On the Role of Depth in Surgical Vision Foundation Models: An Empirical Study of RGB-D Pre-training

Anonymous authors
Paper under double-blind review

Abstract

Vision foundation models (VFMs) have emerged as powerful tools for surgical scene understanding. However, current approaches predominantly rely on unimodal RGB pre-training, overlooking the complex 3D geometry inherent to surgical environments. Although several architectures support multimodal or geometry-aware inputs in general computer vision, the benefits of incorporating depth information in surgical settings remain underexplored. We conduct a large-scale empirical study comparing eight ViT-based VFMs that differ in pre-training domain, learning objective, and input modality (RGB vs. RGB-D). For pre-training, we use a curated dataset of 1.4 million robotic surgical images paired with depth maps generated from an off-the-shelf network. We evaluate these models under both frozen-backbone and end-to-end fine-tuning protocols across eight surgical datasets spanning object detection, segmentation, depth estimation, and pose estimation. Our experiments yield several consistent findings. Models incorporating explicit geometric tokenization, such as MultiMAE, substantially outperform unimodal baselines across all tasks. Notably, geometric-aware pre-training enables remarkable data efficiency: models fine-tuned on just 25% of labeled data consistently surpass RGB-only models trained on the full dataset. Importantly, these gains require no architectural or runtime changes at inference; depth is used only during pre-training, making adoption straightforward. These findings suggest that multimodal pre-training offers a viable path towards building more capable surgical vision systems.

1 Introduction

In recent years, artificial intelligence (AI) has become a driving force in computer-assisted interventions. Data-driven methods have shown remarkable success in computer vision tasks such as surgical scene segmentation (Allan et al., 2020), action recognition (Bian et al., 2024), object detection (Nwoye et al., 2025), and tissue tracking (Schmidt et al., 2024). In the context of robot-assisted minimally invasive surgery (RAMIS), visual and spatial intelligence can enable downstream capabilities such as 3D reconstruction (Wang et al., 2022; Xu et al., 2024; Schmidt et al., 2024), language interfaces (Perez et al., 2025; Honarmand et al., 2024), and autonomy (Da Col et al., 2021; Kim et al., 2024; 2025; Acar et al., 2025). Consequently, developing models with robust visual representations that capture both the semantic and geometric nature of the surgical scene is a necessary step towards intelligent RAMIS systems.

In the computer vision community, vision foundation models (VFMs) have gained significant traction because of their scalability and generalization capabilities. Through large-scale pre-training, VFMs learn transferable representations that perform well across diverse downstream tasks. Self-supervised learning (SSL) has emerged as the dominant pre-training paradigm, relying on pretext tasks rather than explicit labels. This property is particularly important in surgical computer vision, where large-scale expert annotation is costly and difficult to obtain. Motivated by this challenge, recent efforts have focused on compiling large surgical video datasets for VFM development (Hirsch et al., 2023; Wang et al., 2023; Batić et al., 2024; Schmidgall et al., 2024; Jaspers et al., 2025; Yang et al., 2025), producing promising results in global image-level tasks such as phase recognition (Ramesh et al., 2023; Hirsch et al., 2023) and triplet recognition (Batić et al., 2024; Yang et al., 2025).

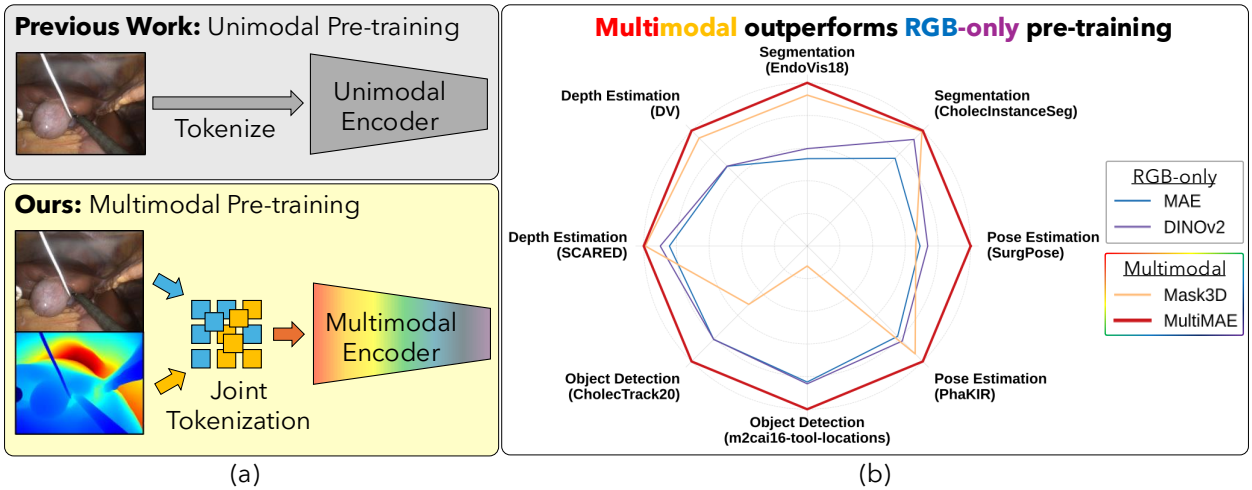

Figure 1: An overview of our contribution. (a) Previous work in surgical VFMs typically use unimodal vision-only pre-training schemes (e.g., DINO, MAE), which limit the geometric/3D understanding of the models. In contrast, we use models which explicitly incorporate geometry into the pre-text task (e.g., Mask3D, MultiMAE), thus improving spatial and semantic understanding of surgical scenes. (b) We fine-tune these models to demonstrate the superiority of multimodal over unimodal pre-training across various geometric, semantic, and dense pixel downstream tasks. All values are normalized; please see Table 11 for numerical results.

However, surgical VFM pre-training remains predominantly unimodal, relying solely on RGB images. While such representations are often sufficient for image-level tasks such as tool presence detection, they frequently struggle with fine-grained spatial and geometric tasks. In other words, existing models may excel at answering "what is happening" but lack spatial 3D scene understanding for dense pixel tasks. In contrast, previous research has shown that incorporating additional modalities during pre-training, such as depth maps, can substantially improve scene understanding and geometric reasoning (Bachmann et al., 2022; Mizrahi et al., 2023; Bachmann et al., 2024; Girdhar et al., 2023; Lu et al., 2022). However, the empirical impact of incorporating explicit geometric information into surgical visual representation learning remains insufficiently explored. To this end, we systematically compare unimodal and multimodal pre-training strategies for surgical vision. Throughout this paper, we use the term *multimodal* to refer specifically to models that jointly process RGB images and aligned depth maps (RGB-D).

In this work, we present an empirical evaluation of RGB and RGB-D self-supervised pre-training strategies for surgical vision foundation models. We curate an in-house dataset of 1.4 million images from the da Vinci (DV) system clinical procedures and generated pseudo-depth labels using an off-the-shelf stereo disparity model. We build and benchmark eight models, disentangling the effects of pre-training data source (e.g., natural vs. surgical images), modality (RGB vs. RGB-D), and pretext task objective (e.g., MAE, DINO, MultiMAE). To evaluate representation quality, we probe frozen backbones and perform end-to-end fine-tuning across 8 datasets spanning 4 distinct downstream tasks.

Our contributions are summarized as follows:

- We conduct a large-scale empirical study comparing eight ViT-based VFMs and demonstrate that multimodal (RGB-D) pre-training substantially outperforms RGB-only alternatives across four downstream tasks: object detection, semantic segmentation, pose estimation, and depth estimation.
- We show that these findings generalize beyond our in-house data by repeating a subset of experiments on the publicly available LEMON dataset (Che et al., 2025), confirming the superiority of multimodal pre-training regardless of the pre-training corpus or depth source.

- We provide evidence that geometry-aware models exhibit remarkable data efficiency: models fine-tuned on just 25% of labeled data consistently outperform RGB-only baselines trained on the full dataset.

- We analyze the factors driving these improvements and provide actionable recommendations for geometry-aware surgical VFM development.

2 Related Work

2.1 Vision Foundation Models for Surgery

Developing VFMs typically requires large-scale data, a resource that is notoriously difficult to curate in the surgical domain due to privacy concerns. Consequently, early literature focused on benchmarking architectures using aggregated public datasets. Ramesh et al. (2023) pre-trains various architectures on the Cholec80 dataset (Twinanda et al., 2016) to evaluate tool detection and phase recognition. Subsequent work has focused on scaling this paradigm by aggregating diverse public sources. For instance, Batić et al. (2024), Schmidgall et al. (2024), and Yang et al. (2025) leverage large-scale public data compilations to fine-tune models for semantic segmentation, action triplet recognition, and phase recognition.

More recently, efforts have shifted toward scaling with proprietary data. Wang et al. (2023) and Jaspers et al. (2025) combined large-scale private repositories with public benchmarks to train VFMs for colonoscopy and laparoscopy, respectively, demonstrating that data scale correlates with downstream performance. However, a critical limitation persists across these works: they rely exclusively on unimodal RGB pre-training. Although effective for global image-level tasks (e.g., tool recognition), RGB-only models lack explicit geometric inductive biases. Although task-specific approaches such as that of Jamal & Mohareri (2024) have explored fusing RGB and depth for segmentation, this work only applies to a specific application rather than a general-purpose representation learning strategy. In contrast, our work introduces depth as an explicit signal during the pre-training stage to enable 3D scene understanding across a plethora of downstream tasks. Furthermore, our multimodal models do not require depth during inference or fine-tuning, making adoption straightforward and demonstrating clear superiority over RGB-only baselines.

2.2 Self-supervised RGB-D Pre-training

Recently, multimodal pre-training has emerged as a popular alternative to unimodal VFMs. This shift is driven by the availability of powerful off-the-shelf models for generating pseudo-labels and the need for richer physical representations. Indeed, studies such as Chen et al. (2025) demonstrate via probing that feedforward reconstruction models such as DUSt3R (Wang et al., 2024) learn richer representations for 3D geometric tasks than traditional unimodal VFMs.

A dominant paradigm in this space is the "masked modeling" of multimodal tokens. For instance, MultiMAE (Bachmann et al., 2022) extends the Masked Autoencoder (MAE) (He et al., 2022) by projecting RGB, depth, and semantic maps into a unified latent space, training the model to reconstruct missing patches across all modalities simultaneously. Similarly, Mask3D (Hou et al., 2023) employs a masked prediction objective specifically designed to learn 3D priors by reconstructing depth from masked RGB-D inputs. This paradigm has been pushed further by 4M (Mizrahi et al., 2023; Bachmann et al., 2024), which scales to 7 and even 21 modalities using modality-specific tokenizers such as VQ-VAE (Van Den Oord et al., 2017), effectively turning multimodal representation learning into a discrete sequence modeling problem.

The DINO family has shown that geometric properties can emerge implicitly from teacher-student distillation. Previous works have successfully probed DINOv2 features for depth estimation (Yang et al., 2024b), 3D segmentation (Zeid et al., 2025), and spatial understanding (El Banani et al., 2024; Man et al., 2024). However, these capabilities are emergent rather than explicit. In this study, we hypothesize and demonstrate that while DINOv2 captures implicit geometry, models explicitly pre-trained with RGB-D signals (like MultiMAE) yield superior representations for fine-grained geometric surgical tasks.

