SCOPE-4D: ENDOSCOPIC 4D GEOMETRY FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

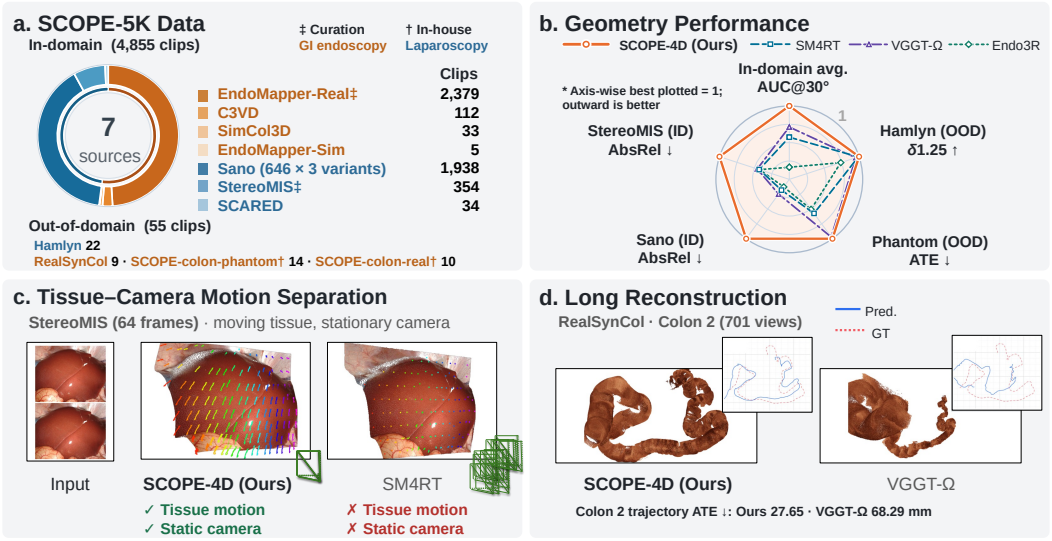

Figure 1: **SCOPE-4D overview.** (a) Our curation and annotation pipeline builds SCOPE-5K for geometric supervision and cross-domain evaluation. SCOPE-4D (b) achieves strong cross-domain geometry, (c) recovers tissue motion while preserving camera stability in a stationary-camera sequence, unlike SM4RT, and (d) enables stable long-sequence reconstruction.

ABSTRACT

Geometric understanding supports endoscopic navigation and robotic assistance, but learning reliable endoscopic geometry faces two challenges: scarce geometric annotations and ambiguity between camera motion and tissue deformation. We present SCOPE-4D, an endoscopic 4D geometry foundation model that jointly predicts camera parameters, dense geometry, and 3D tissue trajectories from monocular RGB video in a single forward pass. Our curation and annotation pipeline constructs SCOPE-5K, a collection of approximately 5,000 clips spanning real and synthetic gastrointestinal endoscopy and laparoscopy. The collection provides rich geometric supervision and includes newly collected phantom and real-colonoscopy evaluation sets. Geometric supervised fine-tuning on SCOPE-5K learns endoscopic priors that improve camera and depth estimation. Common–Residual Motion (CRM) further constrains local deformation relative to common tissue movement. Together with geometric supervision, CRM and trajectory supervision further improve camera and depth estimation over geometric fine-tuning alone while enabling dense 3D tissue tracking. Evaluations on public and newly collected benchmarks demonstrate strong in-domain and out-of-domain geometry, superior 3D tracking, and more stable long-sequence colon reconstruction. A blinded user study further supports the perceived reconstruction quality on real clinical video. Together, these results demonstrate the value of large-scale endoscopic supervision and motion constraints for joint geometry estimation and tissue tracking. **Anonymous project page:** https://scope4dreview.github.io/scope4d-anonymous/.

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.

1 INTRODUCTION

Endoscopic navigation, surgical planning, and robotic assistance require spatial understanding within a confined workspace and limited field of view. Recovering camera poses, depth, and dense point clouds from video can help clinicians assess tissue shape and position instruments.

Traditionally, Structure-from-Motion (SfM) and Multi-View Stereo (MVS) recover scene geometry from multiple images (Schönberger & Frahm, 2016; Schönberger et al., 2016). Geometry foundation models instead learn transferable priors through large-scale pretraining on natural scenes. DUSt3R predicts pointmaps from image pairs (Wang et al., 2024); VGGT and VGGT-Ω jointly predict cameras and dense geometry in a single forward pass (Wang et al., 2025a; 2026), and D4RT extends feed-forward prediction to dynamic 4D reconstruction (Zhang et al., 2026).

Endoscopy adds tissue deformation to this reconstruction problem. EndoNeRF and EndoGaussian model deforming tissue with camera and depth guidance, but require per-sequence optimization (Wang et al., 2022; Liu et al., 2024). Feed-forward methods avoid this fitting: AF-SfMLearner uses self-supervised image reconstruction (Shao et al., 2022), while Endo3R and Endo-FASt3R adapt pretrained general-domain models (Guo et al., 2025; Sheikh Zeinoddin et al., 2025). However, photometric supervision remains ambiguous under weak texture and deformation, and pretrained priors have limited endoscopic coverage.

Two challenges remain in building endoscopic geometry foundation models. **First, endoscopic geometric annotations are scarce and limited in coverage.** Datasets such as SCARED and C3VD provide valuable reference geometry, yet cover a limited range of subjects and acquisition conditions (Allan et al., 2021; Bobrow et al., 2023). Obtaining reliable geometric supervision from raw clinical video at scale is difficult: weak texture, specular reflections, liquids, bleeding, and illumination changes can disrupt the correspondences on which SfM relies. Limited annotations restrict the geometric priors learned across endoscopic domains. **Second, geometry supervision does not explicitly constrain how tissue points move over time.** Endoscopic scenes combine camera motion, coherent tissue motion such as breathing, and local deformation from tool interaction. These processes can produce similar image changes, creating ambiguity between camera movement and scene deformation. A model can therefore predict camera motion from deforming tissue even with a stationary camera, misaligning observations and distorting reconstructed geometry.

To expand reliable geometric supervision, we develop a data curation and annotation pipeline that combines vision-language screening, tracking-assisted camera reconstruction, and camera-conditioned dense annotation. We assess annotation reliability through geometric quality checks and comparisons between independent correspondence sources. With this pipeline, we construct SCOPE-5K, a collection of approximately 5,000 endoscopic clips covering real and synthetic gastrointestinal (GI) endoscopy and laparoscopy. The collection also includes two newly collected evaluation sets: *SCOPE-colon-phantom* provides externally tracked camera references in a physical colon phantom, while *SCOPE-colon-real* supports a blinded user study of reconstruction quality in clinical videos.

We present SCOPE-4D: Structured Camera and Organ Motion Prediction for Endoscopy, an endoscopic 4D geometry foundation model. To our knowledge, it is the first such model to jointly predict camera parameters, dense geometry, and 3D tissue trajectories from monocular RGB video in a single forward pass. The model combines supervised fine-tuning (SFT) for geometry with Common–Residual Motion (CRM). SFT on SCOPE-5K learns endoscopic geometric priors, improving depth and camera-pose estimation. To address the second challenge, CRM explicitly constrains tissue motion by composing a common SE(3) transform with point-dependent residual motion. With trajectory supervision, CRM further improves camera and depth estimation over geometric fine-tuning alone while enabling dense 3D tissue tracking.

Our main contributions are:

- **Data:** A curation and annotation pipeline constructs SCOPE-5K, approximately 5,000 endoscopic clips for learning and evaluation, including phantom and real-colonoscopy benchmarks.
- **Model:** SCOPE-4D combines geometric SFT with CRM and trajectory supervision to improve camera and depth estimation while enabling dense 3D tissue tracking.
- **Evaluation:** We demonstrate leading geometry performance across in-domain and held-out endoscopic benchmarks, improved 3D tissue tracking, and more accurate long-sequence camera reconstruction.

2 Related Work

General geometry and 4D models. Structure-from-Motion (SfM) and Multi-View Stereo (MVS) recover geometry through image correspondence and costly per-scene optimization (Schönberger & Frahm, 2016; Schönberger et al., 2016). Geometry foundation models learn feed-forward reconstruction through large-scale pretraining (Wang et al., 2024; Tang et al., 2025; Yang et al., 2025; Wang et al., 2025b;a; 2026; Lin et al., 2025; Keetha et al., 2026). DUSt3R predicts paired pointmaps without calibration; VGGT jointly predicts cameras and geometry through within-image and cross-image attention; VGGT-Ω improves attention efficiency to scale training. Depth Anything 3 represents single-image and multi-view geometry through depth and ray maps, while MapAnything incorporates available camera and depth information. Applying these advances to endoscopy requires broader data coverage and reliable annotations.

General 4D models extend reconstruction to dynamic scenes through joint geometry and motion prediction (Zhang et al., 2026; Karhade et al., 2025; Xiao et al., 2025). 4RC (Luo et al., 2026) separates base geometry from relative motion. SM4RT (Lin et al., 2026) introduces Structure-of-Motion (SoM) to represent point motion through shared motion bases with point-dependent assignments. In endoscopy, a tissue surface can move coherently with breathing while deforming locally under tool interaction. SCOPE-4D separates these motions through common–residual composition, constraining local deformation while preserving shared movement.

Endoscopic geometry models. Neural fields and Gaussian representations model deforming tissue, but require optimizing a new scene for each sequence (Wang et al., 2022; Zha et al., 2023; Liu et al., 2024; Hayoz et al., 2024). Methods that also estimate camera motion must distinguish it from tissue deformation; ambiguity between the two can lead to tracking drift (Shan et al., 2026). Feed-forward models instead learn across sequences and avoid per-scene fitting. Early approaches estimate depth and camera motion through self-supervised image reconstruction (Ozyoruk et al., 2021; Shao et al., 2022). AF-SfMLearner (Shao et al., 2022) compensates for appearance changes, but photometric supervision remains ambiguous in weakly textured or deforming tissue. More recent methods adapt pretrained general-domain models to endoscopic reconstruction. EndoDAC (Cui et al., 2024) uses parameter-efficient depth adaptation, while Endo3DAC (Cui et al., 2025) jointly estimates depth, pose, and intrinsics. Endo-FASt3R (Sheikh Zeinoddin et al., 2025) combines pretrained depth and pose models. Endo3R (Guo et al., 2025) supports online pointmap reconstruction through short-term and long-term memories. These models benefit from general-domain pretraining, but limited endoscopic supervision restricts their ability to handle anatomical and visual variation. Geometry supervision alone also leaves tissue motion insufficiently constrained. SCOPE-4D combines geometric SFT on SCOPE-5K with CRM to improve geometry estimation and constrain tissue deformation.

Endoscopic datasets. Endoscopic geometry learning is constrained by limited data scale and a shortage of dense, accurate geometric annotations (Allan et al., 2021; Bobrow et al., 2023; Azagra et al., 2023; Rau et al., 2024; Hayoz et al., 2023). SCARED (Allan et al., 2021) obtains depth through structured light in porcine cadavers and propagates it to intermediate frames under a static-scene assumption. C3VD (Bobrow et al., 2023) obtains reference geometry by registering colonoscopy images of physical phantoms to their 3D models. Both provide valuable supervision under controlled conditions, with limited coverage of clinical appearance and deformation. EndoMapper (Azagra et al., 2023) broadens clinical coverage through complete calibrated colonoscopy and gastroscopy procedures, but lacks dense geometric ground truth throughout its real videos. SimCol3D (Rau et al., 2024) provides synthetic depth and camera poses, while StereoMIS (Hayoz et al., 2023) provides in-vivo stereo video and camera poses from robot kinematics. Synthetic appearance and specialized acquisition setups limit the coverage of routine clinical conditions. To broaden coverage, SCOPE-5K combines existing datasets with curated and annotated clinical video. Phantom and real-colonoscopy benchmarks assess camera accuracy and reconstruction quality, respectively.

3 SCOPE-5K: Endoscopic Geometry Data

SCOPE-5K provides depth and camera supervision for learning endoscopic geometric priors, together with benchmarks for evaluating reconstruction quality. SCOPE-5K contains 4,855 in-domain clips across seven sources and 55 out-of-domain clips, spanning real and synthetic GI endoscopy and laparoscopy (Figure 1a). Our pipeline adds camera and dense geometry annotations to EndoMapper-

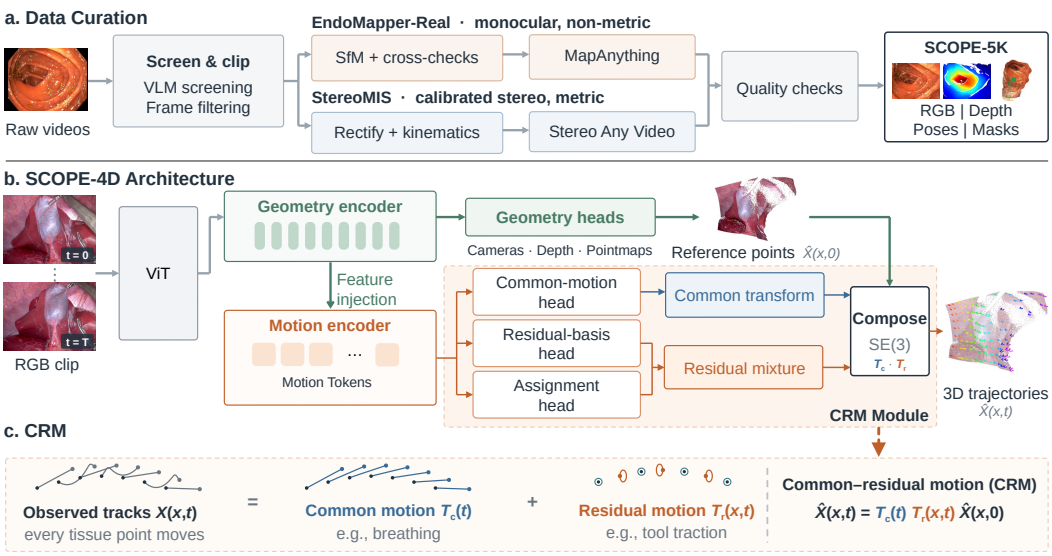

Figure 2: **Data annotation and SCOPE-4D.** (a) SCOPE-5K curation and annotation. (b) Joint camera, geometry, and 3D tissue-track prediction. (c) Common–Residual Motion composes common tissue movement with local deformation.

Real and dense depth to StereoMIS. We also introduce *SCOPE-colon-phantom* and *SCOPE-colon-real* for controlled camera evaluation and clinical reconstruction assessment. Appendix A.1 details source composition, annotation coverage, and splits.

3.1 Geometry Annotation

Our annotation pipeline combines clip selection, geometry reconstruction, and quality control, with reconstruction tailored to the available camera information (Figure 2a).

Monocular geometry annotation. Unlike datasets with dedicated capture guidelines such as DL3DV-10K (Ling et al., 2024), routine clinical videos contain obstructions, poor visibility, and abrupt camera motion. For EndoMapper-Real, a VLM selects segments with visible anatomy and sufficient camera movement, excluding obstructed or irrelevant content. Image-based filters reject frames with blur, exposure failures, specular reflections, weak texture, or abrupt illumination changes. We split the filtered videos at temporal discontinuities into reconstructable clips.

Weak texture, deformation, and specular reflections make local-feature matches unreliable for endoscopic SfM (Barbed et al., 2022). We instead convert MFT (Neoral et al., 2024) temporal tracks into pairwise matches for incremental SfM. LoMa (Nordström et al., 2026) and SIFT with LightGlue (Lowe, 2004; Lindenberger et al., 2023) provide independent reconstructions to cross-check the recovered camera trajectory after similarity alignment.

Conventional MVS estimates depth by photometric matching (Schönberger et al., 2016), which is unreliable under endoscopic illumination changes, reflections, and weak texture (Liu et al., 2020). We instead condition MapAnything (Keetha et al., 2026) on the recovered poses and intrinsics to generate dense depth and point clouds using its learned geometric priors.

Quality control checks multi-view consistency of dense predictions and agreement between independent camera reconstructions. Pixel masks exclude depth-edge artifacts and regions without multi-view support. Across 93 recordings, screening produces 3,034 clean clips; geometric QC and held-out filtering retain 2,379. Appendix A.4 details the checks, thresholds, and verification results.

Stereo geometry annotation. StereoMIS provides calibrated stereo video and camera poses from robot kinematics, allowing us to annotate depth without reconstructing cameras. We use Stereo Any Video (SAV) (Jing et al., 2025) to estimate disparity from rectified stereo pairs and convert it to depth using the known focal length and baseline. Clip selection and quality control follow the same

principles as the monocular pipeline, with checks tailored to stereo geometry. Appendices A.2–A.4 detail the VLM prompt, annotation procedures, and QC criteria.

