and ultrasonic scalpel examples similarly show more complete target coverage across time. Together, these comparisons demonstrate the effectiveness of our language-embedded Gaussian representation in associating language descriptions with the corresponding scene regions during online reconstruction. Please refer to our anonymous project page for more video demonstrations.



(a) Two unseen scenes from EndoVis 17 (Allan et al., 2019) and EndoVis 18 (Allan et al., 2020).



(b) Two unseen scenes from StereoLUNG (Zhou et al., 2025b).

Figure 10: **Additional online reconstruction results.** Four test scenes are shown at three time steps. Each scene compares reference observations, Stream3R, and *Endo4DLR* in RGB, depth, and back-projected point clouds. Predicted and reference point clouds are obtained from rendered and reference depth, respectively. Zoom in for details.

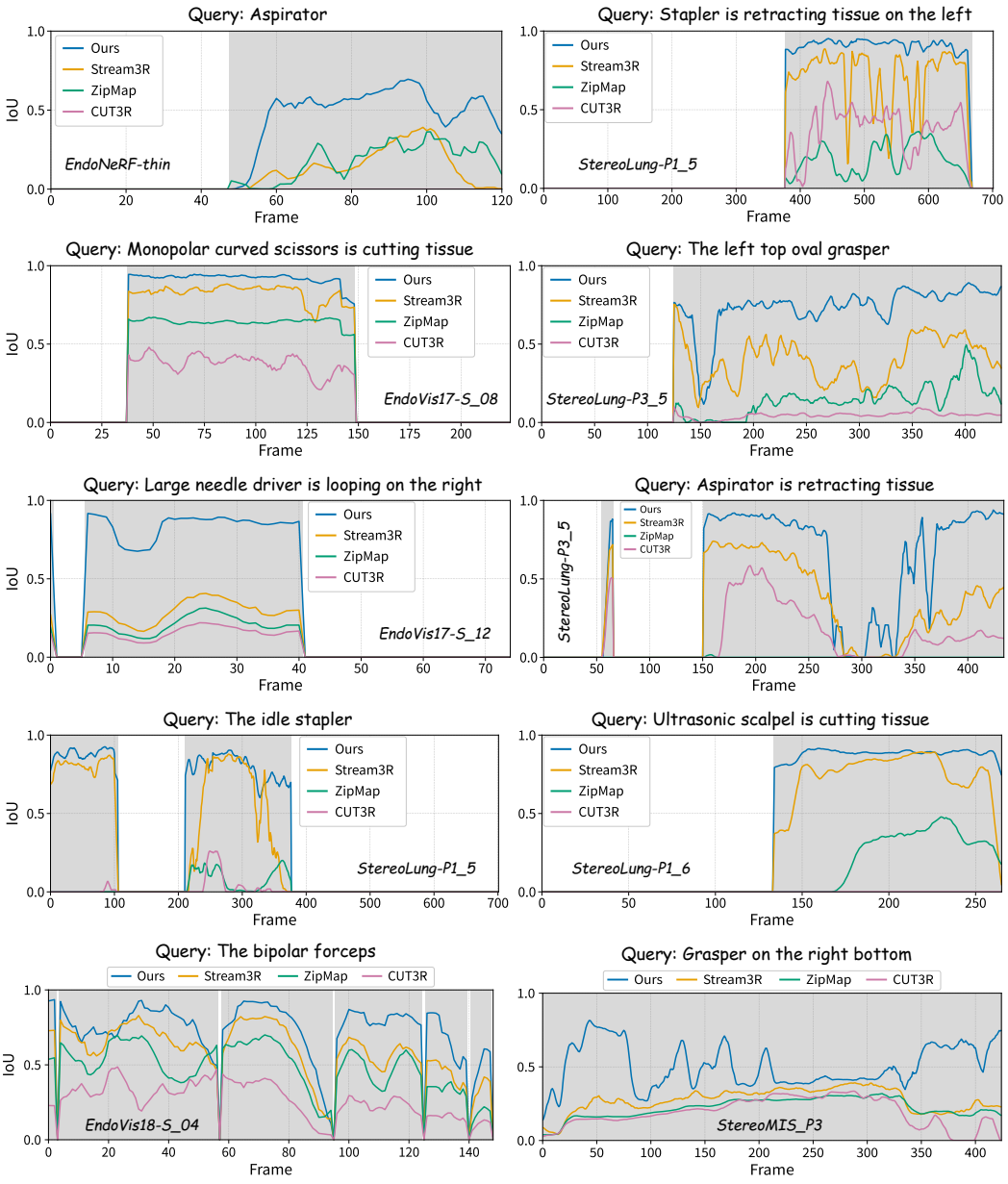
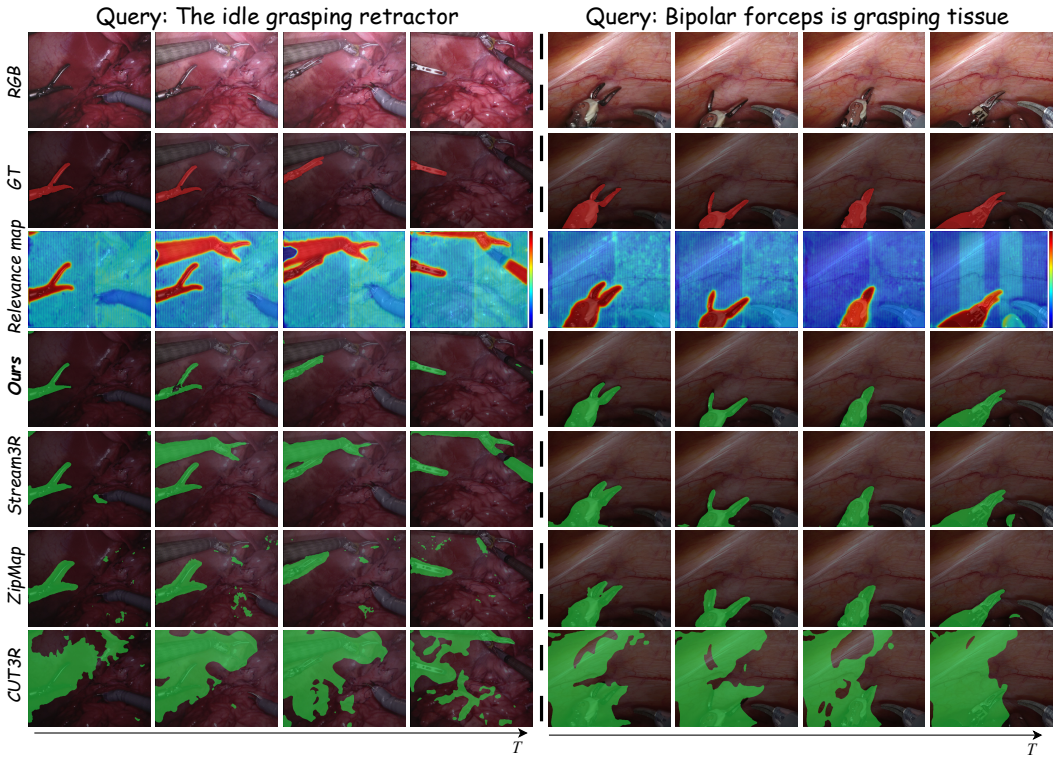
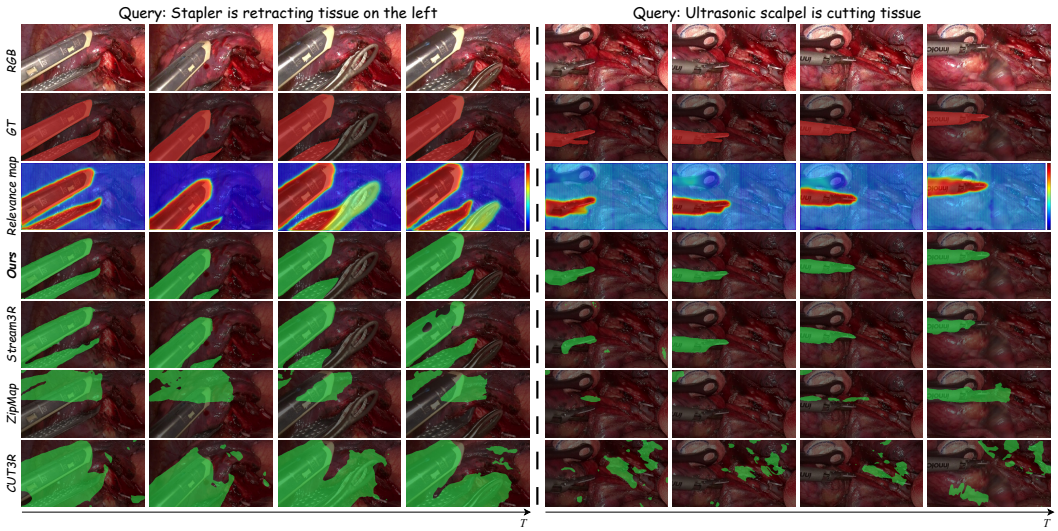