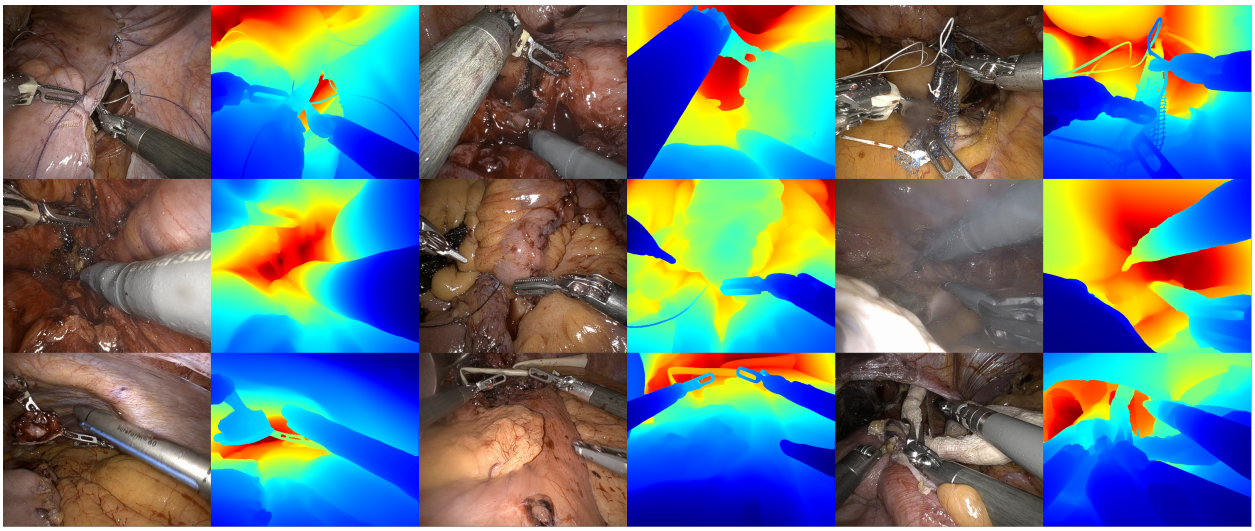

Figure 2: Samples from the pre-training dataset. We curate an in-house RGB-D dataset from videos collected via the da Vinci (DV) system and generate pseudo-labeled depth maps via FoundationStereo (Wen et al., 2025), which is robust against corruptions such as blur and smoke.

3 Methods

3.1 Pre-training Data

We compile an in-house dataset composed of endoscopic robot-assisted stereo surgical videos acquired from the da Vinci (DV) Surgical System (Intuitive Inc., Sunnyvale, CA). To curate the dataset, we sample about 1.4 million frames from the recorded video dataset. This is to ensure that the dataset scale is similar to that of ImageNet-1k. We use the left camera view for all RGB images.

To generate depth images, we leverage the recently released FoundationStereo (Wen et al., 2025) model, which generates disparity maps given stereo rectified pairs of images. Although disparity can be trivially converted to metric depth maps via triangulation, we did not observe significant differences in performance when using depth maps. Thus, we simply maintain them as disparity maps and normalize them to $[0, 1]$ during pre-training, following previous work (Bachmann et al., 2022). For clarity, we refer to these normalized disparity maps as depth throughout the remainder of the paper. Figure 2 displays examples of RGB-D pairs from the pre-training dataset.

In order to verify whether or not our takeaways are generalizable across different pre-training data, we repeat some of our experiments on the LEMON (Che et al., 2025) dataset, composed of 3.4M surgical image frames extracted from publicly available surgical videos. Because LEMON is a dataset of monocular surgical images, we use the ViT-L variant of Depth Anything V2 (Yang et al., 2024c) to extract pseudolabels for depth maps. We will release our models trained on this dataset for public use.

3.2 Models

Table 1 provides an overview of the eight models evaluated in this study. For all models, we follow the implementation and training details described in the original works, unless stated otherwise.

Pre-trained ViTs We evaluate three pre-trained vision encoders from prior works: MAE (He et al., 2022) pre-trained on ImageNet-1k, DINOv2 (Oquab et al., 2023) (without registers) pre-trained on LVD-142M, and MultiMAE (Bachmann et al., 2022) pre-trained on ImageNet-1k with pseudo-labeled multimodal data. Henceforth, these will be referred as "MAE (IN-1k)", "DINOv2 (LVD)", and "MultiMAE (IN-1k)" respectively.

Table 1: Overview of the 8 self-supervised models evaluated in this study. The models vary by pre-training architecture, data source (ImageNet-1k, LVD-142M, or in-house DV), and modality usage. We use the ViT-Base variant for all models. When unspecified in parenthesis, the pre-training data source is the in-house DV dataset.

Model Name	**Data**	**Epochs**	**Pretext Task Description**
RGB-Only Baselines			
MAE (IN-1k)	ImageNet-1k	1600	Reconstructs masked RGB patches
MAE	da Vinci	400	Reconstructs masked RGB patches
DINOv2 (LVD)	LVD-142M	100	Self-distillation with no labels
DINOv2-RGB	da Vinci	100	Self-distillation with no labels
Multimodal - Naive Data Augmentation			
DINOv2-RGBD	da Vinci	100	Self-distillation with RGB-D input (See App A)
Multimodal - Explicit Tokenization			
MultiMAE (IN-1k)	ImageNet-1k	1600	Input: RGB, Depth, & Semantic Output: All modalities reconstructed
MultiMAE	da Vinci	400	Input: RGB & Depth Output: Reconstructed RGB & Depth
Mask3D	da Vinci	100	Input: RGB & Depth Output: Reconstructed Depth

Surgical domain models. We train each architecture on our in-house da Vinci (DV) dataset using the original training configurations. MAE and DINOv2 are trained on RGB images, while MultiMAE is trained on RGB-D pairs. We refer to these as "MAE", "DINOv2-RGB", and "MultiMAE" respectively. Although the original MultiMAE model was trained with 3 modalities (RGB, depth, and semantic maps), we train only on RGB and depth. Due to computational constraints, we train MAE for 400 epochs, DINOv2 for 100 epochs, and MultiMAE for 400 epochs.

Mask3D. We additionally train Mask3D (Hou et al., 2023) on the surgical dataset. Since the original authors have not released their code or model weights, we implement our own version by building on the original MAE codebase. Following the original paper, we initialize the RGB encoder with MAE (IN-1k) weights, only mask depth patches where no RGB patches are masked, and train for 100 epochs. Although the original paper uses both the encoder and decoder for downstream tasks, we opt to only use the RGB encoder to ensure fair comparison across models.

DINOv2-RGBD. While multimodal masked autoencoders have been explored, extending DINO-style self-distillation to additional data modalities remains underexplored. To bridge this gap, we introduce DINOv2-RGBD, a straightforward adaptation of DINOv2 with a joint-modality data augmentation strategy. In standard DINOv2 training, random local and global crops of the RGB image are provided to the teacher and student networks. In our quasi-multimodal variant, crops from either RGB image or depth image can be provided to either network. Consequently, the student and teacher often receive corresponding crops from different modalities, encouraging the model to learn representations that are invariant across RGB and depth. See Appendix A and Fig. 5 for further details.

Additional pre-training details are provided in Appendix B. Furthermore, Appendix C presents feature map visualizations for the DINOv2 variants and qualitative reconstruction results for the masked image modeling approaches.

Table 2: Overview of the 8 downstream datasets used for evaluating transfer performance. The benchmarks cover four distinct geometric and semantic tasks across both laparoscopic and robotic surgical domains. *Because non-keyframes in the SCARED dataset are inaccurate due to kinematic and calibration errors, we correct these frames with the method outlined in Appendix D.

Task	Dataset	Domain
Object Detection	CholecTrack20 (Nwoye et al., 2025)	Laparoscopic (Cholecystectomy)
	m2cai16-tool-locations (Jin et al., 2018)	Laparoscopic (Cholecystectomy)
Segmentation	CholecInstanceSeg (Alabi et al., 2025)	Laparoscopic (Cholecystectomy)
	EndoVis18 (Allan et al., 2020)	Robotic (da Vinci)
Depth Estimation	DV (Pseudo-labeled)	Robotic (da Vinci)
	SCARED-C* (Allan et al., 2021)	Robotic (da Vinci)
Pose Estimation	PhaKIR (Rueckert et al., 2025)	Laparoscopic (Cholecystectomy)
	SurgPose (Wu et al., 2025)	Robotic (da Vinci)

3.3 Downstream Tasks

The goal of this paper is to analyze the impact of explicit geometric priors in pre-training vision foundation models. As a result, we select downstream tasks that can evaluate the models' spatial, semantic, and geometric understanding rather than global or image-level tasks. We evaluate models on four representative tasks: **object detection**, **semantic segmentation**, **depth estimation**, and **pose estimation**. Each task is evaluated using two datasets, which are detailed in Table 2.

Among these tasks, depth estimation presents unique challenges in the surgical domain. Due to physical constraints of minimally invasive surgery, it is difficult to acquire ground truth depth maps in surgical scenes. The SCARED dataset uses an external structured light sensor in ex vivo porcine tissue. To generate a video dataset with paired depth, the endoscope is moved using the robotic arm and the keyframe depth map is reprojected into the non-keyframe camera views. Due to calibration and kinematics errors, these depth maps are not reliable for training or evaluation, shown in Fig. 10. We use COLMAP (Schönberger & Frahm, 2016) to correct the non-keyframe depth maps; see Appendix D for details. We will release this corrected dataset for public use.

For the remaining depth estimation dataset (DV), we generate pseudo-labeled depth maps using disparity maps obtained from FoundationStereo (Wen et al., 2025). Notably, among the eight datasets evaluated, DV is the only one that is in-domain with respect to the pre-training. For both depth estimation datasets, we normalize the depth maps to perform relative depth estimation.

For all downstream tasks and datasets, we follow the official data splits provided by the original benchmarks. Since the PhaKIR dataset (Rueckert et al., 2025) does not publicly release a validation set, we choose one training video sequence (Video 13) out of the 8 as the validation set.

3.4 Implementation Details and Task Specific Heads

We conduct all experiments using PyTorch on NVIDIA H200s and A100s. When available, we use the original codebases for pre-training. For downstream evaluation, we use the OpenMMLab ecosystem: mmpose (Contributors, 2020a) for pose estimation, mmdetection (Chen et al., 2019) for object detection, and mmsegmentation (Contributors, 2020b) for segmentation and depth estimation.

Frozen backbone evaluation. For segmentation and depth estimation, we train a single linear layer on top of frozen backbone features. For pose estimation and object detection, we use task-specific heads: ViTPose (Xu et al., 2022) for pose estimation and Mask-RCNN (He et al., 2017) for object detection, following the ViTDet (Li et al., 2022) setup. In all cases, only the task head is trained while the backbone remains frozen.

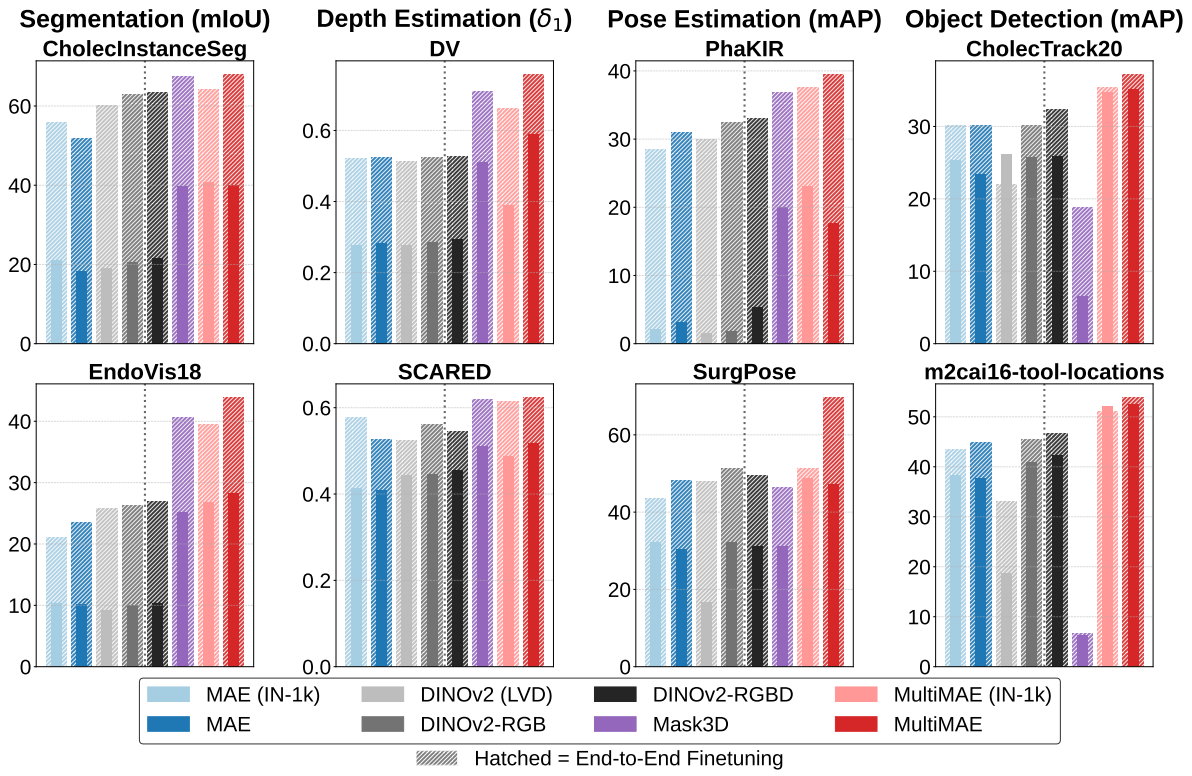

Figure 3: Comparison of downstream task performance on eight surgical datasets. Solid inner bars represent performance with a frozen backbone (linear probing), while hatched outer bars represent end-to-end fine-tuning. Vertical dotted lines distinguish between unimodal and multimodal models. Exact numerical results are provided in Appendix G.