3.2 SCOPE-colon Benchmark

We collect *SCOPE-colon-phantom* and *SCOPE-colon-real* to evaluate colonoscopy reconstruction. For *SCOPE-colon-phantom*, we synchronize video captured in a physical colon phantom with a five-degree-of-freedom electromagnetic tracker. The tracker provides camera position and viewing-direction references for quantitative evaluation. *SCOPE-colon-real* uses clinical colonoscopy videos for a blinded assessment of reconstruction quality. Together, they support controlled camera evaluation and clinical validation (Section 5); Appendix A.5 details their acquisition and annotations.

4 SCOPE-4D: Endoscopic 4D Geometry Foundation Model

SCOPE-4D combines geometric supervised fine-tuning on SCOPE-5K with Common–Residual Motion (CRM) to jointly predict cameras, dense geometry, and 3D tissue tracks from monocular RGB video (Figure 2b–c).

4.1 Geometry Supervised Fine-Tuning

To learn geometric priors suited to endoscopy, we apply supervised fine-tuning (SFT) to a pretrained geometry model using camera and depth annotations from SCOPE-5K. Given an uncalibrated monocular RGB video $\mathcal{I} = (I_0, \dots, I_{T-1})$, where $I_t \in \mathbb{R}^{H\times W\times 3}$, we use the first frame, $t = 0$, as the reference frame. An image encoder extracts frame features, which the backbone aggregates across views into geometry features $\{F_t\}_{t=0}^{T-1}$. Camera and depth heads decode these features into intrinsics $\mathbf{K}_t \in \mathbb{R}^{3\times 3}$, camera poses $C_t \in \mathrm{SE}(3)$, and depth maps $D_t \in \mathbb{R}^{H\times W}$. Here C_t maps camera coordinates into a shared reference frame. Back-projecting each depth map with $\mathbf{K}_t$ and applying C_t produces a dense pointmap $P_t \in \mathbb{R}^{H\times W\times 3}$ in this frame. The resulting mapping is

$$f : \mathcal{I} \mapsto \{\mathbf{K}_t, C_t, D_t, P_t\}_{t=0}^{T-1}. \tag{1}$$

We denote this geometry-only model as SCOPE-4D$_{\mathrm{SFT}}$ and the full geometry–motion model as SCOPE-4D. Tracking tissue points requires explicit motion modeling.

4.2 Common–Residual Motion

Common–Residual Motion (CRM) constrains tissue movement beyond camera and depth supervision, reducing ambiguity between camera motion and tissue deformation.

CRM representation. Structure-of-Motion (SoM) in SM4RT (Lin et al., 2026) mixes shared motion bases using point-dependent assignments fixed over time. In endoscopy, breathing induces coherent tissue movement while tool interaction causes local deformation. A single mixture can spread local deformation into surrounding tissue, yet suppressing it can weaken shared movement. CRM instead composes a shared rigid transform with a point-dependent residual transform (Figure 2c), constraining local deformation relative to common motion. Each transform is parameterized by a six-dimensional twist $\xi = (v, \omega)$, where $v, \omega \in \mathbb{R}^3$ are its translational and angular components. CRM uses a common twist $\xi_c(t)$ and K residual bases $\{\Delta\xi_k(t)\}_{k=0}^{K-1}$. The first basis is a fixed zero-residual anchor, $\Delta\xi_0(t) \equiv 0$, which introduces no local deformation. The model predicts the common twist and the remaining $K-1$ residual twists per frame. All twists are zero at $t = 0$, so each track begins at its reference point. For a reference-frame pixel x, logits $\ell_k(x)$ define time-invariant weights $q_k(x)$ that mix the residual bases:

$$q_k(x) = \frac{\exp \ell_k(x)}{\sum_{j=0}^{K-1} \exp \ell_j(x)}, \qquad \Delta\xi(x,t) = \sum_{k=1}^{K-1} q_k(x)\Delta\xi_k(t). \tag{2}$$

We convert twist coordinates into SE(3) transforms using the exponential map (Solà et al., 2018), defined in Appendix C. Composing the common and residual transforms gives

$$\widehat{X}(x,t) = T_c(t)T_r(x,t)X(x,0), \qquad T_c(t) = \mathrm{Exp}(\xi_c(t)), \quad T_r(x,t) = \mathrm{Exp}(\Delta\xi(x,t)). \tag{3}$$

Here $X(x,0)$ is the reference 3D point, and its predicted positions $\widehat{X}(x,t)$ form the 3D tissue track.

For a reference point $p = X(x,0)$, CRM bounds displacement relative to common motion:

$$\|\widehat{X}(x,t) - T_c(t)p\|_2 \le (1 - q_0(x))B(p,t), \tag{4}$$

where $B(p,t)$ depends only on the residual bases and the reference point (Appendix C). For fixed bases, increasing $q_0(x)$ tightens this bound. As $q_0(x) \to 1$, the point follows $T_c(t)p$. Thus, the anchor suppresses local deformation without freezing the tissue, whereas a zero mixed twist in SoM suppresses all point motion. Appendix C provides the proof.

4.3 Joint Geometry and Motion Prediction

SCOPE-4D decodes geometry and motion from shared features (Figure 2b). Tissue tracks are represented in the reference camera frame, separately from camera poses C_t. Joint trajectory supervision gives the shared features an explicit representation of tissue movement, helping camera and depth prediction account for deformation. Linear projections map four intermediate geometry-backbone layers into a motion transformer that combines spatial and temporal information. Common-motion, residual-motion, and dense assignment heads predict $\xi_c(t)$, $\Delta\xi_k(t)$, and reference-frame logits $\ell_k(x)$, respectively. Applying CRM to the reference pointmap P_0 produces dense tracks $\widehat{X}_t \in \mathbb{R}^{H\times W\times 3}$, where $\widehat{X}_t(x) = \widehat{X}(x,t)$ is the position at time t of reference point x. The complete mapping is

$$f_{4D} : \mathcal{I} \mapsto \{\mathbf{K}_t, C_t, D_t, P_t, \widehat{X}_t\}_{t=0}^{T-1}. \tag{5}$$

4.4 Training

We jointly supervise scene geometry and tissue motion:

$$\mathcal{L} = \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{tracking}}, \qquad \mathcal{L}_{\text{tracking}} = \lambda_{\text{tr}}\mathcal{L}_{\text{tr}} + \lambda_c\mathcal{L}_c + \lambda_q\mathcal{L}_q, \tag{6}$$

where the λ terms balance the motion objectives. Following the MapAnything training framework (Keetha et al., 2026), geometry supervision uses $\mathcal{L}_{\text{geo}} = \mathcal{L}_{\text{pts}} + \mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{depth}}$ to constrain predicted pointmaps, camera parameters, and depth maps against their annotations. Motion supervision also updates the shared backbone, shaping the features used for camera and depth prediction.

Tracking labels combine trajectories from the synthetic Sano dataset (Nelson et al., 2026), derived from rendered flow and depth, with MFT correspondences lifted into 3D using StereoMIS depth and camera poses. StereoMIS tracks exclude occlusions and instruments. Trajectory supervision $\mathcal{L}_{\text{tr}}$ compares predicted and annotated 3D displacements with a weighted squared loss, balancing points with larger and smaller residual motion. For the common-motion target, we robustly select inlier correspondences $\mathcal{J}_t$ from annotated tracks $X_i(t)$ and fit a rigid transform:

$$T_c^\star(t) = \arg\min_{G\in \text{SE}(3)} \sum_{i\in\mathcal{J}_t} \|GX_i(0) - X_i(t)\|_2^2. \tag{7}$$

The common-motion loss $\mathcal{L}_c$ compares the point positions produced by predicted and fitted transforms. Assignment supervision $\mathcal{L}_q$ uses binary KL divergence between predicted anchor probabilities and targets derived from track residuals relative to this transform. Small residuals receive high anchor targets, encouraging points to follow common motion without unnecessary local deformation. Appendix B.2 gives the target construction and loss definitions.

5 Experiments

Implementation details. Geometric SFT and the full model are independently initialized from the same pretrained VGGT-Ω geometry backbone (Wang et al., 2026) and trained on SCOPE-5K for 29,000 steps. The full model initializes its motion modules from SM4RT (Lin et al., 2026). Appendices B.1 and D detail training and evaluation.

5.1 3D Geometry Reconstruction and Tracking

3D geometry reconstruction. Geometric SFT on SCOPE-5K substantially improves camera and depth estimation (Table 1). Relative to pretrained VGGT-Ω, in-domain (ID) pose AUC@5° rises

Table 1: **Camera pose and depth across domains.** Hamlyn pose measures camera stability; phantom ARE* excludes roll. Gain compares our best variant per metric with the best competing method.

	In-domain average					Hamlyn (OOD)				*SCOPE-colon-phantom* (OOD)		
	Pose			Depth		Pose		Depth		Pose		
Method	AUC@5° ↑	AUC@30° ↑	ATE↓	AbsRel↓	$\delta_{1.25}$ ↑	ATE↓	ARE↓	AbsRel↓	$\delta_{1.25}$ ↑	ATE↓	RTE↓	ARE* ↓
General-domain												
VGGT	2.94	29.11	0.128	0.160	74.61	4.610	1.964	0.190	69.19	16.53	16.78	12.14
VGGT-Ω	7.81	48.07	0.049	0.093	91.07	4.496	1.746	0.135	80.82	8.54	10.79	8.22
MapAnything	2.63	26.22	0.122	0.141	85.00	4.984	1.960	0.166	74.15	15.98	17.70	26.98
CUT3R	0.27	15.55	0.131	0.300	54.06	4.571	1.693	0.185	70.27	13.64	16.56	22.48
SpatialTrackerV2	6.01	44.85	0.058	0.113	86.98	5.499	2.241	0.142	77.42	10.92	13.80	11.23
4RC	6.12	41.85	0.068	0.138	79.71	3.639	1.438	**0.120**	84.26	13.98	15.33	9.13
Any4D	0.29	11.18	0.160	0.215	65.25	4.755	1.114	0.167	72.23	16.94	17.57	12.15
SM4RT	4.86	38.72	0.087	0.147	77.78	3.517	1.350	**0.120**	83.94	14.46	15.50	9.17
Endoscopy-specific												
Endo3R	0.26	11.14	0.179	0.185	70.06	20.982	9.492	0.223	62.77	16.42	17.05	16.13
AF-SfMLearner	1.45	20.76	0.134	0.253	56.94	15.562	5.979	0.190	68.20	10.35	13.88	10.31
EndoDAC	1.41	19.08	0.130	0.201	66.22	18.057	8.968	0.191	68.87	11.51	15.03	10.54
Endo-FASt3R	4.13	35.02	0.062	0.230	60.39	9.152	3.794	0.211	64.75	8.70	11.39	9.22
SCOPE-4D$_{SFT}$ (Ours)	29.44	67.33	**0.016**	0.042	98.52	**2.927**	1.014	0.129	83.77	8.41	**10.71**	7.37
SCOPE-4D (Ours)	**29.86**	**67.41**	0.017	**0.040**	**98.68**	2.976	**1.013**	0.124	**85.23**	**8.29**	10.82	**7.21**
Gain (%)	+282.2%	+40.2%	+67.2%	+56.8%	+8.4%	+16.8%	+9.0%	-3.6%	+1.2%	+2.9%	+0.7%	+12.3%

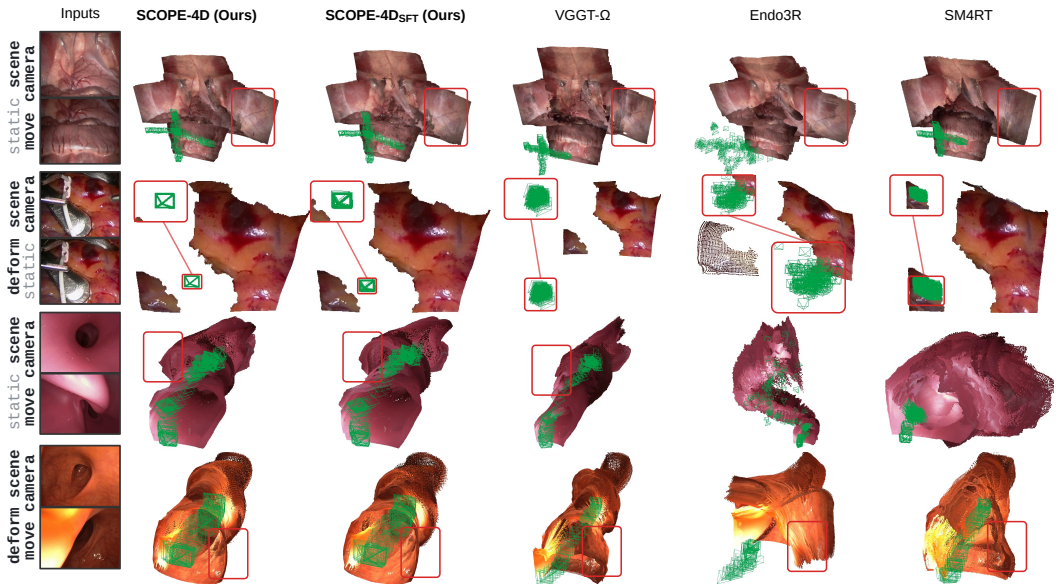

Figure 3: **Qualitative comparison of 3D geometry reconstruction.** Instrument regions are masked out. Highlighted regions show surface consistency and suppression of spurious camera motion.

from 7.81 to 29.44, while depth AbsRel falls from 0.093 to 0.042. Geometric SFT outperforms all compared endoscopy-specific methods on every metric reported in Table 1. Full SCOPE-4D further improves ID pose AUC and depth accuracy. On out-of-domain (OOD) benchmarks, without dataset-specific fine-tuning, the full model outperforms competing methods in camera-stability ATE, ARE, and $\delta_{1.25}$ on the held-out laparoscopic Hamlyn benchmark (Recasens et al., 2021) and achieves the lowest ATE and viewing-direction error on *SCOPE-colon-phantom*. Figure 3 shows more coherent surfaces, fewer duplicated layers, and better-preserved colon structure. In its stationary-camera example, the full model yields more tightly clustered camera predictions than geometric SFT despite tissue deformation, while VGGT-Ω, Endo3R, and SM4RT exhibit more pronounced spurious camera motion. Appendix E provides per-dataset results and further comparisons.

3D tracking. SCOPE-4D achieves the highest APD on both Sano and StereoMIS and the lowest EPE on StereoMIS (Table 2). On StereoMIS, it improves EPE by 19.8% over the best competing method. It also outperforms the competing tracking methods in camera and depth metrics on both datasets. Figure 4 shows coherent tissue trajectories with larger local displacements concentrated near instrument interactions. SM4RT and 4RC underestimate shared tissue movement in these examples, whereas SCOPE-4D preserves it while reducing spurious motion around local interactions. APD at individual distance thresholds is reported in Appendix E.1.

Table 2: **3D tracking and geometry on Sano and StereoMIS.** D4RT*: open-source reproduction.