Figure 11: **Language grounding over unseen sequences.** Frame-wise IoU is compared with Stream3R (Lan et al., 2026), ZipMap (Jin et al., 2026), and CUT3R (Wang et al., 2025b) for different language queries. Each panel identifies the query and test sequence. Shaded regions indicate intervals with ground-truth annotations.



(a) Qualitative comparisons of language grounding on two unseen scenes with varying queries from EndoVis17 (Allan et al., 2019) and EndoNeRF (Wang et al., 2022).



(b) Qualitative comparisons of language grounding on two unseen scenes with varying queries from StereoLung (Zhou et al., 2025b).

Figure 12: **Additional language-grounding comparisons.** Rows display input RGB frames, ground-truth masks, relevance maps produced by *Endo4DLR*, and segmentation results from *Endo4DLR*, Stream3R (Lan et al., 2026), ZipMap (Jin et al., 2026), and CUT3R (Wang et al., 2025b). Ground-truth and predicted masks are overlaid in red and green, respectively. Zoom in for details.

D LIMITATIONS, FUTURE WORK AND BROADER IMPACT

While *Endo4DLR* demonstrates effective reconstruction and language grounding on unseen surgical scenes, its joint generalization to surgical procedures and imaging systems beyond the training distribution remains an area for further investigation. Moreover, current grounding evaluation focuses on instruments and descriptions of their actions and locations, as suitable datasets combining dynamic reconstruction data with language annotations for tissues and anatomical structures are still lacking. Our framework can seamlessly extend to these semantic targets using the same language-embedded Gaussian representation with corresponding supervision. Future work will explore adaptation across surgical domains and broader language annotations to support more diverse surgical environments and semantic targets.

By connecting evolving scene geometry with language semantics, this work provides a foundation for interactive understanding of surgical environments. The framework could enable natural-language exploration of surgical recordings, allowing users to locate instruments and examine their interactions within reconstructed scenes, with potential applications in surgical training and spatially grounded surgical embodied assistance systems.