End-to-end fine-tuning. For full fine-tuning experiments, we update both the backbone and task-specific heads. For pose estimation and object detection, we retain the same heads used in the frozen-backbone setting. For semantic segmentation and depth estimation, we replace the linear head with a DPT-based decoder (Ranftl et al., 2021) to better exploit the additional modeling capacity afforded by end-to-end optimization.

Additional fine-tuning details are provided in Appendix E. We further ablate the impact of task-head design choices in Table 9 of Appendix F.

4 Results and Discussion

This section presents a comprehensive analysis of VFM performance on downstream spatial and geometric tasks. We report results on two primary experiments. First, we evaluate eight VFMs across eight datasets spanning four geometric downstream tasks (Fig. 3). Qualitative results can also be seen in Appendix H. Second, we analyze data efficiency through end-to-end fine-tuning on progressively smaller subsets of each downstream dataset, using $n\%$ of the labeled data where $n \in \{25, 50, 75, 100\}$. For the data efficiency experiments, we randomly sample $n\%$ of the training data using three different random seeds and report mean performance across runs (Fig. 4).

We organize our analysis around five key findings that characterize the impact of explicit geometric priors and domain-specific pre-training on surgical scene understanding.

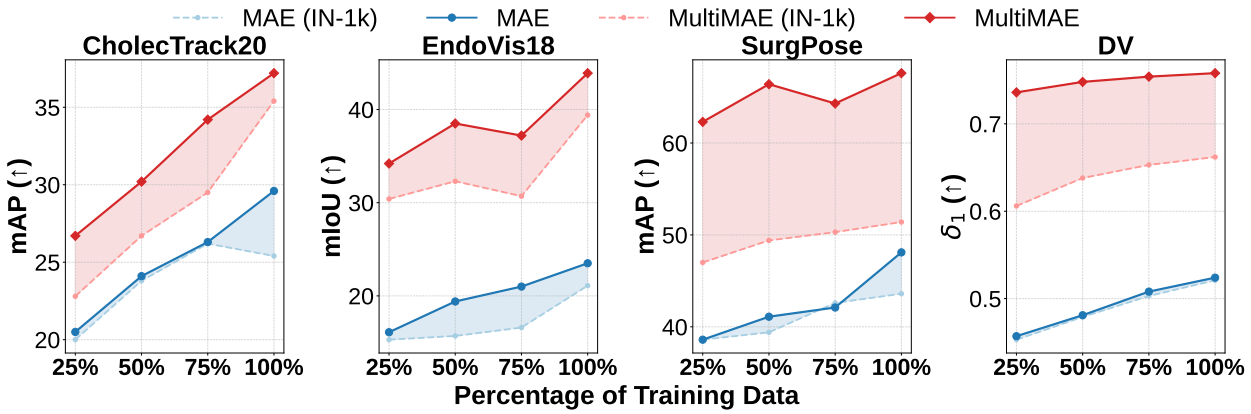

Figure 4: Data efficiency analysis comparing model performance at 25%, 50%, 75%, and 100% training data availability. Solid lines indicate surgical domain (DV) pre-training and dashed lines indicate ImageNet pre-training. The horizontal dotted line marks the baseline performance of MAE (DV) trained on 100% of the data. Numerical results are provided in Appendix G.

Table 3: Downstream performance of MAE vs. MultiMAE encoders pre-trained on two surgical video datasets (DV and LEMON Che et al. (2025)). Each cell reports end-to-end fine-tuned (frozen-encoder) results. MultiMAE consistently outperforms MAE across all tasks, datasets, and evaluation protocols, confirming that multi-modal masked pre-training yields stronger surgical vision representations regardless of the pre-training corpus. Best results per pre-training dataset are shown in **bold**.

		DV Pre-training		LEMON Pre-training	
Task (Metric)	**Dataset**	MAE	MultiMAE	MAE	MultiMAE
Object Det. (mAP↑)	CholecTrack20	30.1 (23.4)	**37.2** (**35.2**)	27.9 (22.4)	**33.3** (**40.1**)
	m2cai16	44.9 (37.7)	**53.9** (**52.5**)	44.9 (37.7)	**46.2** (**51.8**)
Segmentation (mIoU↑)	EndoVis18	23.5 (10.2)	**43.9** (**28.2**)	17.4 (10.1)	**34.9** (**30.8**)
	CholecInstanceSeg	51.7 (18.3)	**68.0** (**40.0**)	54.8 (19.3)	**57.2** (**59.5**)
Pose Est. (mAP↑)	SurgPose	48.1 (29.4)	**69.7** (**47.1**)	40.0 (13.1)	**57.2** (**54.2**)
	PhaKIR	30.9 (3.2)	**39.5** (**17.7**)	21.5 (2.2)	**38.3** (**23.3**)
Depth Est. (δ_1↑)	DV Depth	.524 (.283)	**.758** (**.590**)	.518 (.288)	**.657** (**.469**)
	SCARED	.526 (.408)	**.593** (**.517**)	.534 (.443)	**.545** (**.485**)

Finding 1

Multimodal pre-training consistently outperforms RGB-only pre-training, with larger gains on dense prediction tasks. This advantage holds regardless of the pre-training data corpus.

Among the eight models evaluated on DV pre-training (Fig. 3), MultiMAE achieves the highest performance on every dataset under end-to-end fine-tuning and on most datasets under frozen-backbone evaluation. In end-to-end fine-tuning, MultiMAE outperforms the best RGB-only baseline (DINOv2-RGB) by an average relative margin of 23.7% across all eight benchmarks. The gains are especially pronounced on dense prediction tasks: on EndoVis18 segmentation, MultiMAE reaches 43.9 mIoU compared to 26.2 for DINOv2-RGB, and on DV depth estimation, MultiMAE achieves 0.758 δ_1 compared to 0.526 for DINOv2-RGBD. These trends hold in the frozen-backbone setting as well, where MultiMAE features consistently yield higher readout accuracy than all RGB-only alternatives.

Interestingly, Mask3D exhibits divergent behavior across task types. While it struggles with bounding-box detection (6.5 mAP on CholecTrack20), it achieves strong performance on segmentation (67.5 mIoU on CholecInstanceSeg) and depth estimation (0.709 δ_1 on DV). We hypothesize that this discrepancy arises from Mask3D's pretext task design: reconstructing only depth patches may induce representations well-suited for dense pixel-wise prediction but lacking the semantic abstraction required for object-level reasoning. In contrast, MultiMAE reconstructs all input modalities jointly, which may encourage more balanced representations that transfer effectively across both semantic and geometric tasks. These results further suggest that architectures employing a single shared encoder across modalities, as in MultiMAE, may be preferable to those with modality-specific encoders, such as Mask3D, for learning transferable spatial representations.

To verify that these conclusions are generalizable across different data corpora, we repeat the MAE and MultiMAE comparison experiments using the publicly available LEMON Che et al. (2025) dataset (Table 3). The same pattern holds: MultiMAE leads on all eight benchmarks under both end-to-end and frozen backbone fine-tuning settings regardless of the pre-training dataset. Notably, the frozen-backbone gap is often larger than the end-to-end gap. For instance, on CholecInstanceSeg with LEMON pre-training, frozen MultiMAE reaches 59.5 mIoU compared to 19.3 for frozen MAE. Furthermore, both MultiMAE and MAE pre-trained on the in-house DV data outperform its counterparts pre-trained on the public LEMON dataset, despite the latter being around 2.5 times larger in quantity.

Finding 2

Geometry-aware pre-training improves data efficiency in downstream surgical tasks.

We evaluate data efficiency by fine-tuning models end-to-end on randomly sampled subsets (25%, 50%, 75%) of the training data, reporting mean performance across three random seeds. The results in Fig. 4 reveal a consistent trend: MultiMAE (DV) achieves the strongest performance across all data fractions and evaluated tasks. Notably, MultiMAE (DV) fine-tuned on only 25% of training data frequently matches or exceeds the performance of RGB-only models fine-tuned on the full dataset.

On EndoVis18, MultiMAE trained with 25% of the data achieves 34.2 mIoU, substantially exceeding fully-trained MAE at 23.5 mIoU. The gap is even more pronounced on SurgPose, where MultiMAE at 25% (62.3 mAP) surpasses MAE at 100% (48.1 mAP) by a wide margin. These results indicate that geometric priors provide a favorable initialization that enables effective few-shot learning. This property is particularly valuable in surgical vision, where expert annotations are scarce and expensive to acquire. We further demonstrate this observation in Fig. 11 (Appendix F), where MultiMAE fine-tuned on only 2.5% of EndoVis18 training data still outperforms MAE fine-tuned on the complete dataset.

Finding 3

Frozen multimodal backbones often outperform end-to-end fine-tuned RGB-only models.

A striking result emerges when comparing evaluation protocols: frozen backbones from multimodal models with trainable task-specific heads are competitive with, and in many cases outperform, fully fine-tuned RGB-only models. This finding has significant practical implications. A single frozen backbone can be shared across multiple downstream tasks, reducing memory footprint and enabling shared feature computation when running multiple tasks simultaneously. Additionally, training only the task-specific heads is faster, requires less GPU memory, and avoids catastrophic forgetting when adapting to new tasks.

This phenomenon is particularly evident on CholecTrack20, where a frozen MultiMAE (DV) backbone achieves 35.2 mAP, outperforming all RGB-only models even under end-to-end fine-tuning (best result: DINOv2-RGBD at 32.4 mAP). The performance gap is more pronounced on geometry-focused tasks. For instance, on SurgPose, frozen MultiMAE achieves 47.1 mAP, compared to 32.2 mAP for frozen MAE. These results indicate that geometry-aware pre-training can encode high-fidelity spatial understanding directly into the learned representations, whereas RGB-only backbones require task-specific adaptation to recover comparable performance in downstream tasks.

Finding 4

Multimodal pre-training can mitigate domain shift and reduces dependence on pre-training data diversity.

For RGB-only models, we observe that ImageNet pre-training (IN-1k) often outperforms surgical domain pre-training (DV), particularly in detection tasks. On CholecInstanceSeg, MAE (IN-1k) achieves 55.9 mIoU compared to 51.7 mIoU for MAE (DV) in end-to-end fine-tuning and 21.1 vs. 18.3 mAP in the linear probing setting. This pattern suggests that for RGB-only features, the limited visual diversity of DV data constrains generalization, whereas ImageNet's breadth provides more robust texture-based representations.

However, this trend reverses with multimodal pre-training. MultiMAE (DV) outperforms MultiMAE (IN-1k) on most linear probing tasks and all end-to-end fine-tuning tasks. On SurgPose, surgical pre-training yields an 18.3 mAP improvement over ImageNet pre-training (69.7 vs. 51.4 mAP). We hypothesize that depth information acts as a regularizer that reduces reliance on texture-based features, which are highly domain-specific. Geometric structure (3D shape of instruments, tissue surfaces, and their spatial relationships) exhibits greater invariance across visual domains than appearance-based cues. Consequently, multimodal models can learn robust representations even from relatively homogeneous datasets, as the geometric signal provides complementary information that compensates for limited visual diversity. Examples from each dataset can be seen in Appendix H.

Finding 5

Explicit geometric tokenization outperforms naive multimodal augmentation.

To disentangle whether performance gains stem from the geometric data itself or the modeling strategy, we compare DINOv2-RGBD (naive augmentation, where depth is provided as an additional input without architectural modification) against MultiMAE (explicit geometric tokenization with cross-modal reconstruction objectives).

Across all datasets except SurgPose, DINOv2-RGBD provides only marginal improvements over DINOv2-RGB in end-to-end fine-tuning, with an average relative gain of just 3.08%. On SurgPose, naive depth augmentation actually degrades performance (49.4 vs. 51.4 mAP), suggesting that without appropriate architectural constraints, depth information can interfere with representation learning. In contrast, MultiMAE achieves an average relative improvement of 39.1% over MAE across the same comparisons.