	Sano							StereoMIS						
	Tracking		Pose			Depth		Tracking		Pose			Depth	
Method	EPE↓	APD↑	ATE↓	RTE↓	RRE↓	AbsRel↓	$\delta_{1.25}$ ↑	EPE↓	APD↑	ATE↓	RTE↓	RRE↓	AbsRel↓	$\delta_{1.25}$ ↑
D4RT*	2.035	72.94	0.014	0.239	0.215	0.193	65.76	7.540	34.12	0.043	0.940	0.344	0.147	87.20
SpatialTrackerV2	0.933	90.58	0.017	0.223	0.125	0.145	78.76	6.870	38.98	0.049	0.677	0.195	0.107	95.24
4RC	0.523	94.72	0.006	0.100	0.062	0.186	66.15	6.266	44.28	0.039	0.481	0.146	0.114	90.42
Any4D	0.535	94.41	0.010	0.376	0.309	0.188	66.01	7.277	42.45	0.061	2.203	0.765	0.144	82.72
SM4RT	**0.411**	95.64	0.007	0.104	0.081	0.182	67.31	6.276	43.95	0.025	0.427	0.158	0.116	92.57
SCOPE-4D (Ours)	0.442	**96.11**	**0.001**	**0.054**	**0.051**	**0.033**	**99.75**	**5.022**	**53.08**	**0.010**	**0.211**	**0.114**	**0.050**	**98.80**
Gain (%)	-7.5%	+0.5%	+81.6%	+45.9%	+18.4%	+77.1%	+26.7%	+19.8%	+19.9%	+59.4%	+50.7%	+21.9%	+53.3%	+3.7%

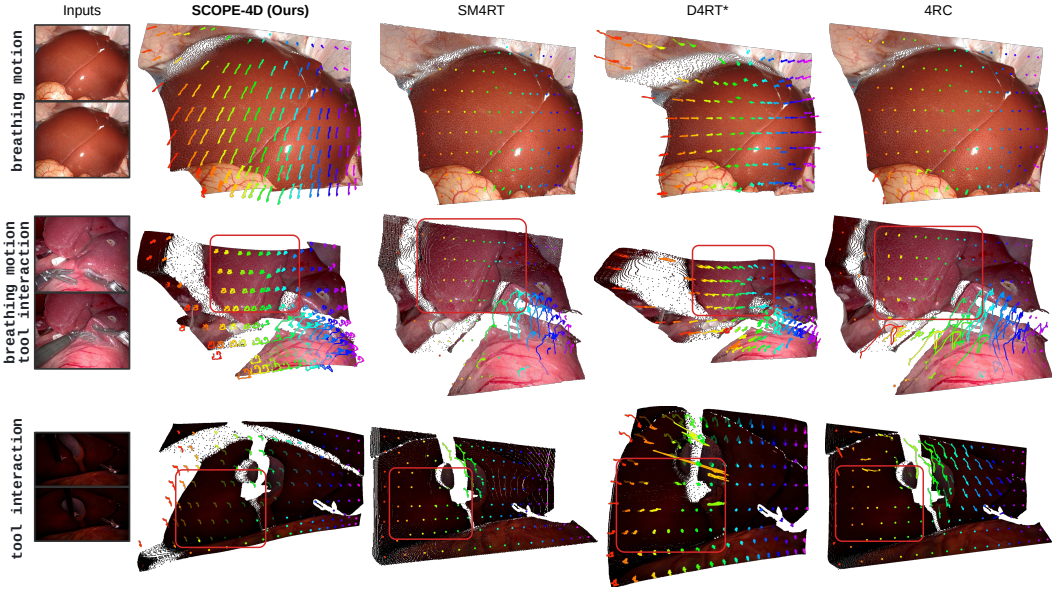

Figure 4: **3D tissue tracking under different motion conditions.** Instrument regions are masked out. Red boxes highlight the suppression of spurious motion.

5.2 Ablation Studies

Data composition. To assess our annotation pipeline's value, we add StereoMIS (SM), EndoMapper-Real (EM-R), or both to geometry-only SFT on five datasets: SCARED (SC), SimCol3D (SIM), C3VD, EndoMapper-Sim (EM-S), and Sano (SA) (Table 3(a)). ID averages cover these five datasets. Both added sources improve depth accuracy on unseen Hamlyn data and provide complementary gains in phantom camera estimation. Combining them yields the strongest overall depth results among the variants. These results demonstrate the value of SCOPE-5K's diverse geometric supervision for generalization across endoscopic domains.

Motion design. On StereoMIS (Table 3(b)), full CRM improves pose AUC@5° by 22.8% and tracking APD by 13.0% over SoM, with comparable depth accuracy. Adding a zero-motion anchor and assignment supervision improves EPE and APD; common–residual composition and its supervision further improve pose AUC and APD. These results support motion modeling for both geometry estimation and tissue tracking. Table 3(c) shows reduced spurious motion in surrounding tissue while retaining deformation near the interaction. Appendix E.1 reports APD at all thresholds.

5.3 Downstream Tasks and User Study

Long-sequence reconstruction. We evaluate on RealSynCol (Lena et al., 2026) using a framework similar to VGGT-Long (Deng et al., 2025). All models reconstruct overlapping chunks and align them sequentially with Sim(3) transforms, without loop closure or global alignment. Both SCOPE-4D variants outperform the baselines (Table 4). Gain compares the full model with the best competing

Table 3: **Data and motion ablations.** (a) Geometry-only SFT by data source. (b) StereoMIS motion design. (c) SoM vs. CRM. Gain in (a): full dataset versus pretrained VGGT-Ω.

(a) Data composition

	In-domain average				Hamlyn (OOD)		*SCOPE-colon-phantom* (OOD)	
	Pose		Depth		Depth		Pose	
	AUC@5° ↑	AUC@30° ↑	AbsRel↓	$\delta_{1.25}$ ↑	AbsRel↓	$\delta_{1.25}$ ↑	RTE↓	ARE* ↓
No finetune (VGGT-Ω weights)	9.55	53.03	0.077	94.03	0.135	80.82	10.79	8.22
w/ SC+SIM+C3VD+EM-S+SA	35.99	**74.08**	0.031	99.50	0.143	80.93	11.28	7.12
w/ SC+SIM+C3VD+EM-S+SA+SM	33.10	70.02	0.032	99.53	0.142	81.40	10.51	**6.83**
w/ SC+SIM+C3VD+EM-S+SA+EM-R	**36.62**	74.07	0.032	99.57	**0.128**	83.04	**10.37**	7.37
Full dataset	36.02	73.68	**0.030**	**99.59**	0.129	**83.77**	10.71	7.37
Gain (%) vs No finetune	+277.1%	+38.9%	+61.6%	+5.9%	+4.3%	+3.7%	+0.7%	+10.4%

(b) Motion design

Metric	SoM (baseline)	Full model + anchor	Full model + CRM	Gain (%)
EPE↓	5.94	5.30	**5.27**	+11.2%
APD↑	46.58	50.11	**52.63**	+13.0%
AUC@5° ↑	5.67	6.09	**6.96**	+22.8%
AUC@30° ↑	26.86	27.17	**27.66**	+3.0%
AbsRel↓	0.050	**0.048**	0.049	+0.3%
$\delta_{1.25}$ ↑	98.82	98.82	**98.85**	+0.0%

(c) Qualitative motion comparison

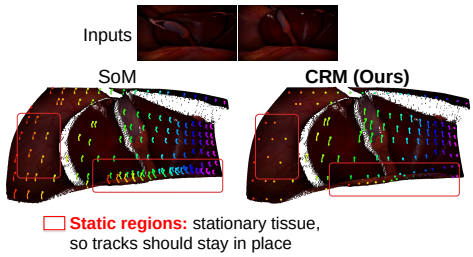

Table 4: **Long-sequence reconstruction.**

Method	ATE↓	RTE↓	RRE↓
VGGT	93.21	23.65	15.11
VGGT-Ω	85.92	21.67	31.21
SM4RT	88.95	21.29	16.98
Endo3R	77.87	14.75	22.92
SCOPE-4D$_{\text{SFT}}$ (Ours)	73.52	**10.38**	**13.38**
SCOPE-4D (Ours)	**71.17**	10.94	13.61
Gain (%)	+8.6%	+25.9%	+9.9%

Table 5: **Clinical expert assessment.**

Method	Camera↓	Point↓	Depth↓
VGGT-Ω	1.93 [1.65, 2.25]	1.95 [1.65, 2.28]	1.98 [1.67, 2.32]
SM4RT	2.90 [2.60, 3.08]	3.22 [2.80, 3.57]	3.33 [2.87, 3.73]
Endo3R	3.83 [3.53, 4.00]	3.52 [3.23, 3.78]	3.37 [3.00, 3.70]
SCOPE-4D (Ours)	**1.33** [1.12, 1.62]	**1.32** [1.13, 1.55]	**1.32** [1.13, 1.53]

method for each metric. The full model achieves lower ATE than geometric SFT, supporting joint geometry and motion modeling for long-sequence camera estimation (Appendix Figure 16).

User study. Six clinical and biomedical experts evaluate ten *SCOPE-colon-real* clips with method identities hidden, ranking four reconstructions separately for camera trajectories, point clouds, and depth maps (1 best, 4 worst). Table 5 reports mean ranks with 95% bootstrap confidence intervals in brackets. SCOPE-4D achieves the best mean rank and ranks first in 70% of evaluations in each dimension. In paired comparisons with pretrained VGGT-Ω, participants prefer SCOPE-4D in 75.0%, 75.0%, and 76.7% of evaluations, respectively. Appendix D.3 details the protocol and analysis.

6 Conclusion

We present SCOPE-4D, an endoscopic 4D geometry foundation model that jointly predicts camera parameters, dense geometry, and 3D tissue tracks from monocular RGB video. Our curation and annotation pipeline constructs SCOPE-5K to expand geometric supervision and evaluation coverage. Geometric SFT learns endoscopic priors from this data, while CRM separates common tissue movement from local deformation. Together with trajectory supervision, this representation further improves geometry over geometric SFT alone while enabling dense tissue tracking. Experiments demonstrate accurate geometry across ID and OOD benchmarks, improved 3D tracking, and more stable long-sequence reconstruction. A blinded study with six clinical and biomedical experts also favors SCOPE-4D on real clinical video. The phantom and clinical benchmarks assess tracked camera accuracy and perceived reconstruction quality, respectively. These findings support combining broader endoscopic supervision with explicit tissue-motion modeling for feed-forward 4D reconstruction.

AI USE STATEMENT

Generative AI assisted manuscript drafting, editing, and literature and code inspection. A vision-language model also supports video screening in the data pipeline (Section 3).

REPRODUCIBILITY STATEMENT

We document data splits and sampling, annotation checks, motion-target construction, training settings, and dataset-specific evaluation protocols. Complete per-domain geometry and tracking-threshold results accompany the main comparisons.

ETHICS STATEMENT

This work combines existing endoscopic datasets with in-house phantom recordings and clinical video. The clinical clips in *SCOPE-colon-real* are used for retrospective reconstruction assessment without altering clinical procedures or patient care. Six clinical and biomedical experts participated in a blinded reconstruction-ranking study (Appendix D.3). The reported experiments evaluate reconstruction quality and do not establish safety for clinical deployment.

REFERENCES

Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, Zhanshi Zhu, Huoling Luo, Fucang Jia, Xiran Zhang, Xiaohong Li, Lalith Sharan, Tom Kurmann, Sebastian Schmid, Raphael Sznitman, Dimitris Psychogyios, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021. 2, 3, 14

Pablo Azagra, Carlos Sostres, Ángel Ferrández, Luis Riazuelo, Clara Tomasini, O. León Barbed, Javier Morlana, David Recasens, Víctor M. Batlle, Juan J. Gómez-Rodríguez, Richard Elvira, Julia López, Cristina Oriol, Javier Civera, Juan D. Tardós, Ana C. Murillo, Angel Lanas, and José M. M. Montiel. Endomapper dataset of complete calibrated endoscopy procedures. *Scientific Data*, 10(1): 671, 2023. doi: 10.1038/s41597-023-02564-7. 3, 14

O. León Barbed, François Chadebecq, Javier Morlana, José M. M. Montiel, and Ana C. Murillo. SuperPoint features in endoscopy. In *Imaging Systems for GI Endoscopy, and Graphs in Biomedical Image Analysis*, pp. 45–55, 2022. doi: 10.1007/978-3-031-21083-9_5. 4

Taylor L. Bobrow, Mayank Golhar, Rohan Vijayan, Venkata S. Akshintala, Juan R. Garcia, and Nicholas J. Durr. Colonoscopy 3D video dataset with paired depth from 2D–3D registration. *Medical Image Analysis*, pp. 102956, 2023. 2, 3, 14

Beilei Cui, Mobarakol Islam, Long Bai, An Wang, and Hongliang Ren. EndoDAC: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In *MICCAI*, 2024. 3, 22

Beilei Cui, Long Bai, Mobarakol Islam, An Wang, Zhiqi Ma, Yiming Huang, Feng Li, Zhen Chen, Zhongliang Jiang, Nassir Navab, and Hongliang Ren. Learning to efficiently adapt foundation models for self-supervised endoscopic 3D scene reconstruction from any cameras. *arXiv preprint arXiv:2503.15917*, 2025. 3

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. VGGT-Long: Chunk it, loop it, align it—pushing VGGT's limits on kilometer-scale long RGB sequences. *arXiv preprint arXiv:2507.16443*, 2025. 8, 22

Mayank V. Golhar, Lucas Sebastian Galeano Fretes, Loren Ayers, Venkata S. Akshintala, Taylor L. Bobrow, and Nicholas J. Durr. C3VDv2 – colonoscopy 3D video dataset with enhanced realism. *Scientific Data*, 2026. doi: 10.1038/s41597-026-07471-1. 14

Jiaxin Guo, Wenzhen Dong, Tianyu Huang, Hao Ding, Ziyi Wang, Haomin Kuang, Qi Dou, and Yun-Hui Liu. Endo3R: Unified online reconstruction from dynamic monocular endoscopic video. In *MICCAI*, 2025. 2, 3, 21, 22

Hamlyn Centre. Hamlyn Centre Laparoscopic / Endoscopic Video Datasets. Imperial College London, n.d. URL http://hamlyn.doc.ic.ac.uk/vision/. Dataset collection. 14

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. doi: 10.1007/s11548-023-02919-w. URL https://doi.org/10.1007/s11548-023-02919-w. 3, 14

Michel Hayoz, Christopher Hahne, Thomas Kurmann, Max Allan, Guido Beldi, Daniel Candinas, Pablo Márquez-Neila, and Raphael Sznitman. Online 3D reconstruction and dense tracking in endoscopic videos. In *MICCAI*, 2024. 3

Junpeng Jing, Weixun Luo, Ye Mao, and Krystian Mikolajczyk. Stereo any video: Temporally consistent stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20836–20846, 2025. 4, 17

Jay Karhade, Nikhil Keetha, Yuchen Zhang, Tanisha Gupta, Akash Sharma, Sebastian Scherer, and Deva Ramanan. Any4D: Unified feed-forward metric 4D reconstruction. *arXiv preprint arXiv:2512.10935*, 2025. 3, 22

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *3DV*, 2026. URL https://map-anything.github.io/. 3, 4, 6, 15, 19, 21, 22

Chiara Lena, Davide Milesi, Alessandro Casella, Luca Carlini, Joseph C. Norton, James Martin, Bruno Scaglioni, Keith L. Obstein, Roberto De Sire, Marco Spadaccini, Cesare Hassan, Pietro Valdastri, and Elena De Momi. RealSynCol: A high-fidelity synthetic colon dataset for 3D reconstruction applications. *arXiv preprint arXiv:2602.08397*, 2026. 8, 14

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 3

Shing Ho J. Lin, Wenzhao Zheng, Dong Zhuo, Yuqi Wu, Jie Zhou, and Jiwen Lu. SM4RT: Learning structured motion geometry for 4D reconstruction. *arXiv preprint arXiv:2607.22534*, 2026. 3, 5, 6, 18, 21, 22

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *ICCV*, pp. 17627–17638, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Lindenberger_LightGlue_Local_Feature_Matching_at_Light_Speed_ICCV_2023_paper.html. 4, 15

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024. 4

Xingtong Liu, Ayushi Sinha, Masaru Ishii, Gregory D. Hager, Austin Reiter, Russell H. Taylor, and Mathias Unberath. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE Transactions on Medical Imaging*, 39(5):1438–1447, 2020. doi: 10.1109/TMI.2019.2950936. 4

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. 2, 3

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94. 4, 15

Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, and Chen Change Loy. 4RC: 4D reconstruction via conditional querying anytime and anywhere. *arXiv preprint arXiv:2602.10094*, 2026. 3, 22

Nigel Nelson, Juo-Tung Chen, Jesse Haworth, et al. Open-h-embodiment: A large-scale dataset for enabling foundation models in medical robotics. *arXiv preprint arXiv:2604.21017*, 2026. URL https://arxiv.org/abs/2604.21017. 6, 14

Michal Neoral, Jonáš Šerých, and Jiří Matas. MFT: Long-term tracking of every pixel. In *WACV*, pp. 6837–6847, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Neoral_MFT_Long-Term_Tracking_of_Every_Pixel_WACV_2024_paper.html. 4, 15

David Nordström, Johan Edstedt, Georg Bökman, Jonathan Astermark, Anders Heyden, Viktor Larsson, Mårten Wadenbäck, Michael Felsberg, and Fredrik Kahl. LoMa: Local feature matching revisited. *arXiv preprint arXiv:2604.04931*, 2026. URL https://arxiv.org/abs/2604.04931. 4, 15

Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L. Bobrow, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, Hasan Sahin, Helder Araujo, Henrique Alexandrino, Nicholas J. Durr, Hunter B. Gilbert, and Mehmet Turan. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical Image Analysis*, 71:102058, 2021. doi: 10.1016/j.media.2021.102058. 3

Anita Rau, Sophia Bano, Yueming Jin, Pablo Azagra, Javier Morlana, Rawen Kader, Edward Sanderson, Bogdan J. Matuszewski, Jae Young Lee, Dong-Jae Lee, Erez Posner, Netanel Frank, Varshini Elangovan, Sista Raviteja, Zhengwen Li, Jiquan Liu, Seenivasan Lalithkumar, Mobarakol Islam, Hongliang Ren, Laurence B. Lovat, José M. M. Montiel, and Danail Stoyanov. SimCol3D — 3D reconstruction during colonoscopy challenge. *Medical Image Analysis*, 96:103195, 2024. doi: 10.1016/j.media.2024.103195. 3, 14