These results confirm that effective multimodal learning requires architectures explicitly designed to model the joint distribution of visual and geometric tokens. Simply incorporating depth as an additional input channel fails to enforce geometric consistency in the learned representations. The cross-modal reconstruction objective in MultiMAE, which requires the model to predict masked tokens in one modality given tokens from another, appears essential for learning representations that meaningfully integrate RGB and depth information.

4.1 Limitations and Future Work

Several limitations of our study warrant discussion. First, our depth maps are generated using an off-the-shelf stereo matching model (FoundationStereo) rather than ground-truth sensor data. While this reflects practical deployment scenarios where ground-truth depth is unavailable, disparity estimation errors may propagate into learned representations. Regardless, in Appendix F, we demonstrate comparable performance when using Depth Anything (Yang et al., 2024c) to generate pseudo labeled depth maps for the DV dataset, suggesting that our approach is applicable also to large-scale monocular surgical datasets. Furthermore, we observe notable gains in multimodal models despite imperfect depth maps.

Second, our evaluation focuses on four task categories spanning eight datasets, all derived from laparoscopic cholecystectomy or similar minimally invasive procedures. Generalization to other surgical domains (e.g., open surgery, colonoscopy) remains to be validated. The geometric priors learned from one surgical

context may not transfer uniformly across procedures with substantially different visual characteristics or instruments.

Finally, our analysis solely focuses on downstream tasks in computer vision. An important direction for future work is evaluating whether multimodal pre-training similarly benefits other types of tasks relevant to surgical robotics, such as surgical synthetic data generation, predictive modeling for autonomous actions, or vision-language alignment for instruction following.

5 Conclusion

We present a systematic evaluation of unimodal and multimodal pre-training for surgical computer vision, benchmarking eight foundation models on segmentation, object detection, pose estimation, and depth estimation. Our findings establish that incorporating geometric priors through explicit depth supervision during pre-training consistently and substantially outperforms RGB-only approaches, with relative improvements of up to 35% on downstream tasks. Furthermore, we show that these takeaways generalize across different pre-training corpora (DV vs. LEMON).

Crucially, these gains come at minimal practical cost. Depth maps can be generated from existing RGB data using off-the-shelf stereo or monocular depth estimation models, requiring no additional sensor hardware or data collection. Furthermore, depth is used exclusively during pre-training—at inference time, multimodal models accept standard RGB input and maintain identical architecture and computational cost to their RGB-only counterparts. This means practitioners can use a multimodal backbone instead of an RGB-only backbone with no significant changes to their downstream pipeline, deployment infrastructure, or inference latency.

Three results carry particular practical significance. First, multimodal models exhibit remarkable data efficiency: MultiMAE pre-trained on surgical data and fine-tuned with just 25% of labeled examples frequently surpasses RGB-only models trained on complete datasets. In annotation-scarce surgical domains, this property could substantially reduce the barrier to developing task-specific models. Second, frozen backbones from geometry-aware pre-training outperform fully fine-tuned RGB-only models, enabling computationally efficient deployment without sacrificing performance. Third, geometric pre-training mitigates domain shift, allowing models trained on relatively homogeneous surgical datasets to match or exceed those pre-trained on diverse natural image corpora.

Our analysis also reveals that naive multimodal augmentation fails to capture these benefits; explicit architectural design for cross-modal learning is essential. Simply providing depth as an additional input yields marginal gains, whereas architectures with cross-modal reconstruction objectives during pre-training achieve substantial improvements.

These findings offer clear guidance for the surgical AI community: multimodal pre-training on domain-specific data represents a more effective foundation than scaling RGB-only approaches without adding hardware modifications or inference overhead.

References

Ayberk Acar, Mariana Smith, Lidia Al-Zogbi, Tanner Watts, Fangjie Li, Hao Li, Nural Yilmaz, Paul Maria Scheikl, Jesse F d'Almeida, Susheela Sharma, et al. From monocular vision to autonomous action: Guiding tumor resection via 3d reconstruction. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 21714–21720. IEEE, 2025.

Oluwatosin Alabi, Ko Ko Zayar Toe, Zijian Zhou, Charlie Budd, Nicholas Raison, Miaojing Shi, and Tom Vercauteren. Cholecinstanceseg: A tool instance segmentation dataset for laparoscopic surgery. *Scientific Data*, 12(1):825, 2025.

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.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pp. 348–367. Springer, 2022.

Roman Bachmann, Oğuzhan F Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities. *Advances in Neural Information Processing Systems*, 37:61872–61911, 2024.

Dominik Batić, Felix Holm, Ege Özsoy, Tobias Czempiel, and Nassir Navab. Endovit: pretraining vision transformers on a large collection of endoscopic images. *International Journal of Computer Assisted Radiology and Surgery*, 19(6):1085–1091, 2024.

Gui-Bin Bian, Yaqin Peng, Li Zhang, Jun Li, and Zhen Li. Algorithms in surgical action recognition: A survey. In *2024 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pp. 366–371. IEEE, 2024.

Chengan Che, Chao Wang, Tom Vercauteren, Sophia Tsoka, and Luis C Garcia-Peraza-Herrera. Lemon: A large endoscopic monocular dataset and foundation model for perception in surgical settings. *arXiv preprint arXiv:2503.19740*, 2025.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2gs: Probing visual foundation models with gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6348–6361, 2025.

MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020a.

MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020b.

Tommaso Da Col, Guido Caccianiga, Michele Catellani, Andrea Mariani, Matteo Ferro, Giovanni Cordima, Elena De Momi, Giancarlo Ferrigno, and Ottavio De Cobelli. Automating endoscope motion in robotic surgery: A usability study on da vinci-assisted ex vivo neobladder reconstruction. *Frontiers in Robotics and AI*, 8:707704, 2021.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806, 2024.

Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10406–10417, 2023.

John J Han, Ayberk Acar, Callahan Henry, and Jie Ying Wu. Depth anything in medical images: A comparative study. *arXiv preprint arXiv:2401.16600*, 2024.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Roy Hirsch, Mathilde Caron, Regev Cohen, Amir Livne, Ron Shapiro, Tomer Golany, Roman Goldenberg, Daniel Freedman, and Ehud Rivlin. Self-supervised learning for endoscopic video analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 569–578. Springer, 2023.

Mohammadmahdi Honarmand, Muhammad Abdullah Jamal, and Omid Mohareri. Vidlpro: A video-language pre-training framework for robotic and laparoscopic surgery. In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*, 2024.

Ji Hou, Xiaoliang Dai, Zijian He, Angela Dai, and Matthias Nießner. Mask3d: Pre-training 2d vision transformers by learning masked 3d priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13510–13519, 2023.

Muhammad Abdullah Jamal and Omid Mohareri. Rethinking rgb-d fusion for semantic segmentation in surgical datasets. *arXiv preprint arXiv:2407.19714*, 2024.

Tim JM Jaspers, Ronald LPD de Jong, Yiping Li, Carolus HJ Kusters, Franciscus HA Bakker, Romy C van Jaarsveld, Gino M Kuiper, Richard van Hillegersberg, Jelle P Ruurda, Willem M Brinkman, et al. Scaling up self-supervised learning for improved surgical foundation models. *arXiv preprint arXiv:2501.09436*, 2025.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 691–699. IEEE, 2018.

Ji Woong Kim, Tony Z Zhao, Samuel Schmidgall, Anton Deguet, Marin Kobilarov, Chelsea Finn, and Axel Krieger. Surgical robot transformer (srt): Imitation learning for surgical tasks. *arXiv preprint arXiv:2407.12998*, 2024.

Ji Woong Kim, Juo-Tung Chen, Pascal Hansen, Lucy Xiaoyang Shi, Antony Goldenberg, Samuel Schmidgall, Paul Maria Scheikl, Anton Deguet, Brandon M White, De Ru Tsai, et al. Srt-h: A hierarchical framework for autonomous surgery via language-conditioned imitation learning. *Science robotics*, 10(104):eadt5254, 2025.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pp. 280–296. Springer, 2022.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liangyan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *Advances in Neural Information Processing Systems*, 37:76819–76847, 2024.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

Chinedu Innocent Nwoye, Kareem Elgohary, Anvita Srinivas, Fauzan Zaid, Joël L Lavanchy, and Nicolas Padoy. Cholectrack20: A multi-perspective tracking dataset for surgical tools. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8942–8952, 2025.

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.

Alejandra Perez, Chinedu Nwoye, Ramtin Raji Kermani, Omid Mohareri, and Muhammad Abdullah Jamal. Surglavi: Large-scale hierarchical dataset for surgical vision-language representation learning. *arXiv preprint arXiv:2509.10555*, 2025.

Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Saurav Sharma, Antoine Fleurentin, et al. Dissecting self-supervised learning methods for surgical computer vision. *Medical Image Analysis*, 88:102844, 2023.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149, 2016.

Tobias Rueckert, Raphaela Maerkl, David Rauber, Leonard Klausmann, Max Gutbrod, Daniel Rueckert, Hubertus Feussner, Dirk Wilhelm, and Christoph Palm. Video dataset for surgical phase, keypoint, and instrument recognition in laparoscopic surgery (phakir). *arXiv preprint arXiv:2511.06549*, 2025.

Samuel Schmidgall, Ji Woong Kim, Jeffrey Jopling, and Axel Krieger. General surgery vision transformer: A video pre-trained foundation model for general surgery. *arXiv preprint arXiv:2403.05949*, 2024.

Adam Schmidt, Omid Mohareri, Simon DiMaio, Michael C Yip, and Septimiu E Salcudean. Tracking and mapping in medical computer vision: A review. *Medical Image Analysis*, 94:103131, 2024.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hongchao Shu, Mingxu Liu, Lalithkumar Seenivasan, Suxi Gu, Ping-Cheng Ku, Jonathan Knopf, Russell Taylor, and Mathias Unberath. Seamless augmented reality integration in arthroscopy: a pipeline for articular reconstruction and guidance. *Healthcare Technology Letters*, 12(1):e12119, 2025.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272, 2021.

Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36, 02 2016. doi: 10.1109/TMI.2016.2593957.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome 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 *International conference on medical image computing and computer-assisted intervention*, pp. 431–441. Springer, 2022.

Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 101–111. Springer, 2023.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5249–5260, 2025.

Zijian Wu, Adam Schmidt, Randy Moore, Haoying Zhou, Alexandre Banks, Peter Kazanzides, and Septimiu E Salcudean. Surgpose: a dataset for articulated robotic surgical tool pose estimation and tracking. *arXiv preprint arXiv:2502.11534*, 2025.

Mengya Xu, Ziqi Guo, An Wang, Long Bai, and Hongliang Ren. A review of 3d reconstruction techniques for deformable tissues in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 157–167. Springer, 2024.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10371–10381, 2024b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024c.

Shu Yang, Fengtao Zhou, Leon Mayer, Fuxiang Huang, Yiliang Chen, Yihui Wang, Sunan He, Yuxiang Nie, Xi Wang, Ömer Sümer, et al. Large-scale self-supervised video foundation model for intelligent surgery. *arXiv preprint arXiv:2506.02692*, 2025.

Karim Abou Zeid, Kadir Yilmaz, Daan de Geus, Alexander Hermans, David Adrian, Timm Linder, and Bastian Leibe. Dino in the room: Leveraging 2d foundation models for 3d segmentation. *arXiv preprint arXiv:2503.18944*, 2025.

A Naive Multimodal DINOv2 Implementation

To investigate whether naive multimodal augmentation can induce geometric understanding without architectural modifications, we extend the standard DINOv2 framework to incorporate depth as an additional input modality. Our approach, which we term DINOv2-RGBD, treats RGB and depth images as interchangeable views within the self-distillation objective, requiring no changes to the underlying ViT architecture. Fig. 5 displays the workflow of DINOv2-RGBD.

During pre-training, we generate paired RGB and depth crops that share identical spatial transformations (crop coordinates, flips, and geometric augmentations), ensuring pixel-wise correspondence between modalities. At each training iteration, we independently sample which modality each view will use via per-view Bernoulli draws. Specifically, for each global crop fed to the teacher network, we select depth with probability p_{teacher}; similarly, for student global and local crops, we select depth with probabilities $p_{\text{student}}^{\text{global}}$ and $p_{\text{student}}^{\text{local}}$, respectively. In our experiments, all probabilities are set to 0.5, giving equal likelihood to either modality. This stochastic modality assignment means that within a single batch, the teacher may receive an RGB crop while the student receives the corresponding depth crop for the same spatial region, or vice versa. Since DINOv2's objective encourages the student to match the teacher's output distribution regardless of which specific views they observe, this setup implicitly encourages the model to learn representations that are invariant across modalities. This aligns RGB and depth embeddings for the same spatial content.