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714. 21

David Recasens, José Lamarca, José M. Fácil, J. M. M. Montiel, and Javier Civera. Endo-Depth-and-Motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robotics and Automation Letters*, 6(4):7225–7232, 2021. 7, 14

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pp. 4104–4113, 2016. 2, 3

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, pp. 501–518, 2016. doi: 10.1007/978-3-319-46487-9_31. 2, 3, 4

Jiwei Shan, Zeyu Cai, Yirui Li, Yongbo Chen, Lijun Han, Yun-hui Liu, Hesheng Wang, and Shing Shin Cheng. NRGS-SLAM: Monocular non-rigid SLAM for endoscopy via deformation-aware 3D gaussian splatting. *arXiv preprint arXiv:2602.17182*, 2026. 3

Shuwei Shao, Zhongcai Pei, Weihai Chen, Wentao Zhu, Xingming Wu, Dianmin Sun, and Baochang Zhang. Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. *Medical Image Analysis*, 77:102338, 2022. 2, 3, 22

Mona Sheikh Zeinoddin, Mobarak I. Hoque, Zafer Tandogdu, Greg L. Shaw, Matthew J. Clarkson, Evangelos B. Mazomenos, and Danail Stoyanov. Endo-FASt3r: Endoscopic foundation model adaptation for structure from motion. In *MICCAI*, 2025. 2, 3, 22

Joan Solà, Jeremie Deray, and Dinesh Atchuthan. A micro Lie theory for state estimation in robotics. *arXiv preprint arXiv:1812.01537*, 2018. 5

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. MV-DUSt3R+: Single-stage scene reconstruction from sparse views in 2 seconds. In *CVPR*, pp. 5283–5293, 2025. 3

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025a. 2, 3, 21, 22

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026. 2, 3, 6, 18, 21, 22

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *CVPR*, 2025b. 3, 22

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, 2024. 2, 3

Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery. In *MICCAI*, 2022. 2, 3

Mateusz Wójcikowski, Przemysław Korzeniowski, Sabina Martyniak, Michał Naskręt, Filippo Filicori, Aditya Amit Godbole, Maria Clara Morais, and Mattia Ballo. SANO_VR_Cholecystectomy. Open-H-Embodiment dataset, Hugging Face, 2025. URL https://huggingface.co/datasets/nvidia/PhysicalAI-Robotics-Open-H-Embodiment/blob/main/Surgical/sanoscience/sanoscience_v1_2_merged/nonexpert_partial_modalities/README.md. Simulated Domain: Robotic Cholecystectomy. 14

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: 3D point tracking made easy. *arXiv preprint arXiv:2507.12462*, 2025. 3, 22

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *CVPR*, pp. 21924–21935, 2025. 3

Ruyi Zha, Xuelian Cheng, Hongdong Li, Mehrtash Harandi, and Zongyuan Ge. EndoSurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In *MICCAI*, 2023. 3

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently reconstructing dynamic scenes one D4RT at a time. In *CVPR*, 2026. URL https://d4rt-paper.github.io/. 2, 3, 22

APPENDIX

In the appendix, we provide the following:

- Section A: Dataset construction.
- Section B: Implementation details.
- Section C: Proof of the CRM residual bound.
- Section D: Evaluation protocols.
- Section E: Additional results.

A DATASET CONSTRUCTION

A.1 DATA COMPOSITION AND SPLITS

SCOPE-5K covers real and synthetic GI endoscopy and laparoscopy. Table 6 lists sources and annotation coverage; Table 7 gives collection sizes, disjoint training, validation, and test splits, and sampling proportions. SCARED and EndoMapper-Real hold out source recordings for testing; C3VD holds out the descending region. SimCol3D and EndoMapper-Sim split by synthetic sequence, StereoMIS by non-overlapping temporal segment, and Sano by simulation attempt.

Table 6: **Data composition and annotation coverage of SCOPE-5K.** Annotations reflect those available in SCOPE-5K. † denotes annotations generated by our pipeline.

Data source	Domain		Data type		Annotations		Clips
	GI	Laparo	Synthetic	Real	Depth	Camera	
EndoMapper-Real (Azagra et al., 2023)	✓	–	–	✓	✓†	✓†	2,379
C3VD (Bobrow et al., 2023)	✓	–	–	✓	✓	✓	112
SimCol3D (Rau et al., 2024)	✓	–	✓	–	✓	✓	33
EndoMapper-Sim (Azagra et al., 2023)	✓	–	✓	–	✓	✓	5
Sano (Nelson et al., 2026)	–	✓	✓	–	✓	✓	1,938
StereoMIS (Hayoz et al., 2023)	–	✓	–	✓	✓†	✓	354
SCARED (Allan et al., 2021)	–	✓	–	✓	✓	✓	34
Hamlyn (Hamlyn Centre, n.d.)	–	✓	–	✓	✓	–	22
RealSynCol (Lena et al., 2026)	✓	–	✓	–	✓	✓	9
SCOPE-colon-phantom	✓	–	–	✓	–	✓	14
SCOPE-colon-real	✓	–	–	✓	–	–	10
SCOPE-5K (Ours)	✓	✓	✓	✓	✓	✓	**4,910**

Annotation sources. SCARED provides structured-light-derived 3D points, camera poses, and calibration (Allan et al., 2021). C3VD provides depth, camera poses, and calibration registered to phantom geometry (Bobrow et al., 2023; Golhar et al., 2026). SimCol3D and EndoMapper-Sim provide rendered depth, camera poses, and intrinsics (Rau et al., 2024; Azagra et al., 2023). Sano denotes the SANO_VR_Cholecystectomy subset of Open-H-Embodiment (Nelson et al., 2026), which provides rendered depth, camera parameters, and segmentation labels. The contributor's dataset card provides the source description and access link (Wójcikowski et al., 2025). We use these annotations to supervise depth, cameras, and pointmaps.

EndoMapper-Real provides calibrated clinical video without dense geometric ground truth (Azagra et al., 2023), while StereoMIS provides calibrated stereo video and camera poses from robot kinematics (Hayoz et al., 2023). Our pipeline adds cameras and dense geometry to EndoMapper-Real and dense depth to StereoMIS to supervise depth, cameras, and pointmaps.

For evaluation, Hamlyn provides depth and calibration but no camera-pose ground truth for the selected clips (Recasens et al., 2021); RealSynCol provides rendered depth, camera trajectories, and intrinsics (Lena et al., 2026). *SCOPE-colon-phantom* provides tracked camera positions and viewing directions; *SCOPE-colon-real* supports specialist reconstruction assessment.

Table 7: **Dataset sizes, splits, and sampling proportions.**

Dataset	Collected		Experimental split			Sampling (%)
	Clips	Views	Train	Val.	Test	
In-domain						
SCARED	34	24,046	24	2	8	13.86
SimCol3D	33	37,833	22	2	9	13.25
C3VD	112	43,164	99	2	11	18.07
EndoMapper-Sim	5	1,610	3	1	1	3.01
EndoMapper-Real	2,379	184,796	2,347	4	28	24.10
StereoMIS	354	54,770	346	2	6	13.86
Sano	1,938	223,761	519	63	64	13.86
Out-of-domain						
Hamlyn	22	3,300	0	0	22	0
SCOPE-colon-phantom	14	1,792	0	0	14	0
RealSynCol	9	2,680	0	0	9	0
SCOPE-colon-real	10	1,280	0	0	10	0
Total	**4,910**	**579,032**	**3,360**	**76**	**182**	**100.00**

Algorithm 1 EndoMapper-Real curation and annotation

```
def annotate(videos, held_out):
    dataset = []
    for video in videos:
        segments = vlm_judge(video)
        for clip in filter_and_split(segments):
            cameras = sfm(clip, matcher=MFT)
            if not valid(cameras):
                continue
            alternatives = [sfm(clip, matcher=m)
                            for m in (LoMa, SIFT_LightGlue)]
            rgb = undistort(clip, cameras)
            depth, points = mapanything(rgb, cameras)
            mask = valid_mask(depth, points, cameras)
            status = qc(cameras, alternatives, depth, points, mask)
            if status == REJECT:
                continue
            if video in held_out and status == BORDERLINE:
                continue
            dataset.append((rgb, cameras, depth, points, mask))
    return dataset
```

A.2 EndoMapper-Real Curation and Annotation

We convert raw clinical video into reconstructable clips with camera and dense geometry annotations (Algorithm 1). Gemini 3.5 Flash acts as a VLM judge, assessing tissue visibility, camera travel, obstruction, and image quality with the prompt below.

Clip selection. Frame-level filters refine the VLM-selected segments by removing blur, exposure failures, strong highlights, and frames with insufficient texture or tissue visibility. We then split the retained frames at abrupt viewpoint changes and temporal gaps to form coherent 50–300-frame clips. Source frame indices keep the clips aligned with their annotations.

Geometric annotation. For camera reconstruction in low-texture tissue, MFT tracks provide the primary SfM correspondences (Neoral et al., 2024). Independent reconstructions from LoMa and SIFT with LightGlue allow us to check camera agreement across correspondence sources (Nordström et al., 2026; Lowe, 2004; Lindenberger et al., 2023). MapAnything then uses the recovered cameras and undistorted images to predict dense depth and point clouds (Keetha et al., 2026). Pixel masks exclude unreliable predictions, depth-edge artifacts, and regions without multi-view support. Clip-level geometric checks then determine which annotations to retain.

VLM prompt. We use the following prompt to define the screening criteria and output format.

```
You are segmenting a FULL colonoscopy video into coarse SEMANTIC CHAPTERS
and SCORING each chapter for whether it can be used to 3D-RECONSTRUCT the
colon.  Watch the entire video, then return ONLY a JSON array (no prose).
Each element:
{"start":"M:SS","end":"M:SS","state":"<one of:  outside_body, insertion,
withdrawal_inspection, wall_redout, stationary, cleaning_irrigation,
therapy_tool, unclear>", "scores":{"structure_visibility":0-3,
"camera_travel":0-3,"unobstructed":0-3,"image_quality":0-3}, "reasons":
{"structure_visibility":"<=8 words","camera_travel":"<=8 words",
"unobstructed":"<=8 words","image_quality":"<=8 words"}}
PRINCIPLES for 3D reconstruction:  a segment is reconstructable only if
textured colon structure is in view AND the scope TRANSLATES (producing
parallax).  Brightness is NOT geometry -- specular highlights and the wet
non-Lambertian sheen MOVE WITH the scope and are noise, not structure;
judge structure by the lumen recess and haustral-fold ridges, NOT by how
bright/wet a frame looks.  Pure rotation (no translation) gives zero
baseline = unreconstructable.  A featureless red wall has no trackable
texture even under motion.  The colon deforms (peristalsis / CO2
insufflation), so shorter contiguous segments are preferred.
Rules:  MERGE adjacent same-activity moments into ONE chapter (~5-20,
temporal order, cover the whole timeline).  Score each chapter 0-3 on
FOUR INDEPENDENT axes (do NOT fuse them; do NOT output keep/discard):
structure_visibility -- is the colon LUMEN / haustral folds / textured
mucosa VISIBLE? Judge by GEOMETRY (a DARK lumen recess / forward tunnel,
concentric haustral-fold ridges, submucosal vessels), NOT by brightness,
and INDEPENDENT of travel direction (insertion or withdrawal).  A bright,
wet, sharp frame full of mucosa can still have NO lumen (scope pressed on
the wall) = 0.
0 = 'red-out' / wall-apposition:  scope pressed on / pointed straight at
the mucosa, uniform red-orange, no lumen recess
1 = lumen only glimpsed occasionally; mostly wall
2 = lumen / haustral folds in view for MOST of the segment
3 = clear lumen tunnel with concentric folds in view throughout
camera_travel -- does the scope TRANSLATE along the colon (advancing OR
withdrawing), giving a baseline / parallax?
0 = no net travel:  static, dwelling on one spot, OR retroflexion (tip
angulated ~180 deg to view backward, scope shaft often visible) =
rotation in place -- score 0 EVEN IF the lumen is clearly visible
1 = small wobble, little net advance or withdrawal
2 = clear advance OR withdrawal through the lumen
3 = sweeps a long extent of colon (much new tissue continuously
enters/leaves view)
unobstructed -- free of endoscopic artifacts?  Obstructions:  simethicone
bubbles/foam, bile (yellow-green), residual stool/debris, active water
jet/submersion, blood, instruments (forceps/snare).  NOT obstructions:
thin clear mucus, a light water film, a few sparse bubbles.
0 = fluid / foam / stool / tool covers most of the field
1 = frequent heavy obstruction
2 = minor moisture / a few bubbles / mild bile; tissue still clearly
visible
3 = clean, unobstructed
image_quality -- sharpness and exposure?  NOTE: a DARK deep-lumen recess
is EXPECTED and is the signal -- do NOT penalize darkness from a far
lumen; penalize only true black / washed-white frames and motion blur.
0 = mostly severely blurred or badly over/under-exposed
1 = frequent motion blur or exposure problems
2 = generally sharp and well-exposed, occasional blur
3 = consistently sharp and well-exposed
- Timestamps in M:SS from the start of the video.
```

A.3 StereoMIS Curation and Annotation

For calibrated StereoMIS videos, we use the provided camera poses and estimate depth from stereo pairs. Clip selection follows the same semantic screening and frame filtering as EndoMapper-Real, without requiring camera travel. Non-white-light segments are excluded, and source frame indices keep stereo images, poses, and masks aligned.

We rectify each stereo pair and estimate disparity with Stereo Any Video (SAV) (Jing et al., 2025). Depth follows $d = f_x b / \text{disparity}$, where f_x is the horizontal focal length and b is the stereo baseline. Camera poses from robot kinematics are expressed relative to the first camera in each clip; instrument masks use the rectified image grid. This yields clips with aligned depth and camera annotations.

A.4 Quality Control

Table 8 lists the EndoMapper-Real QC checks and thresholds. A hard-gate failure yields Reject; a soft flag without a hard failure yields Borderline; passing all checks yields Accept. Training uses Accept and Borderline clips, while held-out evaluation uses Accept clips.

Table 8: **Clip-level geometry checks.**

Check	Criterion	Role
Registration	One significant reconstruction; at least 90% of frames registered	Hard gate
Integrity	Readable depth, consistent frame counts, finite cameras	Hard gate
Connectivity	Largest point-cloud component at least 90%	Hard gate
Coverage	Valid depth over at least 20% of clip pixels	Hard gate
Parallax	Median maximum sparse-point parallax at least 1.5°	Soft flag
Camera adoption	Dense-model camera change at most 1.5% translation and 1° rotation	Soft flag
Sparse agreement	Median scale-aligned relative depth discrepancy at most 0.20; scale drift at most 0.25	Soft flag
Outliers	At most 10% of points exceed the mean eight-neighbor distance by two standard deviations	Soft flag
Near-zero baseline	Flag when at least 40% of frames have a translation-induced parallax proxy below 0.004°	Soft flag
Cross-source poses	Reject if at least two alternatives are usable and all have aligned trajectory residuals above 10%	Hard gate
	Flag if only one alternative is usable and its aligned trajectory residual exceeds 10%	Soft flag

Sparse–dense discrepancy is the median absolute relative depth difference following median scale alignment; scale drift is the interquartile range of per-view scale ratios divided by their median. Connectivity uses 26-neighbor voxels of width $1/64$ of the cloud extent between coordinate-wise 2nd and 98th percentiles. Cross-source residuals measure mean camera-center distance after Sim(3) alignment, normalized by the alternative trajectory's bounding-box diagonal, using at least eight shared frames. Unavailable comparisons do not affect the QC label.

Table 9 reports clip and frame counts. Excluding 11 Borderline clips from held-out recordings leaves 2,379 retained clips. Human verification assigned a Reject label to 28 of the 2,379 clips (1.18%): 2 previously labeled Accept and 26 Borderline. Figure 5 shows Borderline cases with single-source camera disagreement or sparse–dense depth inconsistency, and Reject cases with disagreement across both alternatives or insufficient camera registration.

Table 9: **EndoMapper-Real curation counts.** Frame counts precede training conversion.