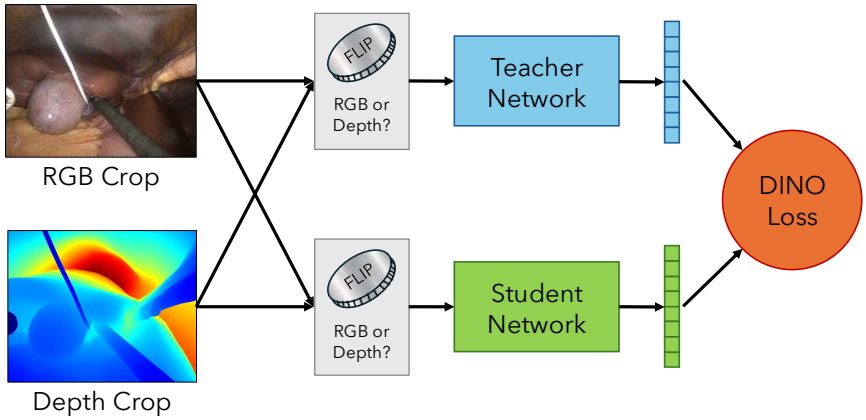

Figure 5: Demonstration of a naive multimodal DINOv2 training scheme. During training, both the teacher and student can receive crops of either modality, thus inducing learning of a joint probability distribution.

Importantly, this approach requires no modifications to the architecture, the self-distillation loss, or the inference pipeline. Depth maps are simply treated as three-channel images (replicated across channels) and processed identically to RGB inputs. The resulting model accepts standard RGB images at inference time with the same computational cost as the baseline DINOv2. However, as our results demonstrate, this naive augmentation strategy yields only marginal improvements over RGB-only pre-training, suggesting that explicit cross-modal reconstruction objectives, rather than implicit alignment through view augmentation, are necessary to fully leverage geometric information.

B Pre-training Details

This section provides training details for pre-training models, shown in Table 4. All trained models were initialized from scratch, uses the ViT-B model, and trained at (224, 224) resolution with patch size = 16. Data augmentations are HorizontalFlip($p = 0.5$) and ColorJitter. We use the LR linear scaling rule: $\text{lr} = \text{blr} \times \text{batch\_size}/256$. We normalize the RGB images with custom dataset statistics to zero mean and unit standard deviation.

C Feature Visualization and Masked Image Modeling Qualitative Results

In this section, we provide qualitative results on our models' pretext task in either feature map via principal component analysis (PCA) or masked image modeling.

C.1 DINOv2-RGB vs DINOv2-RGBD Features

We visualize the feature maps via (PCA) of our two DINOv2 models: DINOv2-RGB and DINOv2-RGBD, shown in Fig. 6. As can be seen, the feature maps of the DINOv2-RGBD model are identical regardless of modality because their features were implicitly aligned during pre-training. In contrast, DINOv2-RGB is inconsistent in its visualized features between the two modalities since it has only seen one modality during training.

C.2 MultiMAE Reconstruction Visualization

We display qualitative results from the pretext task for MultiMAE (Bachmann et al., 2022), which receives unmasked tokens of all modalities and reconstructs missing patches of all modalities. The number of visible patches is set constant throughout training, and the number of masked patches per modality is calculated via sampling from a symmetric Dirichlet distribution.

Table 4: Pre-training Hyperparameters.

(a) MAE Configuration

Config	Value
Epochs	400
Mask Ratio	75%
Optimizer	AdamW
Base Learning Rate	1×10^{-5}
Weight Decay	0.05
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.95$
Batch Size	5120
Learning Rate Schedule	Cosine Decay
Warmup Epochs	40

(b) MultiMAE Configuration

Config	Value
Epochs	400
Optimizer	AdamW
Base Learning Rate	1.0×10^{-4}
Weight Decay	0.05
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.95$
Batch Size	3072
Learning Rate Schedule	Cosine Decay
Warmup Epochs	40

(c) DINOv2 Configuration. Both RGB and RGB-D models have the same hyperparameters. For specific details, we use the default configuration from the DINOv2 codebase. We use 0 register tokens

Config	Value
Epochs	100
Optimizer	AdamW
Base Learning Rate	1.0×10^{-3}
Weight Decay	0.04
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.999$
Batch Size	1200
Drop Path Rate	0.3
LayerScale	1.0×10^{-5}
FFN	MLP
QKV Bias	True
Proj Bias	True
FFN Bias	True

(d) Mask3D Configuration, implemented from the MAE repository.

Config	Value
Epochs	100
Mask Ratio	80%
Optimizer	AdamW
Base Learning Rate	1×10^{-5}
Weight Decay	0.05
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.95$
Batch Size	4480
Learning Rate Schedule	Cosine Decay
Warmup Epochs	10

On the left side of Fig. 7, we see that a trained MultiMAE can completely synthesize RGB textures given the depth map (top left) while also being able to functionally perform monocular depth estimation given unmasked RGB patches (bottom left).

C.3 Mask3D Reconstruction Visualization

Mask3D's pretext task is to mask patches from both modalities and reconstruct solely missing depth patches. The masking strategy is also distinct from that of MultiMAE; depth patches are masked where there are no RGB patches. We demonstrate qualitative results of this reconstruction in Fig. 8.

C.4 MAE Reconstruction Visualization

MAE's pretext task is to mask RGB patches and reconstruct missing RGB patches. We demonstrate qualitative results of this reconstruction in Fig. 9.

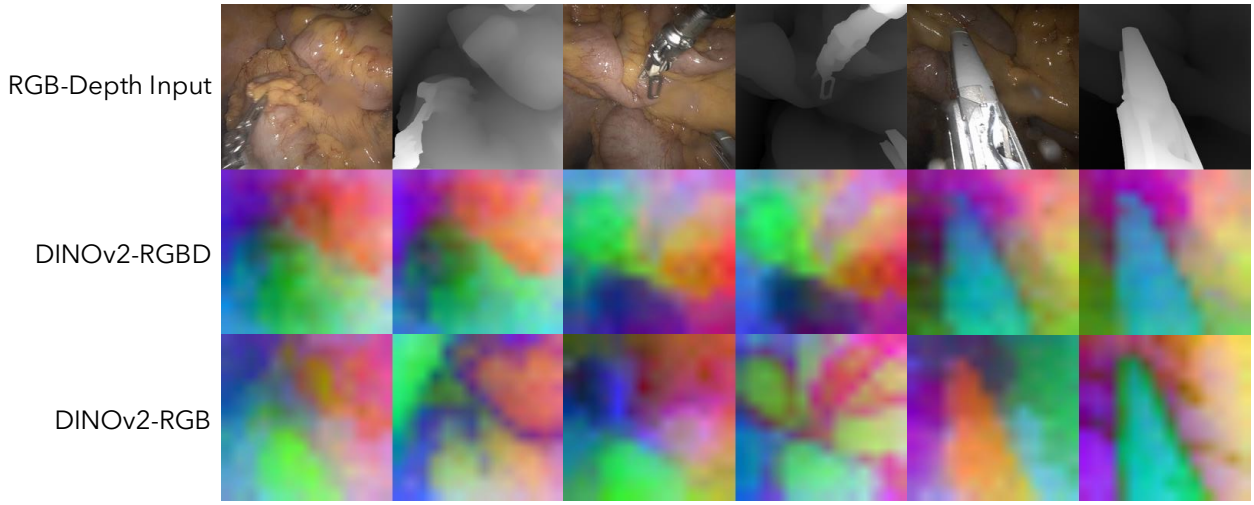

Figure 6: Feature map visualizations via PCA of the multimodally and unimodally trained DINOv2. The multimodal DINOv2, despite its simple data augmentation strategy, extracts similar representations between the two modalities as opposed to the unimodal's features.

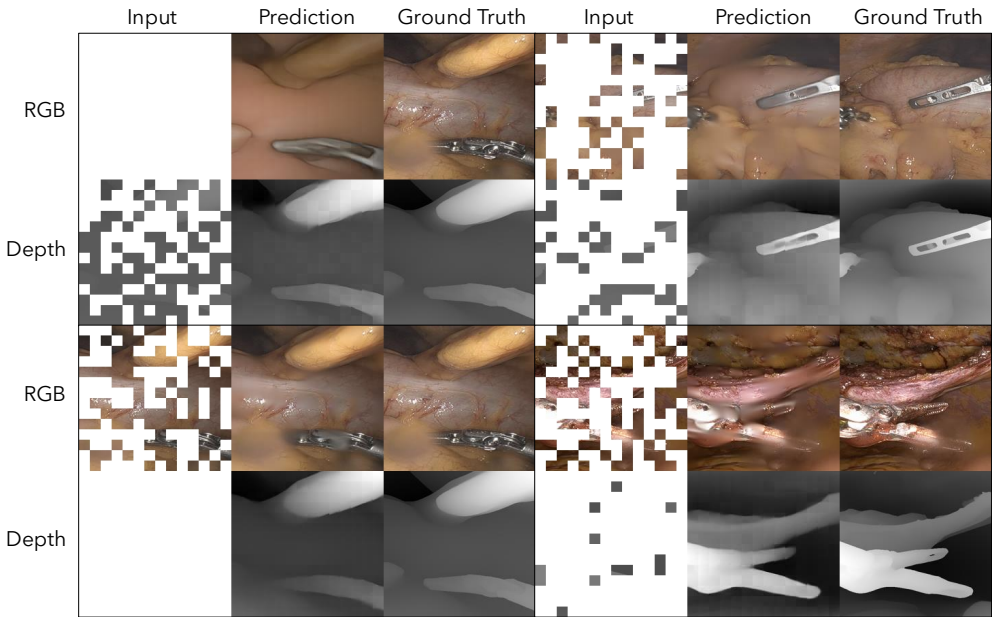

Figure 7: Reconstruction visualization for MultiMAE Bachmann et al. (2022). The model receives unmasked tokens of all modalities and reconstructs the missing patches of all modalities. MultiMAE can perform texture synthesis given unmasked depth patches and monocular depth estimation given unmasked RGB patches

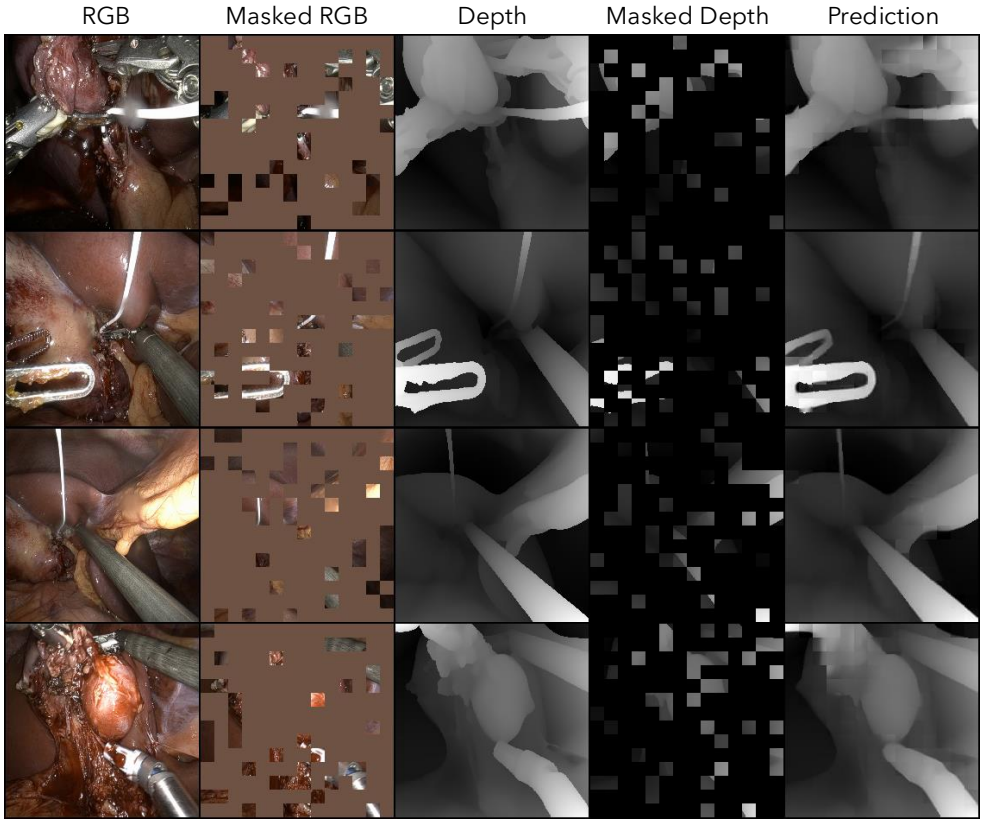

Figure 8: Reconstruction visualization for Mask3D (Hou et al., 2023). The model receives unmasked tokens of both modalities and reconstructs the missing depth patches.

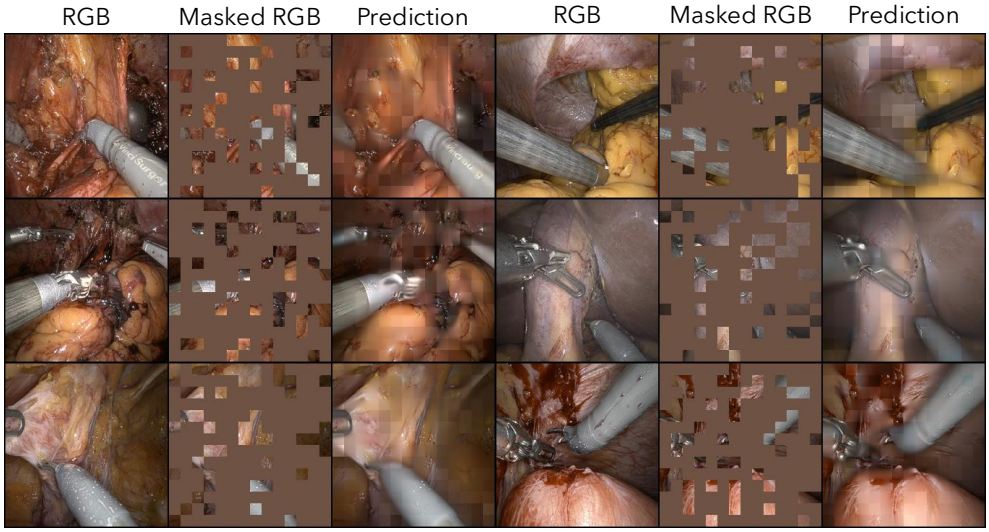

Figure 9: Reconstruction visualization for MAE (Hou et al., 2023). The model receives unmasked tokens of the RGB image and reconstructs the missing patches.

D Correcting the SCARED Dataset

The original authors of the SCARED dataset clarify that the non-keyframe labels are inaccurate (see Section VI. in Allan et al. (2021)) due to calibration and kinematics error. However, only using keyframes is not a viable approach for training depth estimation models since there are only 35 labeled training samples. As a result, there is a need for accurately labeled intermediate frames to expand the dataset.

These issues stem from inaccurate camera poses. If the camera poses are corrected, then these other depth samples can be utilized for training and evaluating depth estimation models. We use COLMAP (Schönberger & Frahm, 2016), a Structure-from-Motion library which estimates camera parameters using SIFT (Lowe, 2004) and bundle adjustment. We run COLMAP with provided camera intrinsics for every video sequence in the dataset to estimate camera extrinsics.

COLMAP estimates camera parameters and a sparse point cloud up to scale, not in metric space. As a result, we devise a simple scale recovery algorithm. We include the keyframe in the image directory prior to running COLMAP to recover the keyframe's unscaled camera pose $\hat{T}_{keyframe}$. We have access to the metric depth map from the structured light sensor provided by the SCARED dataset, D. By projecting the sparse point cloud estimated by COLMAP onto the image plane at $\hat{T}_{keyframe}$, we calculate the unscaled depth map $\hat{D}$. We estimate the scale factor by the median of the ratio between the metric depth map and unscaled depth map:

$$s = \text{median}(D/\hat{D})$$

We use this scale factor to metricize the translation components of the estimated unscaled poses. Let $\hat{T}_i = (R_i, \hat{t}_i)$ denote the scale-ambiguous camera-to-world transformation estimated by COLMAP, where $R_i \in \text{SO}(3)$ and $\hat{t}_i \in \mathbb{R}^3$. Because $\hat{T}_i$ is a camera-to-world transformation, $\hat{t}_i$ represents the camera center. We recover the metric pose T_i by scaling this translation vector:

$$T_i = (R_i, s \cdot \hat{t}_i) \tag{1}$$

We repeat this process for every video sequence. Fig. 10 displays qualitative results of this correction algorithm. We will release the codebase and corrected dataset to the public upon paper acceptance. Table 5 displays the train-validation split we used to conduct experiments.

E Fine-tuning Details

This section provides fine-tuning details for both end-to-end (Table 6) and probing (Table 7). We use the OpenMMLab library to conduct all downstream task experiments. All probing experiments use the last feature map after the last transformer layer of the encoder. We remove class tokens and only use spatial feature maps. We normalize images using the corresponding pre-training dataset statistics during fine-tuning; for instance, we use classic ImageNet dataset statistics $(0.485, 0.456, 0.406) \pm (0.229, 0.224, 0.225)$ to fine-tune MAE (IN-1k), DINOv2 (LVD), and MultiMAE (IN-1k) models. We conduct all fine-tuning experiments at the $(224, 224)$ image resolution.

F Additional Experiments

F.1 Pre-training: FoundationStereo vs. Depth Anything Model

The recently released Depth Anything models (Yang et al., 2024a;c) have demonstrated successful adoption in various surgical computer vision works (Han et al., 2024; Shu et al., 2025). In generating our pre-training RGB-D dataset, we opted to use FoundationStereo (Wen et al., 2025) since the da Vinci systems naturally provide stereo rectified images. However, we show that our approach is also directly applicable for monocular RGB datasets. Specifically, we re-train Mask3D from scratch with the same hyperparameters shown in Table 4 but replace all depth maps with Depth Anything V2 (ViT-L variant) outputs rather than

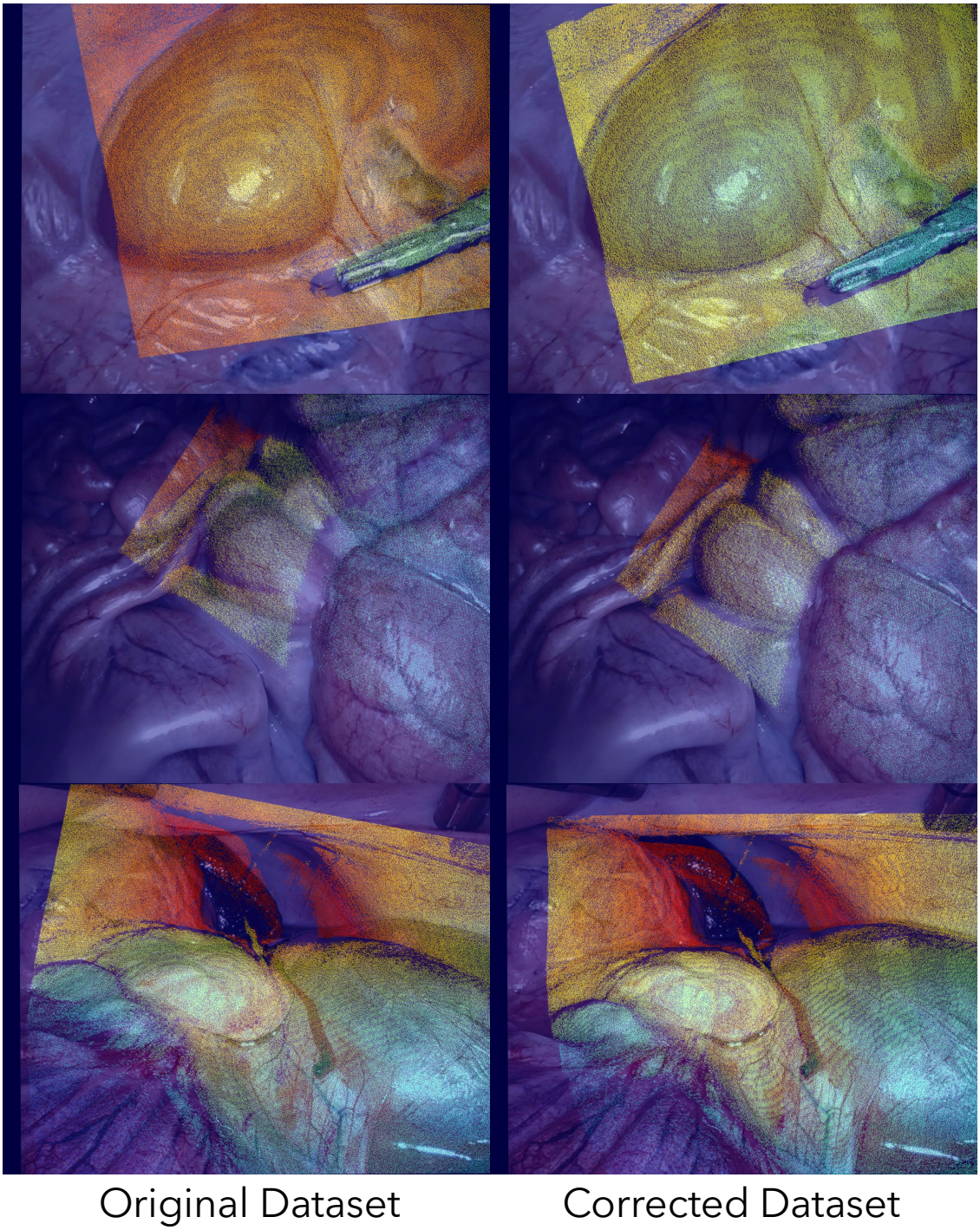

Figure 10: Samples of the original SCARED dataset (left column) and COLMAP-corrected SCARED dataset used in this paper (right column). Note that depth maps are better aligned to the RGB image.

Table 5: Train and validation split for the corrected SCARED dataset. The dataset is split approximately 70-30. The sequence code is {dataset}_{keyframe}.

Train		**Validation**	
Seq	**Frames**	**Seq**	**Frames**
1_1	197	1_2	280
1_3	471	1_4	1
2_2	1,033	1_5	1
2_4	2,114	2_1	88
3_1	329	2_3	1,102
3_2	1,597	2_5	1
3_3	448	3_4	834
6_1	637	3_5	1
6_2	1,087	6_4	1,360
6_3	1,573	6_5	1
7_1	647	7_2	628
7_4	1,548	7_3	584
		7_5	1
Total	**11,681**	**Total**	**4,882**
Grand Total: 16,563			

Table 6: End-to-end fine-tuning training details.

(a) Hyperparameters shared across all 8 datasets and all 4 downstream tasks

Config	**Value**
Epochs	100
Optimizer	AdamW
Learning Rate	1×10^{-4}
Weight Decay	0.05
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.95$
Batch Size	64
Learning Rate Schedule	Cosine Decay
Warmup Epochs	10
Drop Path Rate	0.1
Layer-wise LR decay	0.75

(b) Data augmentations for each task

Config	**Value**
Object Detection	RandomFlip($p = 0.5$)
	MinIoURandomCrop
Pose Estimation	RandomFlip($p = 0.5$)
	RandomBBoxTransform
Segmentation	RandomFlip($p = 0.5$)
	PhotoMetricDistortion
Depth Estimation	RandomFlip($p = 0.5$)

FoundationStereo. Similar to before, we only use the left view and normalize the depth maps to [0,1]. We report linear probing experiments for the segmentation and depth estimation tasks, shown in Table 8.

We observe that there is slightly better performance for FoundationStereo pre-trained models in segmentation whereas Depth Anything pre-trained models have a slight upper hand in depth estimation. This may be because FoundationStereo produces sharper, fine-grained depth maps (making it more conducive for semantic tasks) whereas Depth Anything might be capturing robust relative depth features more useful for depth estimation. Needless to say, we see competitive performance between the two types of models, which suggests that our approach is also applicable for generating multimodal *monocular* surgical datasets.

Table 7: Probing frozen backbone fine-tuning configuration. Data Augmentations are the same as Table 6b.