Stage	Clips	Frames
Raw videos (93 recordings)	–	3,973,969
Clean clips (Stage 1)	3,034	319,727
Accept	1,733	177,358
Borderline	657	70,501
Reject	644	71,868
Retained collection	2,379	246,981

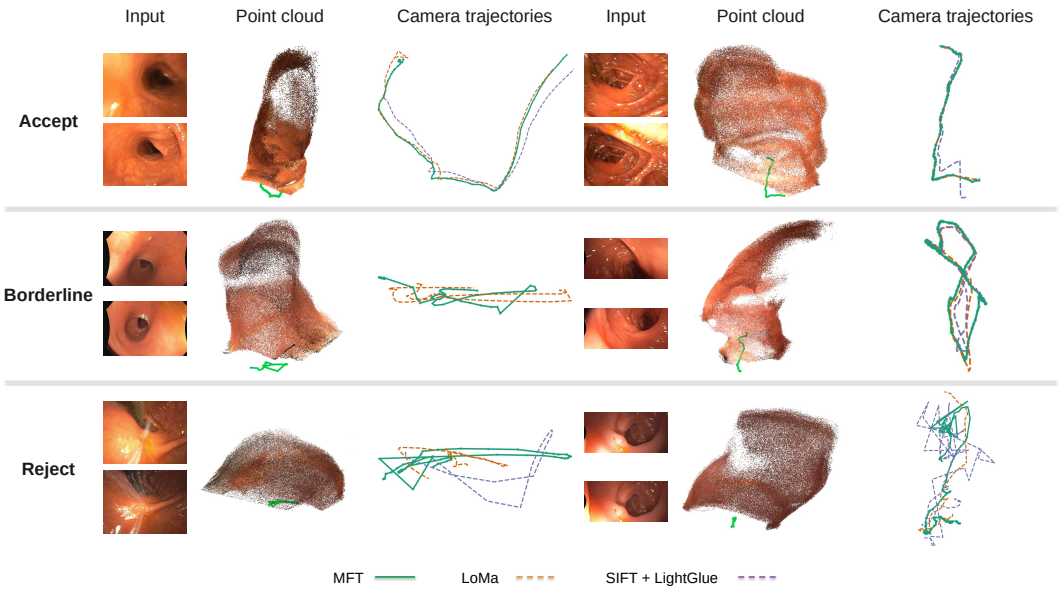

Figure 5: **Quality-control examples.**

A.5 SCOPE-colon Benchmark

SCOPE-colon-phantom contains 14 physical colon phantom clips with 128 evaluated views each. Checkerboard calibration provides intrinsics; a five-degree-of-freedom electromagnetic tracker records camera positions and viewing directions, but not rotation about the viewing axis (roll).

SCOPE-colon-real contains ten clinical colonoscopy clips, with 128 views per clip, for specialist assessment of reconstruction quality. Appendix D.3 describes the study interface and ranking protocol.

Ethical considerations. All data were obtained under institutional review board (IRB) approval. Clinical recordings serve only to evaluate reconstruction quality retrospectively; the study does not alter clinical procedures or patient care. Institutional identifiers are withheld for double-blind review.

B Implementation Details

B.1 Optimization and Trainable Modules

We initialize the geometry backbone from pretrained VGGT-Ω (Wang et al., 2026) and the motion decoder, assignment head, and residual-twist decoder from SM4RT (Lin et al., 2026). The common-motion head has a zero-initialized output layer. We train both geometric SFT and the full model for 29,000 optimizer steps on eight B200 GPUs. We use AdamW with weight decay 0.05, betas $(0.9, 0.95)$, 500 warmup steps, and cosine learning-rate decay. The full model uses a 2×10^{-5} motion-branch learning rate and 10^{-5} for the image encoder and attention backbone, decaying to zero; its geometry heads remain fixed. Geometric SFT updates all parameters, using 10^{-5} for the model and 5×10^{-7} for the patch encoder, with minimum rates of 10^{-7} and 5×10^{-9}, respectively.

Training uses landscape aspect ratios at a nominal 512-pixel resolution and a budget of 24 images per GPU per step.

The full model uses $K = 10$ residual bases, including the zero-residual anchor. The tracking loss weights are $\lambda_{\text{tr}} = 2000$, $\lambda_c = 1$, and $\lambda_q = 0.5$.

Dataset sampling proportions follow Table 7. Geometric SFT uses 3,360 training clips and 76 validation clips; the full model uses 3,349 training clips and 75 validation clips. The full-model training pool excludes nine Sano clips without motion labels and two StereoMIS clips too short for view sampling. Training samples contain 2–24 views, and validation uses 24 views. Geometric SFT uses geometry/covisibility sampling. For Sano and StereoMIS, SCOPE-4D samples ordered views with a target temporal span of 64 source frames; other sources use geometry sampling.

Ablation settings. Data-composition variants use the geometric SFT recipe, each initialized from pretrained VGGT-Ω and trained with the same budget on eight B200 GPUs. Sampling weights are renormalized over the retained datasets. Motion-design variants are trained on StereoMIS for 8,000 steps on four B200 GPUs. The SoM baseline uses sparsemax without an anchor. The anchor variant uses softmax with a zero-motion anchor and static/moving assignment supervision. Full CRM further adds common–residual composition and its corresponding supervision. In this full formulation, the anchor represents zero residual motion: points assigned to it follow the common transform rather than remaining stationary.

B.2 Architecture and Supervision

Features and coordinates. The motion branch reads backbone layers 4, 11, 17, and 23 (zero-indexed) and projects their 2,048-dimensional features to 1,536 dimensions through linear adapters. A motion transformer predicts per-frame twists and temporally shared reference-frame assignments. Subtracting the first-frame common and residual twists from their respective sequences makes both initial transforms identity. Trajectories use first-camera coordinates. Transforms act on annotated reference points during training and predicted reference points at inference. Training coordinates are normalized by the mean reference-track distance, with common translations scaled consistently.

Constructing 3D tracking supervision. For StereoMIS, MFT correspondences are lifted into 3D using stereo depth and camera poses, excluding occlusions and instruments. Sano trajectories fuse rendered optical flow and depth across views using simulator cameras.

We robustly fit a common SE(3) transform to reference-to-target correspondences, using a 3ϵ inlier threshold and rigid refitting; ϵ is a source-specific annotation noise floor.

Let $X_i(t)$ denote an annotated trajectory and $T_c^\star(t)$ the shared common-motion target. The residual amplitude and anchor target are

$$r_i = \text{percentile}_{90,t>0} \left\| X_i(t) - T_c^\star(t) X_i(0) \right\|_2, \tag{8}$$

$$r_i' = \max(r_i - \epsilon, 0), \qquad q_{0,i}^\star = 1 - \frac{r_i'}{r_i' + 2\epsilon}. \tag{9}$$

We compute the percentile over valid observations with fitted common-motion targets. Residuals are measured in millimeters before training normalization, with $\epsilon = 0.03$ mm for Sano and 1.32 mm for StereoMIS. For trajectory weighting, $r_i' > \epsilon$ defines the larger-residual group; the remaining points form the smaller-residual group.

Geometry supervision. We follow the MapAnything training framework for geometry supervision (Keetha et al., 2026). Points, depth, and camera translations are normalized by the mean valid point distance. The losses are

$$\begin{aligned}
\mathcal{L}_{\text{pts}} &= \sum_t \Big[\big\langle c_t\, \rho(\phi(P_t) - \phi(P_t^\star)) - 0.2 \log c_t \big\rangle_{M_t} \\
&\qquad + \big\langle \rho\big(\phi(P_t^{\text{cam}}) - \phi(P_t^{\text{cam},\star})\big) \big\rangle_{M_t}^{\text{trim}} \Big], \\
\mathcal{L}_{\text{cam}} &= \sum_t \Big[\rho(\tau_t - \tau_t^\star) + \min_{s \in \{-1,1\}} \rho(u_t - s u_t^\star) + \big\langle \rho(\mathbf{r}_t - \mathbf{r}_t^\star) \big\rangle_\Omega \Big], \\
\mathcal{L}_{\text{depth}} &= \sum_t \big\langle \rho(\log(1 + D_t) - \log(1 + D_t^\star)) \big\rangle_{M_t}^{\text{trim}}.
\end{aligned} \tag{10}$$

Here P_t and P_t^{cam} are pointmaps in the first-camera and local camera frames; τ_t, u_t, and $\mathbf{r}_t$ denote translation, unit quaternion, and camera ray directions; D_t is camera-axis depth and c_t is predicted confidence. Stars denote supervision targets. Minimizing over quaternion signs handles the two equivalent representations of each rotation.

The point transform and robust penalty are

$$\phi(p) = \frac{\log(1+\|p\|_2)}{\|p\|_2} p, \qquad \rho(e) = \frac{|\alpha-2|}{\alpha}\left[\left(1+\frac{\|e\|_2^2}{\sigma^2|\alpha-2|}\right)^{\alpha/2} - 1\right], \tag{11}$$

with $\phi(0)=0$, $\alpha=0.5$, and $\sigma=0.05$. Angle brackets denote pixel means over the annotation validity mask M_t or image grid Ω. The trimmed mean excludes the largest 5% of valid losses within each real image; synthetic images use the full valid set.

Motion supervision. For predicted and target displacements $\widehat{d}_i(t) = \widehat{X}_i(t) - \widehat{X}_i(0)$ and $d_i(t) = X_i(t) - X_i(0)$, define

$$\ell(\mathcal{A}) = \frac{1}{3|\mathcal{A}|} \sum_{(i,t)\in\mathcal{A}} \|\widehat{d}_i(t) - d_i(t)\|_2^2.$$

The set $\mathcal{A}$ contains observations valid at both the reference and target frames. With the larger- and smaller-residual observation sets $\mathcal{A}_f$ and $\mathcal{A}_b$,

$$\mathcal{L}_{\mathrm{tr}} = \ell(\mathcal{A}_f) + 2.5\ell(\mathcal{A}_b). \tag{12}$$

The common-motion loss compares reference points under predicted and fitted common transforms:

$$\mathcal{L}_c = \frac{1}{|\mathcal{C}|\,|\mathcal{P}|} \sum_{t\in\mathcal{C}} \sum_{i\in\mathcal{P}} \|T_c(t)X_i(0) - T_c^\star(t)X_i(0)\|_2 . \tag{13}$$

Here $\mathcal{C}$ indexes frames with valid common-motion targets, and $\mathcal{P}$ indexes sampled reference points. For valid reference seeds $\mathcal{Q}$, let $a_i = q_0(x_i)$ and $b_i = q^\star_{0,i}$. The assignment loss is

$$\mathcal{L}_q = \frac{1}{|\mathcal{Q}|} \sum_{i\in\mathcal{Q}} \left[b_i \log\frac{b_i}{a_i} + (1-b_i)\log\frac{1-b_i}{1-a_i}\right], \tag{14}$$

where $\mathcal{Q}$ includes only seeds with valid observations and finite targets.

Motion losses are averaged over the full batch, with unlabeled samples contributing zero. The joint objective is scaled by $2/T$ for T input views.

C Proof of the CRM Residual Bound

We bound displacement relative to common motion under the CRM representation.

SE(3) exponential map. The hat operator maps a twist $\xi = (v,\omega) \in \mathbb{R}^6$ to its Lie-algebra matrix:

$$\xi^\wedge = \begin{bmatrix} [\omega]_\times & v \\ \mathbf{0}^\top & 0 \end{bmatrix} \in \mathfrak{se}(3), \qquad \mathrm{Exp}(\xi) := \exp(\xi^\wedge) \in \mathrm{SE}(3). \tag{15}$$

Here $[\omega]_\times p = \omega \times p$, and exp is the matrix exponential.

Claim. Fix a reference point $p = X(x,0)$ and time t. Let $T_c(t) \in \mathrm{SE}(3)$, $q_k(x) \ge 0$, $\sum_{k=0}^{K-1} q_k(x) = 1$, and $\Delta\xi_0(t) = 0$. For finite residual bases $\Delta\xi_k(t) = (v_k(t), \omega_k(t))$, define

$$B(p,t) = \max_{1\le k<K} \left(\|v_k(t)\|_2 + \|\omega_k(t)\|_2\|p\|_2\right).$$

With $\Delta\xi(x,t) = \sum_{k=1}^{K-1} q_k(x)\Delta\xi_k(t)$ and $\widehat{X}(x,t) = T_c(t)\,\mathrm{Exp}(\Delta\xi(x,t))p$,

$$\|\widehat{X}(x,t) - T_c(t)p\|_2 \le (1-q_0(x))B(p,t).$$

Proof. We use the exact SE(3) exponential map defined above. For a twist $\zeta = (v,\omega)$, let $A = [\omega]_\times$. Its action is

$$\mathrm{Exp}(\zeta)p = e^A p + \int_0^1 e^{sA} v\, ds.$$

Since $e^A - I = \int_0^1 e^{sA} A\, ds$ and e^{sA} is a rotation,

$$\begin{aligned} \|\operatorname{Exp}(\zeta)p - p\|_2 &= \left\| \int_0^1 e^{sA}(Ap + v)\, ds \right\|_2 \\ &\le \|Ap\|_2 + \|v\|_2 \le \|\omega\|_2 \|p\|_2 + \|v\|_2. \end{aligned}$$

The common transform $T_c(t)$ preserves distances. Applying the triangle inequality and $\sum_{k>0} q_k(x) = 1 - q_0(x)$ gives

$$\begin{aligned} \|\widehat{X}(x,t) - T_c(t)p\|_2 &= \|\operatorname{Exp}(\Delta\xi(x,t))p - p\|_2 \\ &\le \sum_{k>0} q_k(x)\big(\|v_k(t)\|_2 + \|\omega_k(t)\|_2 \|p\|_2\big) \\ &\le (1 - q_0(x))B(p,t). \end{aligned} \tag{16}$$

For fixed residual bases, displacement relative to common motion vanishes as $q_0(x) \to 1$: the point follows $T_c(t)p$. A zero mixed twist in SoM instead suppresses all point motion.

D Evaluation Protocols

D.1 Compared Methods

Table 10 summarizes the compared methods. D4RT* denotes the open-source reproduction. A single resolution value specifies the target long edge before cropping; paired values specify width × height. Test-time optimization refers to geometric fitting: CUT3R fits focal length; Endo3R fits focal length and estimates camera poses with PnP; D4RT uses closed-form camera fitting from predicted points.

D.2 In-Domain Protocol

We follow Wang et al. (2026); Guo et al. (2025) in evaluating depth with absolute relative error (AbsRel) and $\delta_{1.25}$. After sequence-level median scale alignment, AbsRel measures mean absolute depth error divided by reference depth; $\delta_{1.25}$ reports the percentage of predictions within a factor of 1.25 of the reference.

For camera poses, we follow Wang et al. (2025a; 2026) in using pose AUC at 5° and 30°. Pose AUC is the normalized area under the recall curve for the larger of the rotation and translation-direction angular errors of each image pair. Following Keetha et al. (2026), we report absolute trajectory error (ATE) as the mean camera-position error after scale normalization and rigid alignment. Relative translation error (RTE) and relative rotation error (RRE) assess motion between adjacent frames, as in Guo et al. (2025). We report mean translation-vector error in millimeters and mean rotation-angle error in degrees, respectively.

For 3D tracking, we follow Lin et al. (2026) in reporting endpoint error (EPE) and average percent of points within delta (APD). After sequence-level scale alignment, EPE measures the Euclidean distance between predicted and reference 3D positions. APD averages the percentages of predictions within 1, 2, 5, and 10 mm of their references. We average scores over sequences within each dataset and compute the ID geometry summary over the seven datasets.

D.3 Out-of-Domain Protocol

On Hamlyn, we follow Guo et al. (2025) in evaluating depth with AbsRel and $\delta_{1.25}$. Instrument masks are generated with SAM 2.1 Hiera-Large (Ravi et al., 2024), using box and point prompts, bidirectional video propagation, and visual review with prompt refinement. The same source masks are mapped to each method's evaluation grid; sequence-level median scale alignment and depth metrics use the intersection of valid reference depth and non-instrument pixels. Because camera-pose ground truth is unavailable, we evaluate camera stability on clips with a stationary camera. ATE and absolute rotation error (ARE) measure translation and rotation variation relative to the first predicted pose.

On *SCOPE-colon-phantom*, camera estimates are evaluated against electromagnetic tracker measurements. ATE measures camera-position root mean square error (RMSE) after Sim(3) alignment, and

Table 10: **Compared methods.**

Method	Input	Output	Resolution	Supplied intrinsics	Test-time optimization
General 3D					
VGGT (Wang et al., 2025a)	RGB video	Depth, cameras, points	518	×	×
VGGT-Ω (Wang et al., 2026)	RGB video	Depth, cameras, points	512	×	×
MapAnything (Keetha et al., 2026)	RGB video	Depth, cameras, points	518	×	×
CUT3R (Wang et al., 2025b)	RGB video	Depth, cameras, points	512	×	✓
General 4D					
SpatialTrackerV2 (Xiao et al., 2025)	RGB video	Depth, cameras, 3D tracks	518	×	×
4RC (Luo et al., 2026)	RGB video	Depth, cameras, 3D tracks	512	×	×
Any4D (Karhade et al., 2025)	RGB video	Depth, cameras, 3D tracks	518	×	×
SM4RT (Lin et al., 2026)	RGB video	Depth, cameras, 3D tracks	518	×	×
D4RT* (Zhang et al., 2026)	RGB video	Depth, cameras, 3D tracks	256 × 256	×	✓
Endoscopy					
AF-SfMLearner (Shao et al., 2022)	RGB video	Depth, camera poses	320 × 256	✓	×
EndoDAC (Cui et al., 2024)	RGB video	Depth, camera poses	320 × 256	×	×
Endo-FASt3R (Sheikh Zeinoddin et al., 2025)	RGB video	Depth, camera poses	320 × 256 (depth) 512 (pose)	✓	×
Endo3R (Guo et al., 2025)	RGB video	Depth, cameras, points	320	×	✓

RTE measures one-second relative translation RMSE. ARE* measures mean viewing-direction error after global direction alignment. The asterisk indicates that the 5-DoF tracker does not record roll.