(a) Hyperparameters for object detection (ViTDet (Li et al., 2022) head) and pose estimation (ViTPose (Xu et al., 2022) head)

Config	Value
Epochs	100
Optimizer	AdamW
Learning Rate	1×10^{-4}
Weight Decay	0.05
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.95$
Batch Size	64
Learning Rate Schedule	Cosine Decay
Warmup Epochs	10

(b) Hyperparameters for segmentation and depth estimation linear probing, which both use a single linear layer after the last feature map.

Config	Value
Epochs	10
Optimizer	AdamW
Learning Rate	1×10^{-3}
Weight Decay	0.05
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.999$
Batch Size	64

Table 8: **Impact of Depth Prior Source in Pre-training.** Comparison of linear probing performance when using Depth Anything vs. FoundationStereo for segmentation and depth estimation. Better results are **bolded**.

Depth Model	**Segmentation** (mIoU ↑)		**Depth Est.** (δ_1 ↑)	
	CholecInst	**EndoVis18**	**DV**	**SCARED**
Depth Anything V2	36.4	25.0	**0.516**	**0.515**
FoundationStereo	**39.7**	**25.2**	0.511	0.510

F.2 Extreme Data Scarcity: When does multimodality perform on par with unimodally pre-trained models?

In 3 out of 4 tasks, we observe that MultiMAE trained with 25% of labeled data exceeds the performance of MAE trained with 100% of the data. At what point will the performances of these models intersect? We experiment with the EndoVis18 dataset and perform end-to-end fine-tuning at increasingly extreme data scarcity. The results are shown in Fig. 11.

We observe that MultiMAE pre-trained with surgical multimodal data remains competitive even at extreme data scarcity, where at *2.5%* of the training data, MultiMAE *still* outperforms the MAE baseline (24.9 vs. 23.5 mIoU). Meanwhile, we see the performance of MultiMAE (IN-1k) rapidly deteriorating in performance, suggesting that surgical multimodal models are more data efficient than its ImageNet counterparts.

F.3 Additional Task-specific Heads

Table 9 examines whether the choice of task-specific head influences the relative ranking of pre-trained backbones. We compare the heads we used previously with other common heads in the literature: Mask-RCNN (He et al., 2017) vs. Faster-RCNN (Ren et al., 2016) for object detection, ViTPose (Xu et al., 2022) vs. RTMPose (Jiang et al., 2023) for pose estimation, Linear vs. Segmenter (Strudel et al., 2021) for segmentation, and Linear vs. DPT (Ranftl et al., 2021) for depth estimation. We perform these experiments under the frozen backbone setting.

For object detection, Mask R-CNN and Faster R-CNN yield consistent rankings, with MultiMAE variants achieving the highest performance under both heads. However, DINOv2 (LVD) exhibits a notable discrepancy, dropping from 26.2 mAP with Mask R-CNN to 9.4 mAP with Faster R-CNN, suggesting that features pre-trained on natural images may be more sensitive to head architecture when transferred to surgical do-

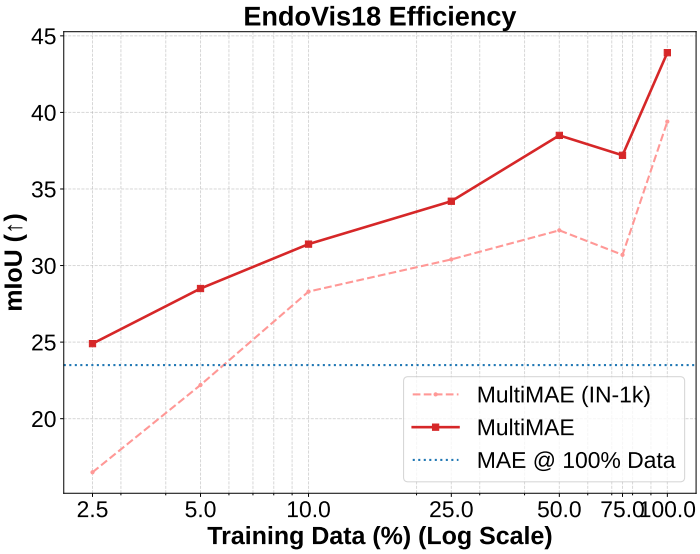

Figure 11: Extreme data scarcity experiments with EndoVis18 for MultiMAE and MultiMAE (IN-1k). The blue dotted line represents MAE fine-tuned with 100% of the data. EndoVis18 has a total of 2,235 training images; thus, the training dataset is only 55 images at 2.5% of the training dataset.

mains. For segmentation, Segmenter outperforms the linear head in seven of eight cases, often by a substantial margin (e.g., 17.5 vs. 10.0 mIoU for DINOv2-RGB). The sole exception is MAE (IN-1k), where the linear head achieves 10.3 mIoU compared to Segmenter's 5.5 mIoU. This anomaly may indicate that MAE features pre-trained on natural images, without domain adaptation, are not well-suited to Segmenter's decoder architecture in the surgical domain. Despite the generally higher absolute performance with Segmenter, the relative ranking of backbones remains consistent: MultiMAE variants achieve the best results under both heads. For pose estimation, ViTPose substantially outperforms RTMPose across all backbones, with the performance gap particularly pronounced for models pre-trained without geometric information (e.g., 32.2 vs. 7.8 mAP for MAE (IN-1k)). This suggests that RTMPose's lightweight design may require higher-quality spatial features to function effectively. For depth estimation, both linear and DPT heads preserve the overall model ranking, with MultiMAE and Mask3D consistently outperforming unimodal baselines, though DPT provides a uniform improvement of approximately 0.1–0.15 δ_1 across all models.

Across all tasks, the key finding is that the relative advantage of multimodal pre-training holds regardless of head architecture. While absolute performance varies with head choice, MultiMAE consistently achieves the best or second-best results under both simple and complex heads. This suggests that the benefits of geometry-aware pre-training stem from the learned representations themselves rather than favorable interactions with specific decoder architectures.

G Quantitative Results

We provide the quantitative results of fine-tuning our models across the eight datasets below. Table 10 displays results when probing a frozen backbone and Table 11 displays results when end-to-end fine-tuning. We also include the numerical values for the data efficiency experiments in Table 12.

H Qualitative Results on Downstream Tasks

We show qualitative results across the 8 datasets and four tasks; object detection in Fig. 12, segmentation in Fig. 13, depth estimation in Fig. 14, and pose estimation in Fig. 15..

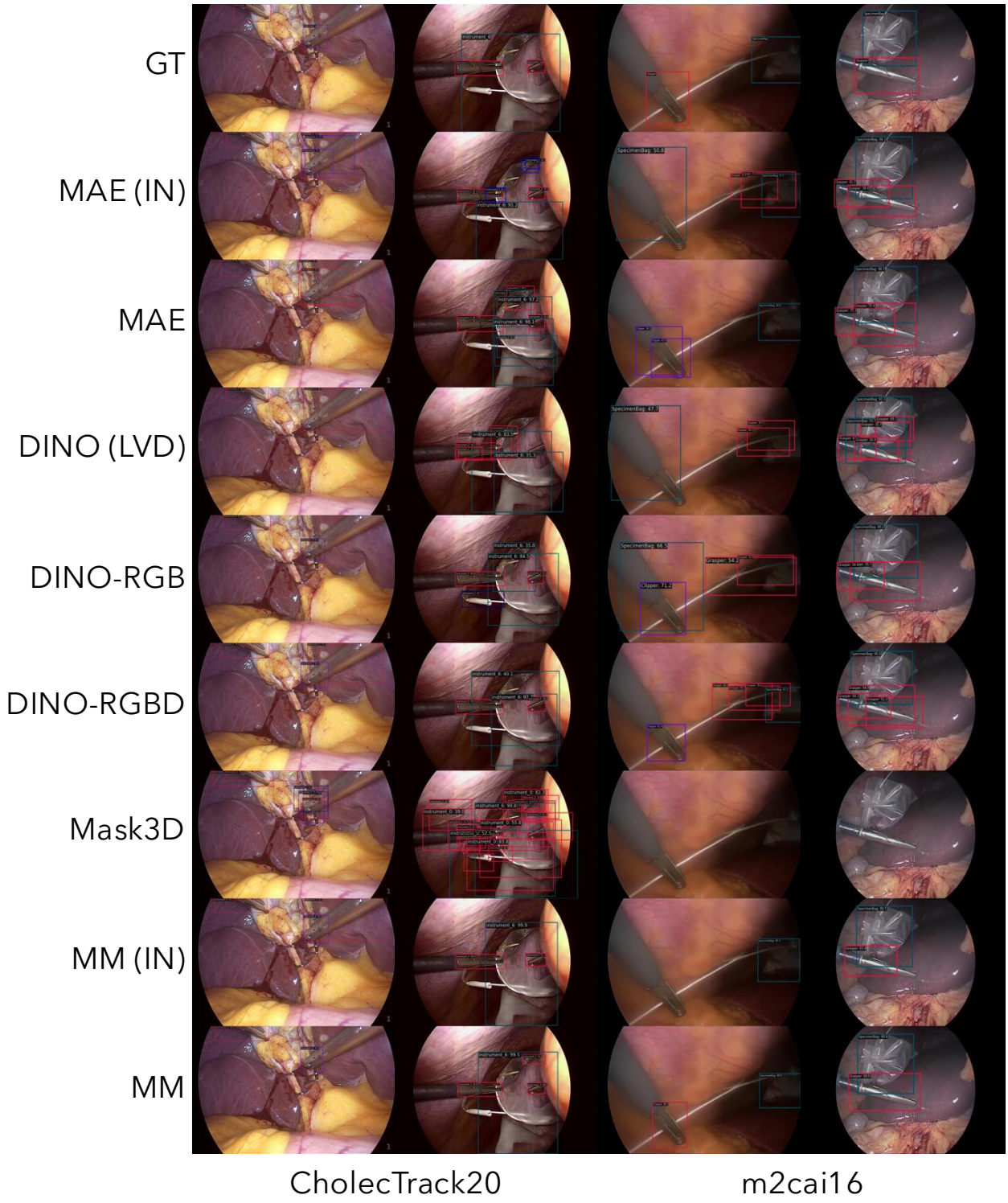

Figure 12: Qualitative results on object detection datasets. GT is ground truth, IN is ImageNet-1k, and MM refers to MultiMAE.

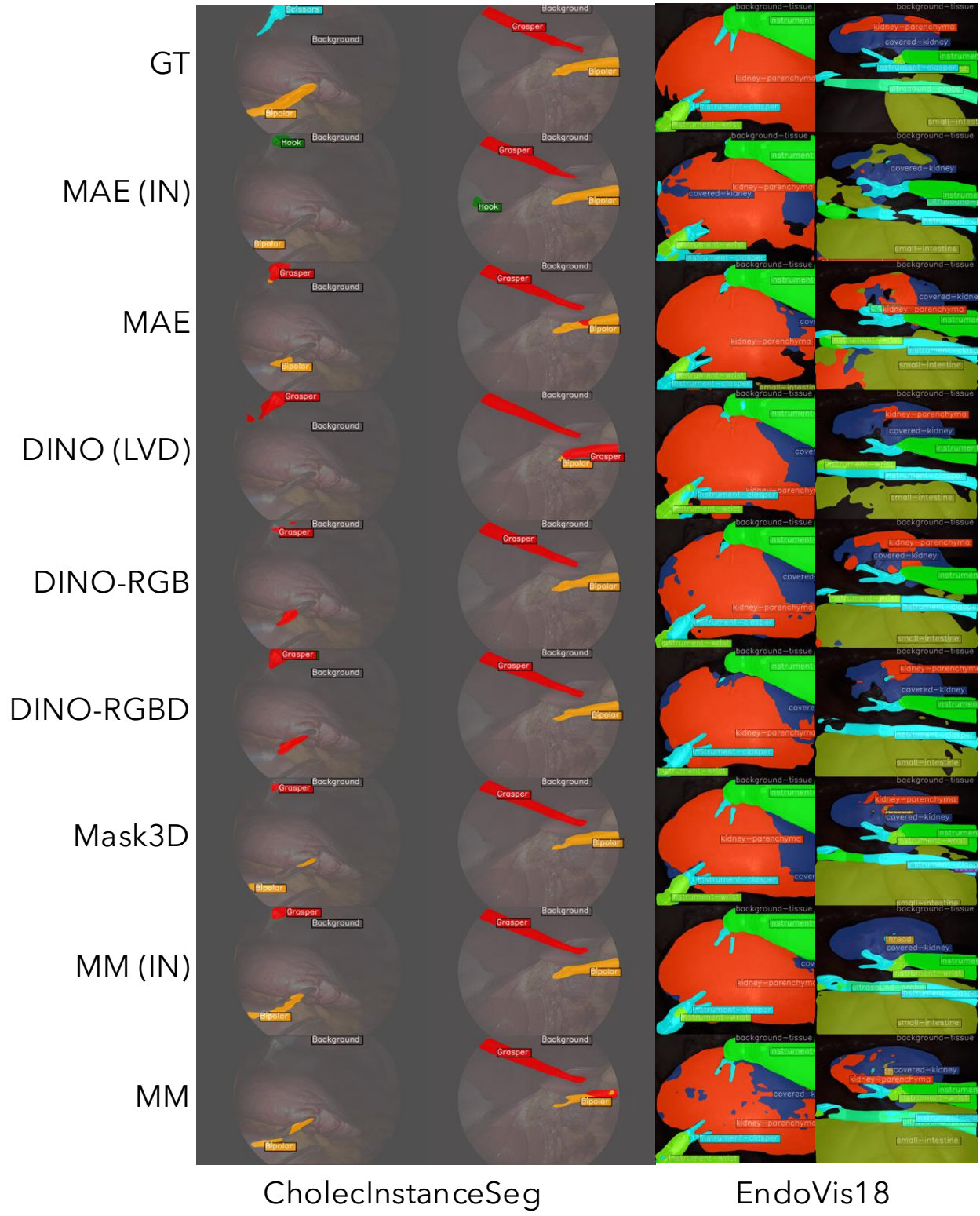

Figure 13: Qualitative results on segmentation datasets. GT is ground truth, IN is ImageNet-1k, and MM refers to MultiMAE.