For long RealSynCol sequences, we follow Deng et al. (2025) in reporting ATE as camera-position RMSE after Sim(3) alignment. Adjacent-frame RTE and RRE are translation-vector RMSE in millimeters and rotation-angle RMSE in degrees, respectively.

Real-colonoscopy user study. All four methods in Table 5 receive the same 128 input frames from each of ten clips. Case and method order are randomized per participant. Rankings run from 1 (best) to 4 (worst), without ties. Participants assess camera motion consistent with the video, coherent colon-wall surfaces, and temporally stable depth with clear fold and lumen boundaries; path length, surface coverage, and absolute depth colors are not criteria. Good/bad examples and an unscored phantom practice view precede the study; examples remain accessible during ranking.

Instrument masks use the same SAM 2.1-assisted procedure as Hamlyn and are shared across methods: RGB masks are overlaid, depth regions are grayed out, and corresponding 3D points are removed before sampling. Figure 6 shows the synchronized comparison interface. The six clinical and biomedical experts completed 180 four-method ranking tasks in total (720 individual ranks), without missing responses or exclusions. Method identities were hidden.

Table 5 reports mean ranks with 95% percentile bootstrap confidence intervals from 50,000 resamples of participants and clips separately, preserving method pairing. Mean per-clip Kendall's W is 0.780, 0.728, and 0.698 for camera trajectories, point clouds, and depth, respectively. Statistical analyses were specified after collection. Exploratory two-sided exact sign tests on participant-level mean rank differences do not reach the 0.05 level after Holm correction across the nine method–dimension comparisons. Given the small study sample, we interpret the results as sample-specific visual preferences rather than population-level superiority.

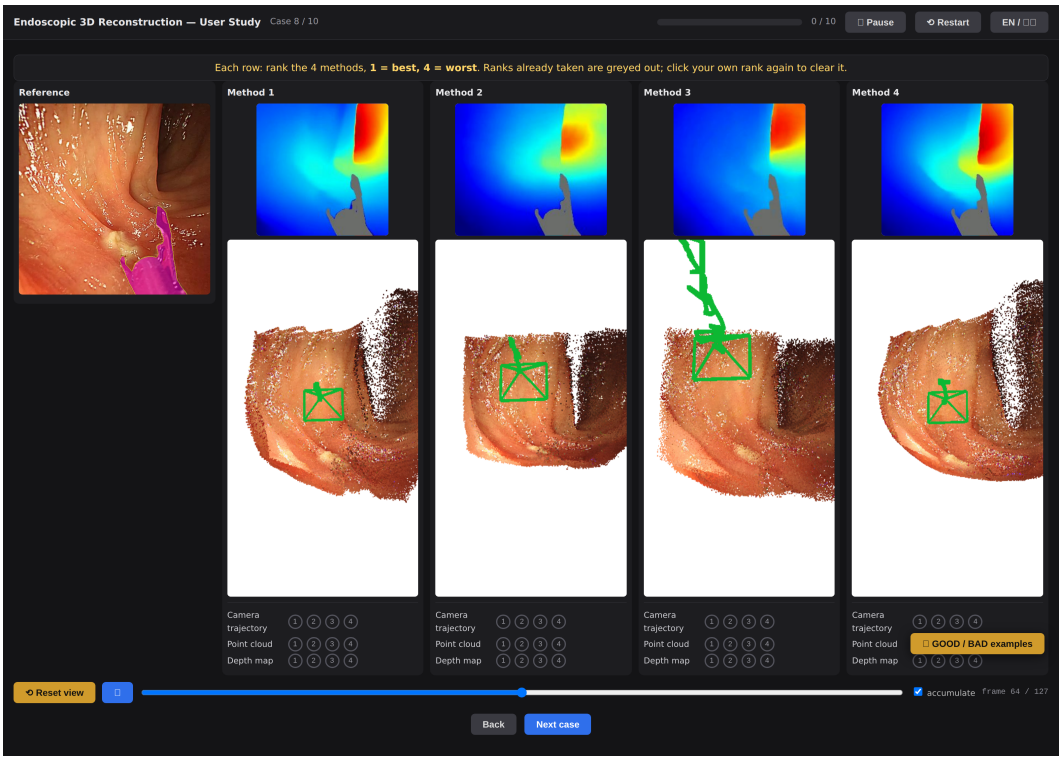

Figure 6: **User-study interface.** Reference video, masked depth, interactive point clouds, and ranking controls for four anonymous methods. No rankings are selected.

E Additional Results

E.1 Additional Quantitative Results

Geometry. We report per-dataset camera pose and depth results on SCARED, SimCol3D, C3VD, EndoMapper-Sim, and EndoMapper-Real. Sano and StereoMIS results are reported in Table 2.

Table 11: **Camera pose and depth estimation on SCARED.**

	Pose			Depth	
Method	AUC@5° ↑	AUC@30° ↑	ATE↓	AbsRel↓	$\delta_{1.25}$ ↑
VGGT	10.31	64.78	0.047	0.047	99.15
VGGT-Ω	17.87	76.51	0.030	0.038	99.36
MapAnything	6.10	56.83	0.050	0.071	95.00
CUT3R	0.35	11.80	0.174	0.135	81.43
SpatialTrackerV2	8.65	68.50	0.049	0.038	99.35
4RC	16.61	77.25	0.032	0.035	<u>99.41</u>
Any4D	0.49	20.43	0.084	0.074	97.45
SM4RT	17.43	76.49	0.060	<u>0.034</u>	**99.47**
Endo3R	0.17	16.07	0.161	0.100	89.38
AF-SfMLearner	0.28	16.18	0.125	0.119	87.59
EndoDAC	0.28	17.22	0.121	0.121	85.79
Endo-FASt3R	14.82	76.52	0.047	0.117	88.16
SCOPE-4D$_{\text{SFT}}$ (Ours)	**29.92**	**83.12**	<u>0.017</u>	<u>0.034</u>	99.15
SCOPE-4D (Ours)	<u>27.89</u>	<u>82.88</u>	**0.016**	**0.033**	99.21

Table 12: **Camera pose and depth estimation on SimCol3D.**

Method	Pose			Depth	
	AUC@5° ↑	AUC@30° ↑	ATE↓	AbsRel↓	$\delta_{1.25}$ ↑
VGGT	0.44	10.82	0.288	0.181	68.03
VGGT-Ω	6.69	52.96	0.072	0.093	92.77
MapAnything	0.22	7.39	0.258	0.172	71.11
CUT3R	0.09	11.28	0.223	0.459	32.57
SpatialTrackerV2	0.79	35.24	0.105	0.155	75.69
4RC	1.19	28.02	0.163	0.153	76.02
Any4D	0.06	3.84	0.227	0.297	43.55
SM4RT	0.95	22.27	0.197	0.165	71.88
Endo3R	0.08	7.03	0.202	0.265	54.12
AF-SfMLearner	0.67	14.14	0.275	0.306	44.98
EndoDAC	0.57	14.55	0.257	0.242	55.23
Endo-FASt3R	0.79	21.50	0.130	0.255	53.13
SCOPE-4D$_{\mathbf{SFT}}$ (Ours)	<u>54.81</u>	<u>91.20</u>	**0.009**	<u>0.025</u>	<u>99.90</u>
SCOPE-4D (Ours)	**59.70**	**91.98**	<u>0.010</u>	**0.023**	**99.91**

Table 13: **Camera pose and depth estimation on C3VD.**

Method	Pose			Depth	
	AUC@5° ↑	AUC@30° ↑	ATE↓	AbsRel↓	$\delta_{1.25}$ ↑
VGGT	7.37	50.19	0.097	0.196	65.00
VGGT-Ω	18.54	70.34	0.042	0.062	97.05
MapAnything	9.37	60.34	0.056	0.144	78.64
CUT3R	0.51	22.06	0.138	0.406	38.45
SpatialTrackerV2	25.32	77.26	<u>0.031</u>	0.076	94.10
4RC	16.40	66.05	0.050	0.117	86.05
Any4D	0.92	27.01	0.178	0.270	51.13
SM4RT	7.68	57.45	0.103	0.152	78.62
Endo3R	0.57	19.60	0.229	0.153	76.45
AF-SfMLearner	8.06	55.69	0.130	0.257	51.23
EndoDAC	7.71	45.12	0.138	0.217	61.30
Endo-FASt3R	11.11	58.56	0.069	0.246	54.69
SCOPE-4D$_{\mathbf{SFT}}$ (Ours)	<u>66.31</u>	<u>90.66</u>	**0.009**	<u>0.031</u>	<u>99.35</u>
SCOPE-4D (Ours)	**67.35**	**90.82**	**0.009**	**0.028**	**99.40**

Table 14: **Camera pose and depth estimation on EndoMapper-Sim.**

Method	Pose			Depth	
	AUC@5° ↑	AUC@30° ↑	ATE↓	AbsRel↓	$\delta_{1.25}$ ↑
VGGT	0.05	19.17	0.228	0.153	79.49
VGGT-Ω	1.86	47.59	0.067	<u>0.062</u>	99.51
MapAnything	0.04	8.53	0.265	0.088	94.00
CUT3R	0.03	21.75	0.120	0.301	49.62
SpatialTrackerV2	0.42	43.41	0.060	0.077	97.55
4RC	0.54	36.93	0.080	0.143	77.74
Any4D	0.03	3.26	0.301	0.146	76.16
SM4RT	0.44	32.78	0.110	0.157	73.17
Endo3R	0.01	9.12	0.240	0.151	76.05
AF-SfMLearner	0.03	11.03	0.120	0.305	44.80
EndoDAC	0.03	11.31	0.108	0.182	68.80
Endo-FASt3R	0.04	14.77	0.051	0.241	54.13
SCOPE-4D$_{\mathbf{SFT}}$ (Ours)	**15.47**	**68.49**	**0.032**	**0.024**	<u>99.83</u>
SCOPE-4D (Ours)	<u>14.06</u>	<u>65.30</u>	<u>0.037</u>	**0.024**	**99.84**

Table 15: **Camera pose and depth estimation on EndoMapper-Real.**

Method	Pose			Depth	
	AUC@5° ↑	AUC@30° ↑	ATE↓	AbsRel↓	$\delta_{1.25}$ ↑
VGGT	1.41	31.43	0.155	0.214	61.56
VGGT-Ω	3.22	49.08	0.089	0.157	76.44
MapAnything	1.49	27.23	0.156	0.274	81.11
CUT3R	0.20	17.54	0.190	0.408	37.70
SpatialTrackerV2	3.03	54.16	0.097	0.194	68.16
4RC	2.16	44.34	0.109	0.219	62.20
Any4D	0.16	12.38	0.256	0.388	39.70
SM4RT	1.58	43.43	0.106	0.224	61.41
Endo3R	0.11	9.92	0.274	0.247	54.90
AF-SfMLearner	0.94	34.59	0.204	0.372	37.38
EndoDAC	1.19	35.16	0.200	0.306	45.84
Endo-FASt3R	1.21	39.23	0.093	0.358	38.99
SCOPE-4D$_{\mathbf{SFT}}$ (Ours)	**23.59**	**77.61**	**0.032**	0.095	92.94
SCOPE-4D (Ours)	22.80	77.35	0.033	**0.091**	**93.89**

3D tracking. Table 16 reports APD at each distance threshold for Sano and StereoMIS, supplementing the mean APD in Table 2.

Table 16: **3D tracking APD at multiple thresholds.**

Method	Sano				StereoMIS			
	APD@1↑	APD@2↑	APD@5↑	APD@10↑	APD@1↑	APD@2↑	APD@5↑	APD@10↑
D4RT*	33.99	66.11	92.98	98.69	4.29	10.56	42.99	78.63
SpatialTrackerV2	77.81	88.97	96.50	99.03	6.87	18.56	50.80	79.70
4RC	86.91	94.26	98.08	99.63	12.25	26.19	56.17	82.50
Any4D	87.29	93.65	97.35	99.34	12.02	25.08	54.46	78.25
SM4RT	89.48	95.28	98.17	99.62	12.20	25.26	56.14	82.21
SCOPE-4D (Ours)	**90.51**	**95.81**	**98.42**	**99.72**	**19.09**	**39.53**	**66.25**	**87.44**

Motion ablation. Table 17 expands Table 3(b) with threshold-wise APD.

Table 17: **Motion ablation on StereoMIS.**

Variant	APD@1↑	APD@2↑	APD@5↑	APD@10↑
SoM (baseline)	11.22	28.89	63.67	82.55
w/ zero-motion anchor	12.73	33.55	66.33	**87.83**
Full CRM	**19.02**	**37.30**	**66.89**	87.30

E.2 Additional Qualitative Results

The following figures compare reconstructed geometry and depth across datasets. Geometry panels use the same tissue-centered crop for all methods within a case; distant camera frusta may lie outside the displayed region.

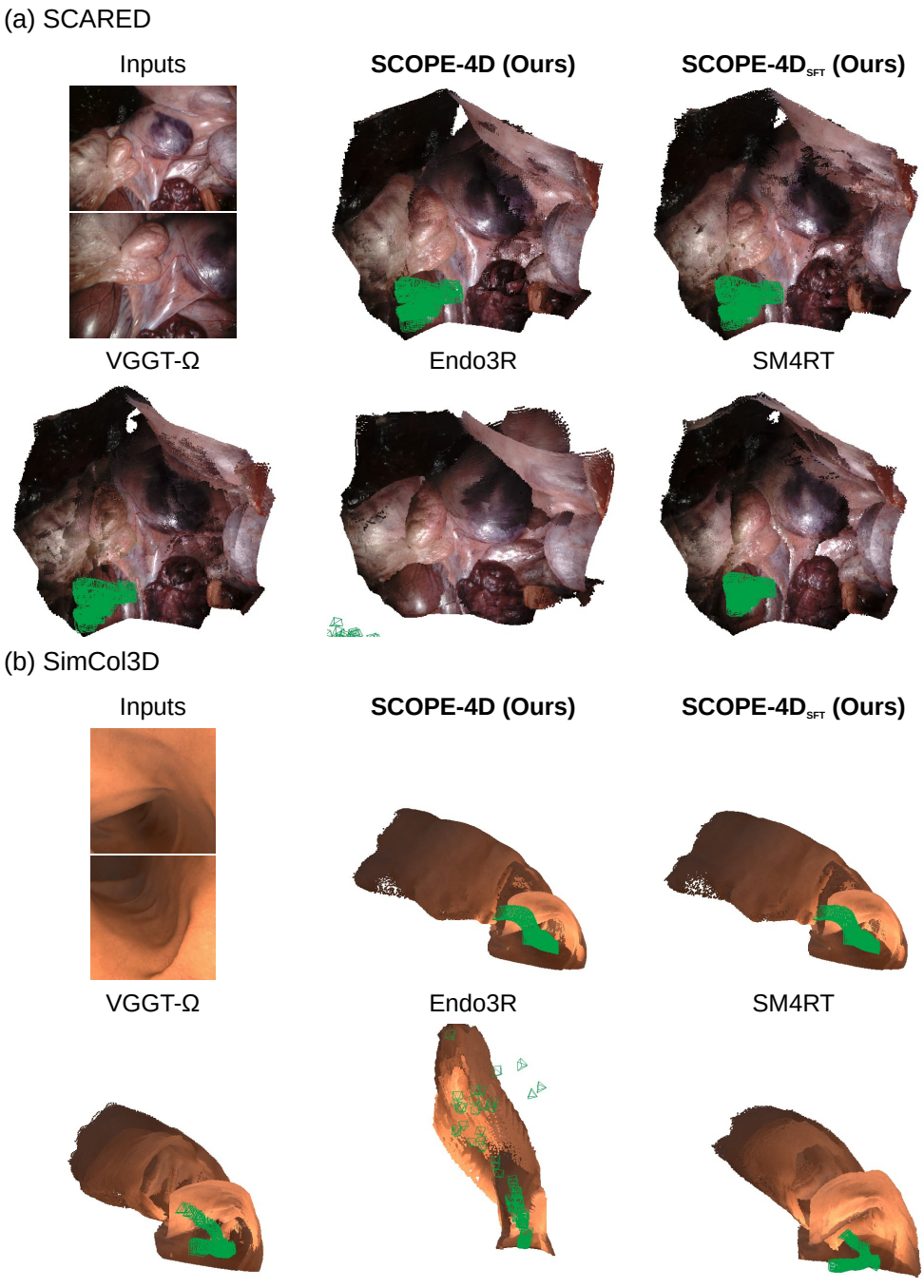

Figure 7: **Additional 3D geometry comparisons.**

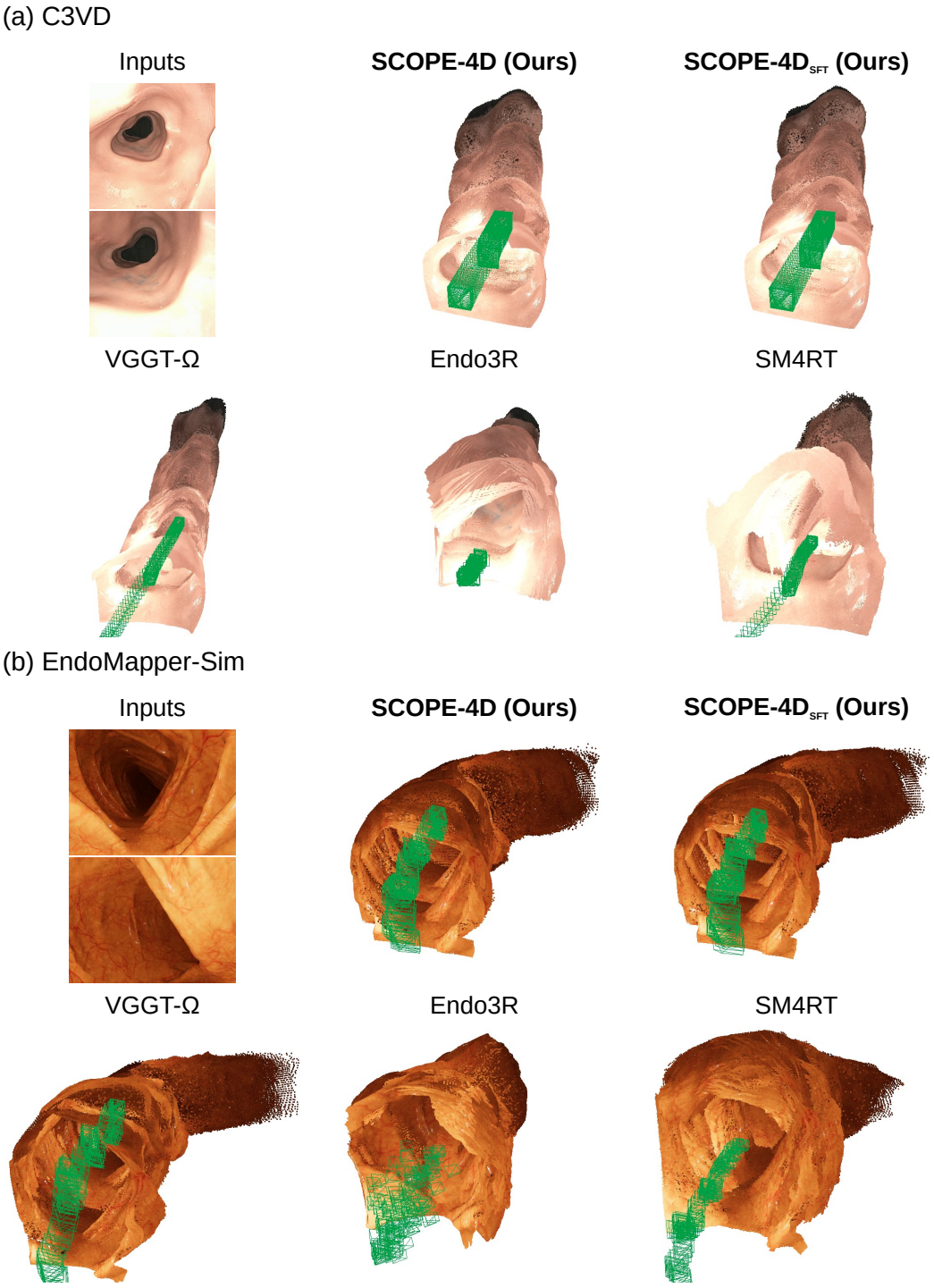

Figure 8: **Additional 3D geometry comparisons.**

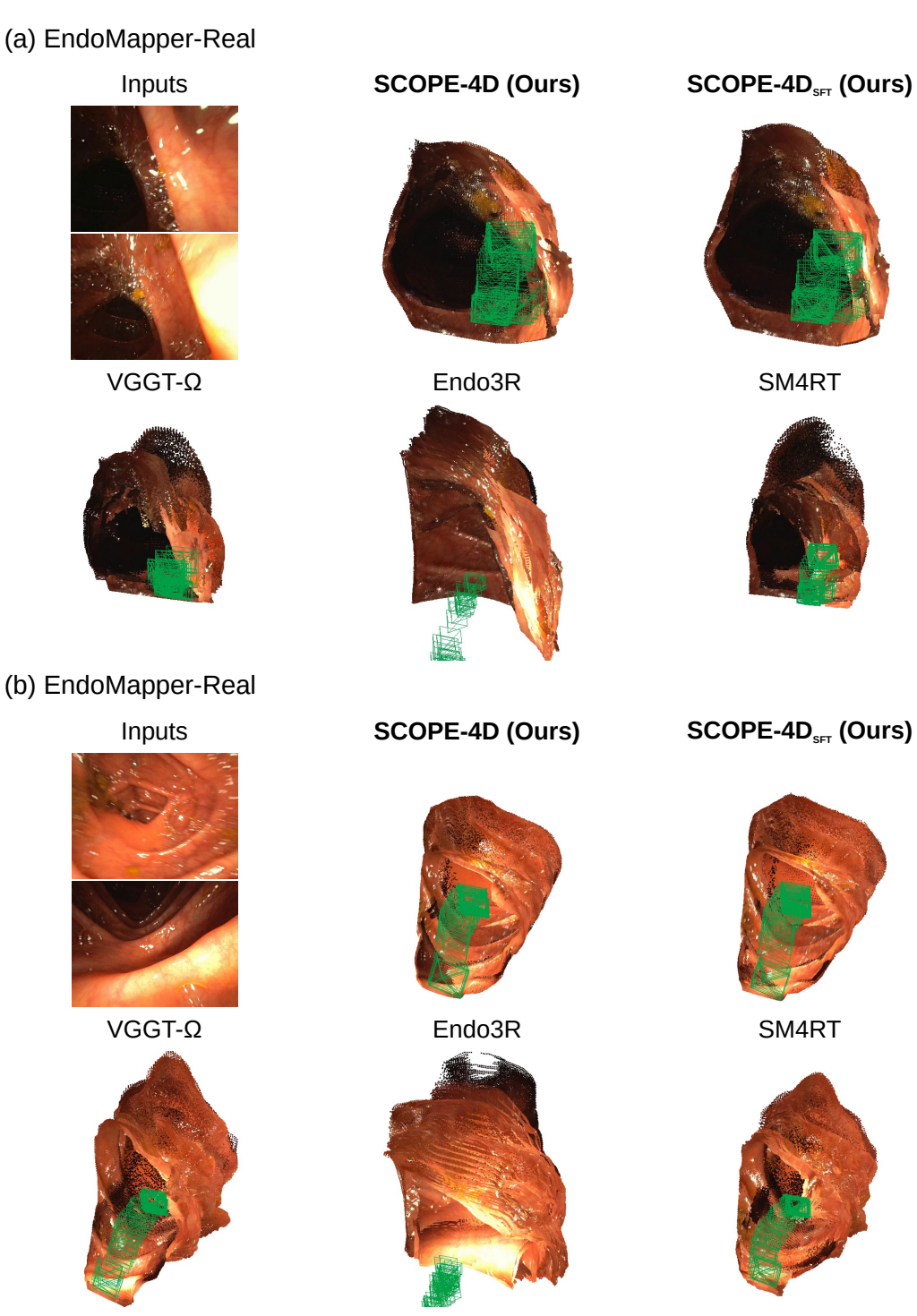

Figure 9: **Additional 3D geometry comparisons.**

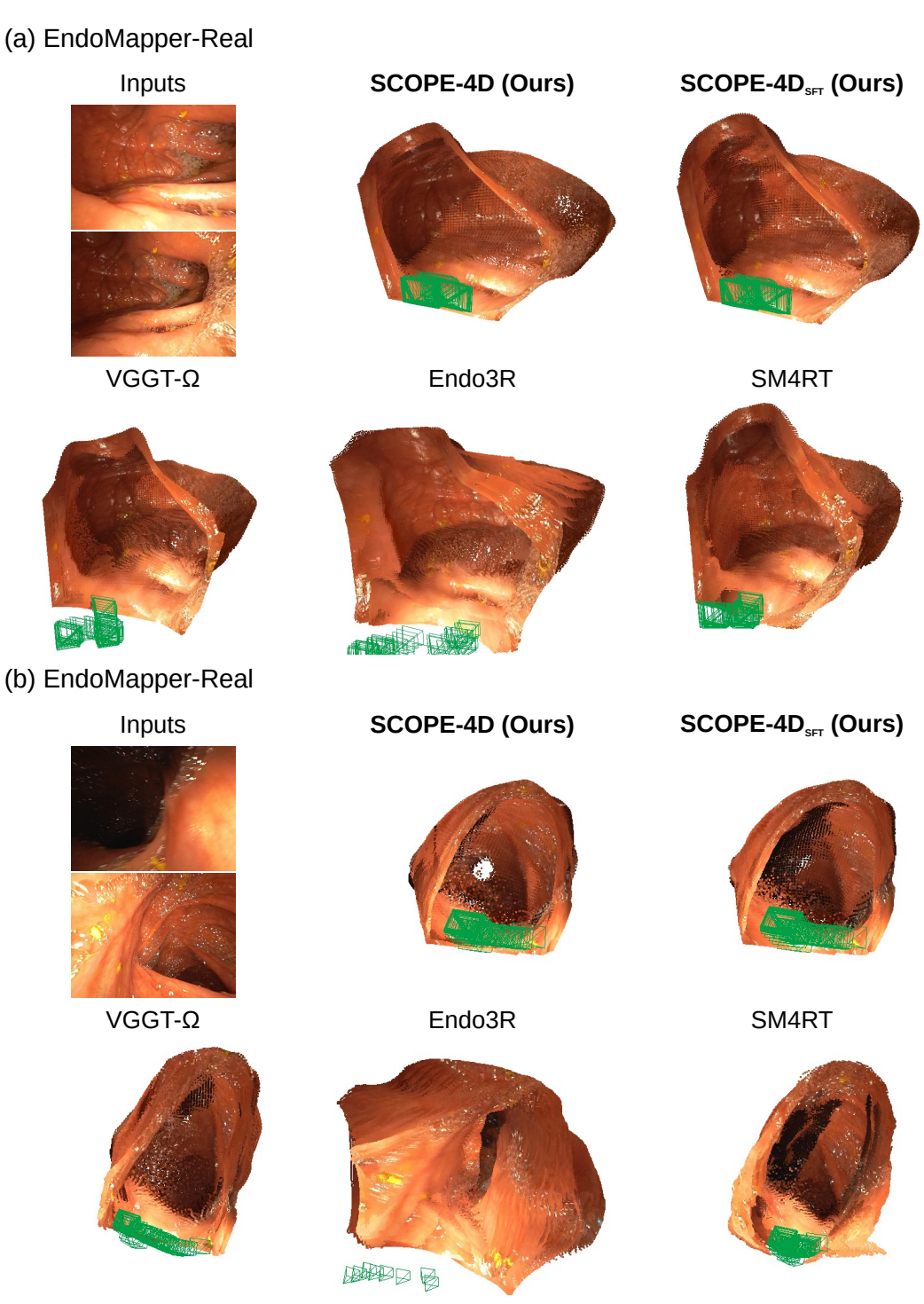

Figure 10: **Additional 3D geometry comparisons.**

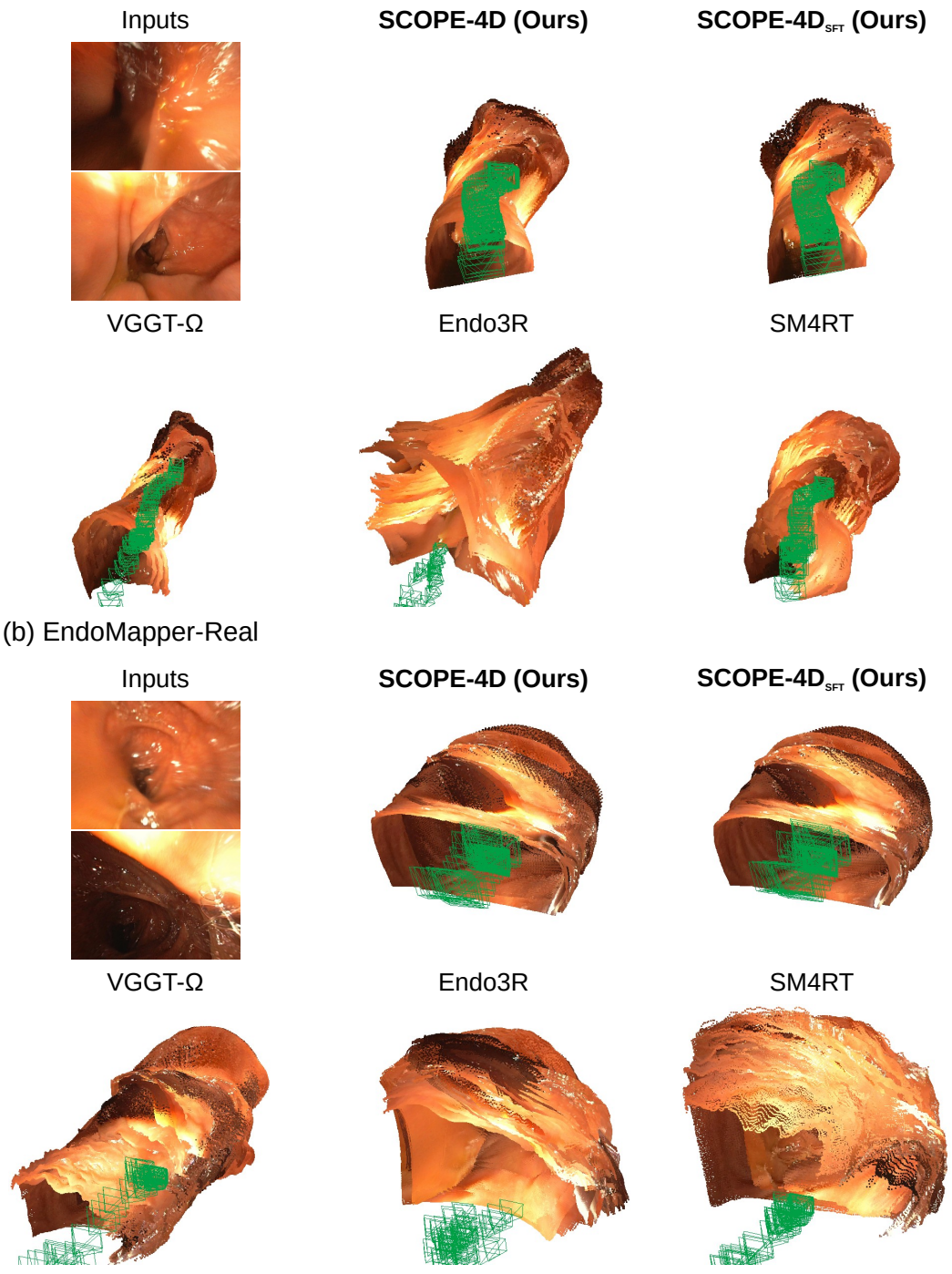

Figure 11: **Additional 3D geometry comparisons.**

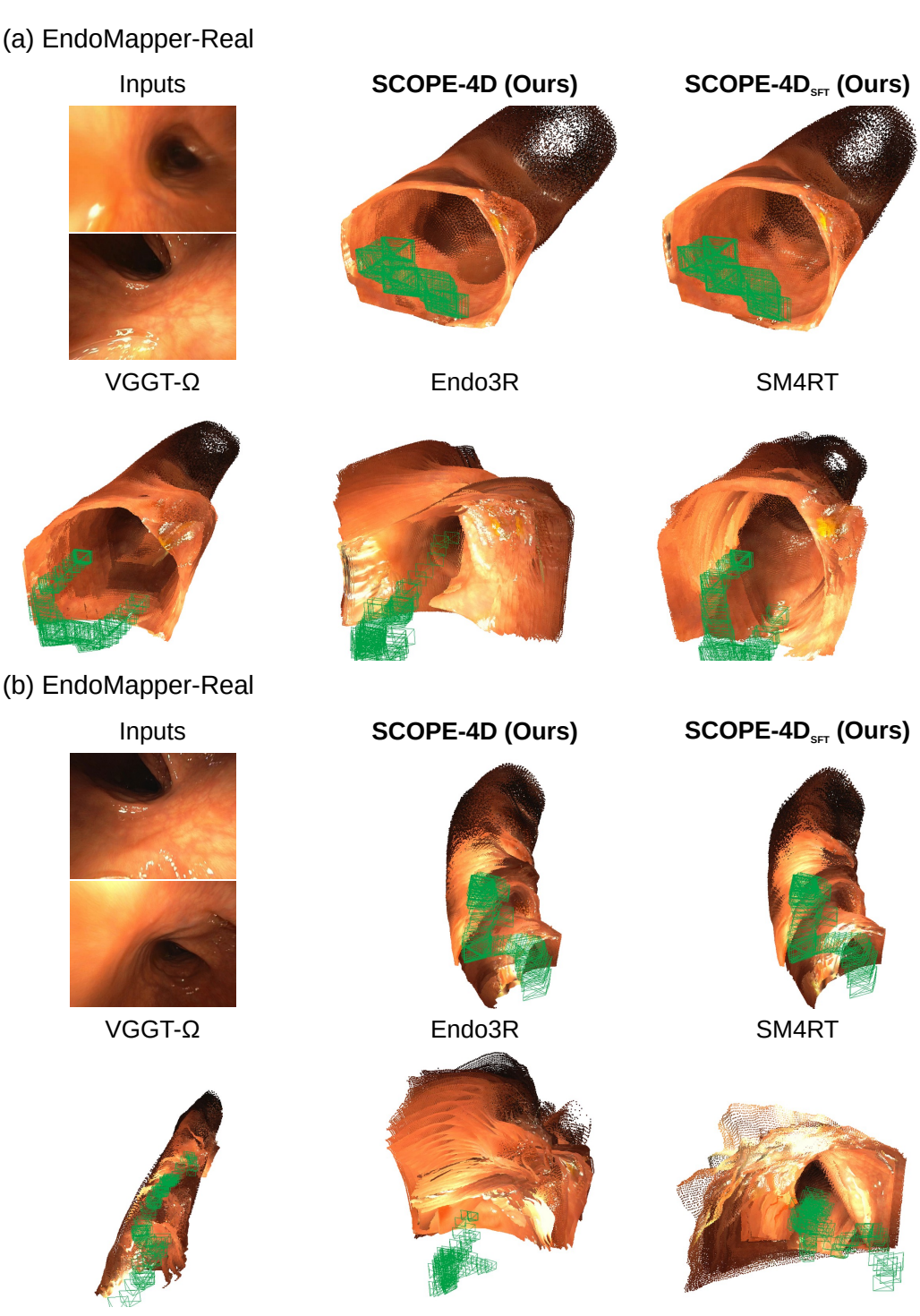

Figure 12: **Additional 3D geometry comparisons.**

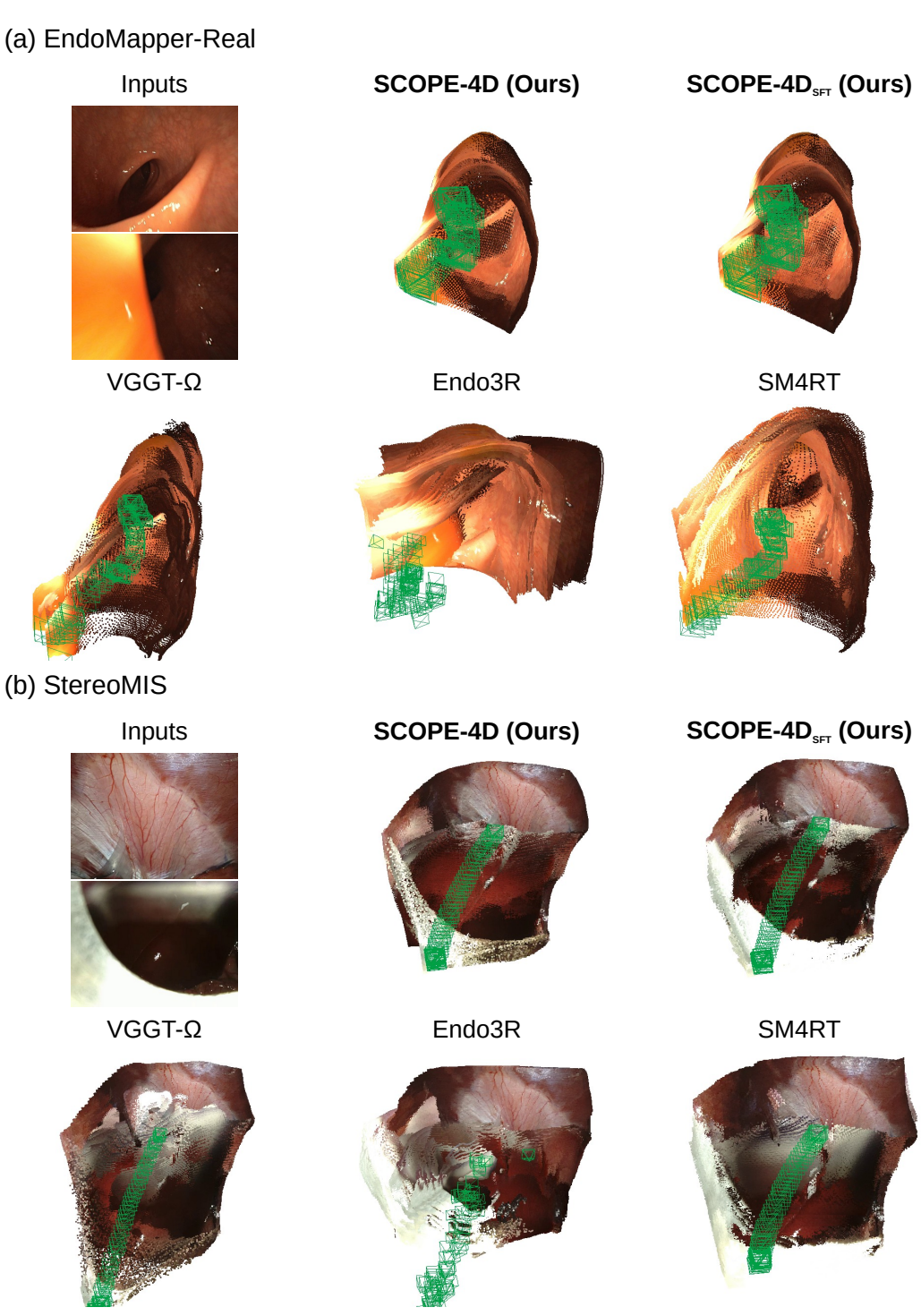

Figure 13: **Additional 3D geometry comparisons.**

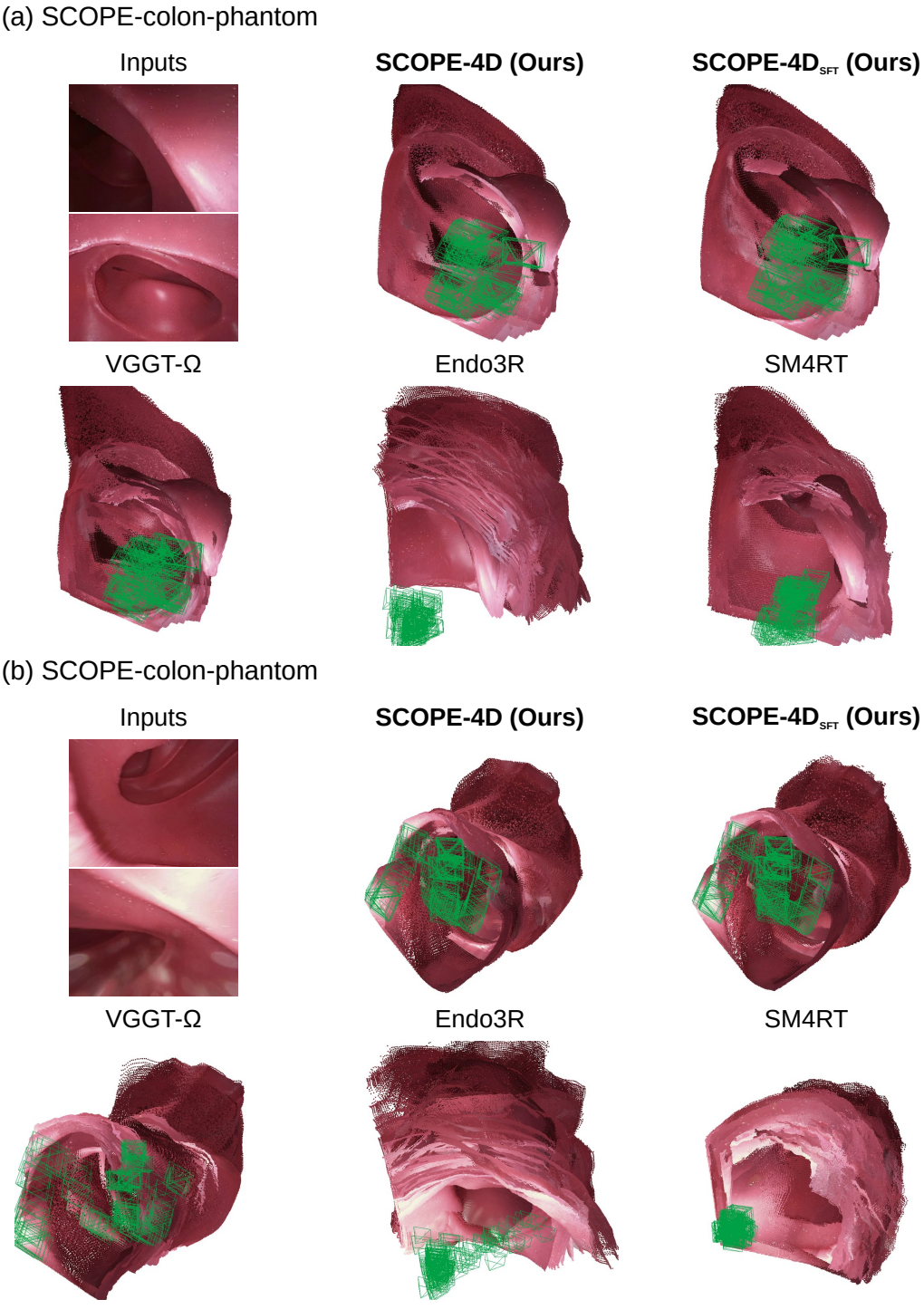

Figure 14: **Additional 3D geometry comparisons.**

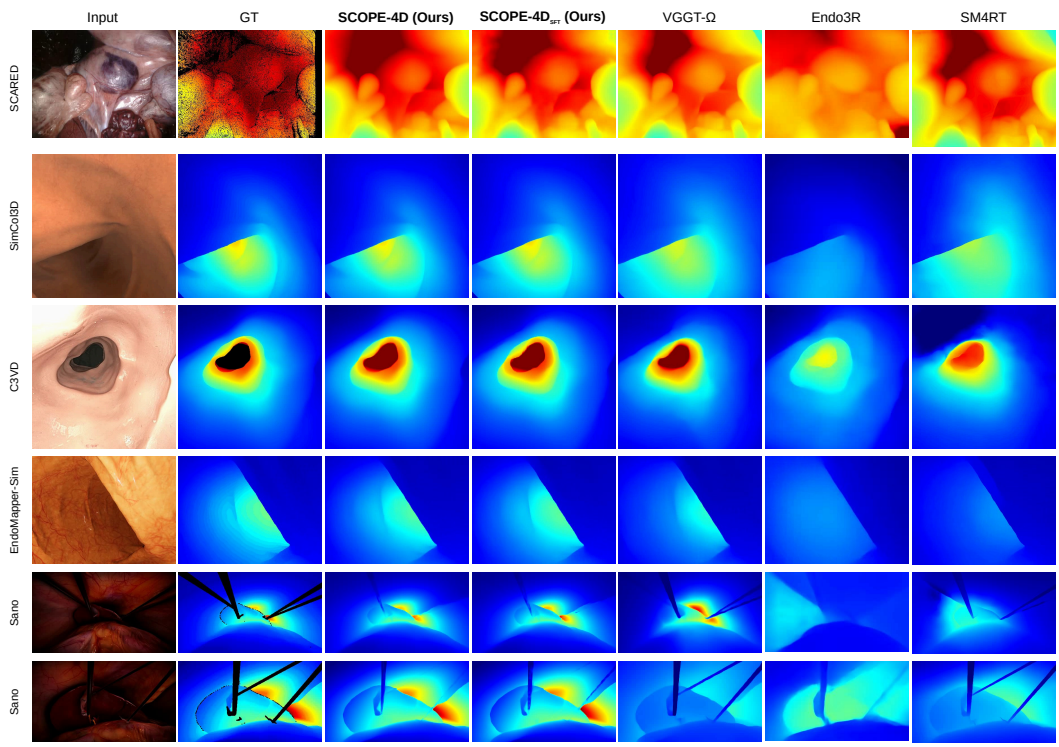

Figure 15: **Additional depth comparisons.**

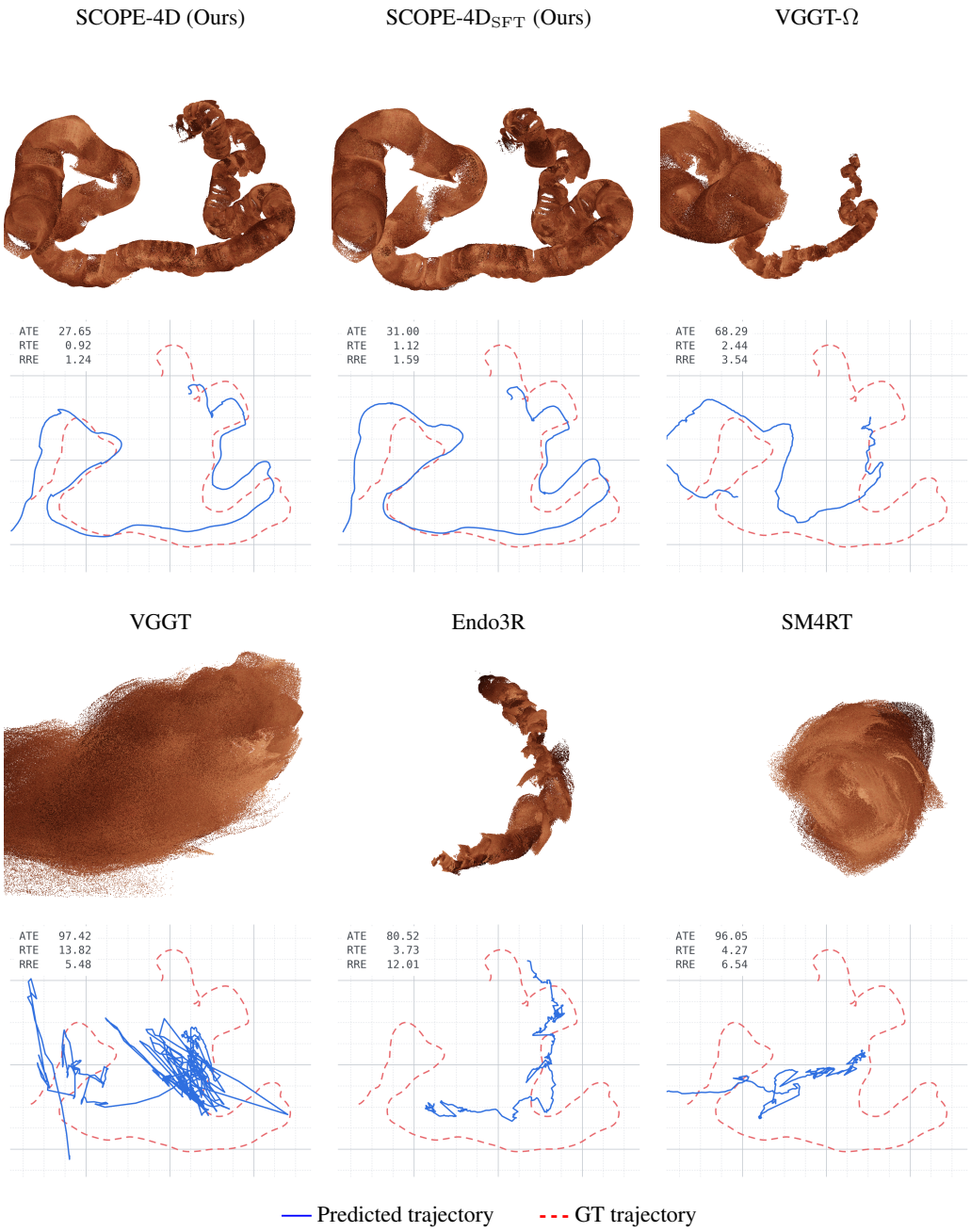

Figure 16: **Long-sequence reconstruction on RealSynCol Colon 2.**