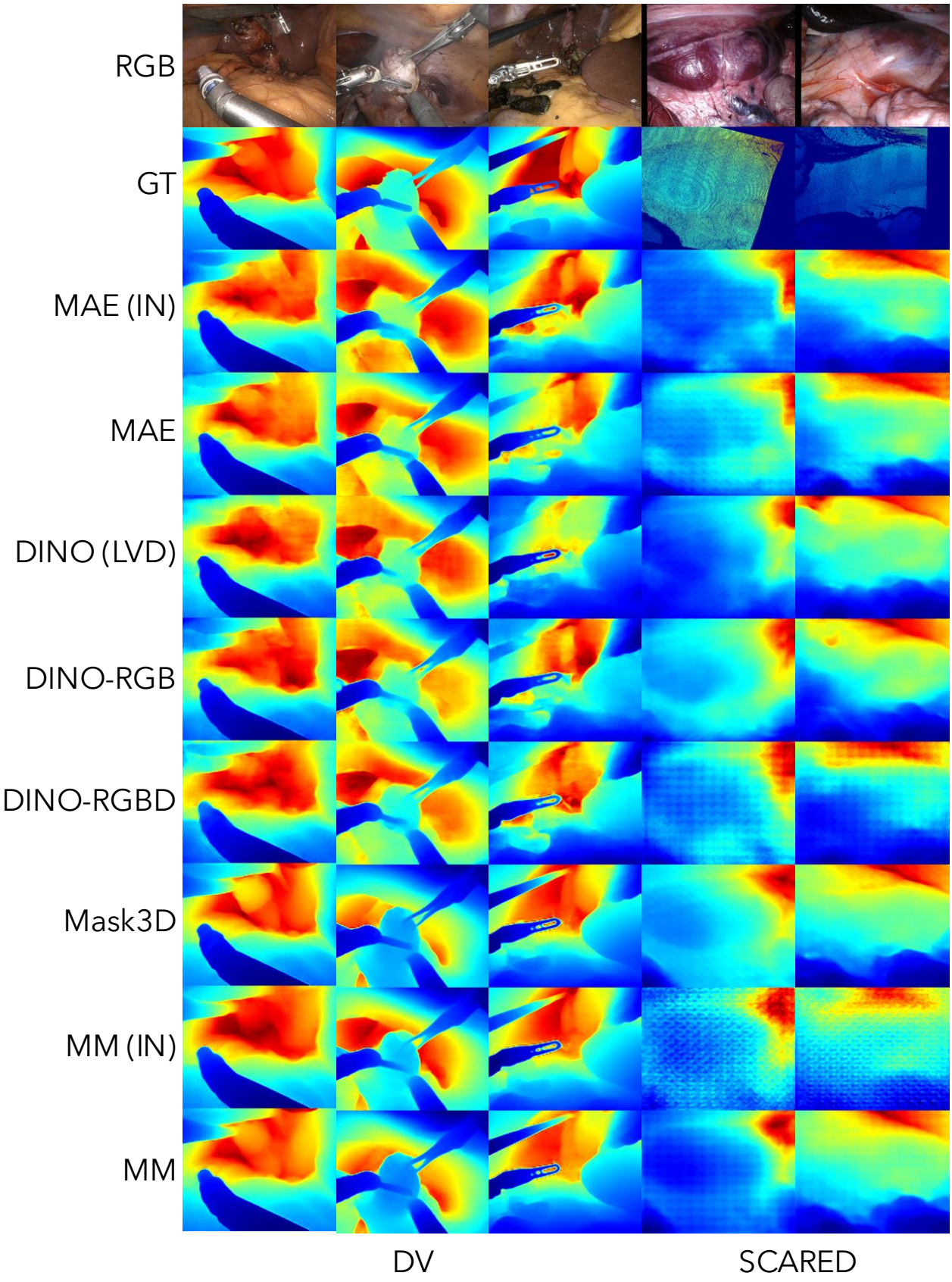

Figure 14: Qualitative results on depth estimation datasets. GT is ground truth, IN is ImageNet-1k, and MM refers to MultiMAE.

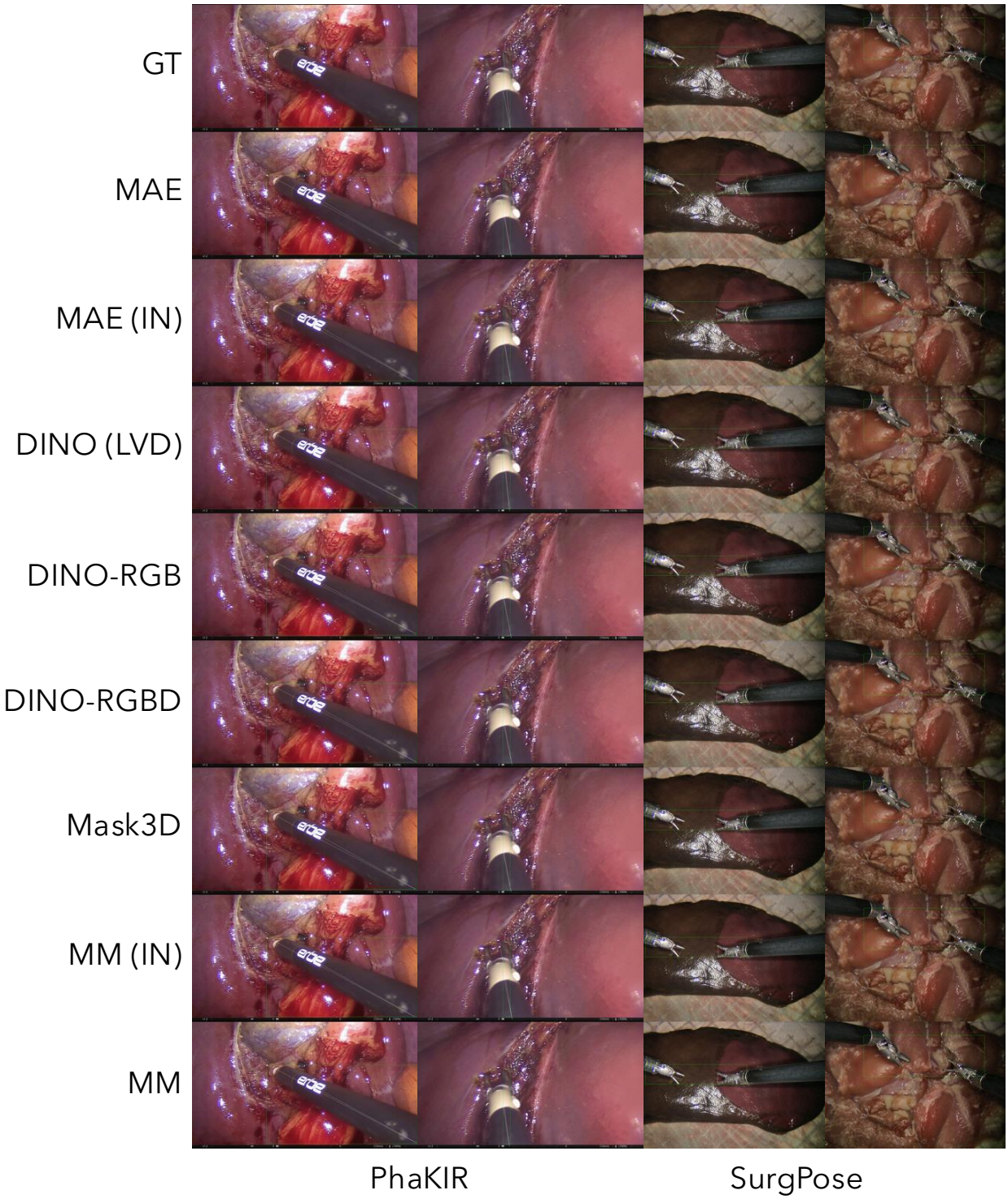

Figure 15: Qualitative results on pose estimation datasets. GT is ground truth, IN is ImageNet-1k, and MM refers to MultiMAE. Please zoom in to view the keypoints.

Table 9: **Frozen Backbone Evaluation (Probing) with different task-specific heads.** Best results are **bolded**, second best are underlined.

Model	**Object Detection** (mAP ↑) **CholecTrack20**		**Segmentation** (mIoU ↑) **EndoVis18**		**Pose Est.** (mAP ↑) **SurgPose**		**Depth Est.** (δ_1 ↑) **DV**	
	Mask-RCNN	Faster-RCNN	Linear	Segmenter	ViTPose	RTMPose	Linear	DPT
MAE (IN-1k)	25.4	25.3	10.3	5.5	32.2	7.8	0.276	0.422
MAE	23.4	24.1	10.2	15.3	30.5	8.4	0.283	0.436
DINOv2 (LVD)	26.2	9.4	9.2	14.4	16.6	3.8	0.278	0.417
DINOv2-RGB	25.7	25.5	10.0	17.5	32.1	9.4	0.287	0.436
DINOv2-RGBD	25.9	26.2	10.3	18.3	31.3	7.5	0.294	0.438
Mask3D	6.5	6.4	25.2	32.7	31.1	8.8	0.511	0.686
MultiMAE (IN-1k)	34.7	34.8	26.8	31.7	**48.8**	31.4	0.390	0.567
MultiMAE	**35.2**	**35.4**	**28.2**	**33.4**	47.1	**31.5**	**0.590**	**0.705**

Table 10: **Frozen Backbone Evaluation (Probing).** Comparison of feature readiness across 8 surgical downstream tasks. Best results are **bolded**, second best are underlined.

Model	**Object Detection** (mAP ↑)		**Segmentation** (mIoU ↑)		**Pose Est.** (mAP ↑)		**Depth Est.** (δ_1 ↑)	
	CholecTrack20	**m2cai16**	**CholecInst**	**EndoVis18**	**PhaKIR**	**SurgPose**	**DV**	**SCARED**
MAE (IN-1k)	25.4	38.4	21.1	10.3	2.1	32.2	0.276	0.414
MAE	23.4	37.7	18.3	10.2	3.2	30.5	0.283	0.408
DINOv2 (LVD)	26.2	18.8	19.0	9.2	1.5	16.6	0.278	0.444
DINOv2-RGB	25.7	41.0	20.5	10.0	1.9	32.1	0.287	0.447
DINOv2-RGBD	25.9	42.3	21.6	10.3	5.3	31.3	0.294	0.455
Mask3D	6.5	6.3	39.7	25.2	20.0	31.1	0.511	0.510
MultiMAE (IN-1k)	34.7	52.2	**40.7**	26.8	**23.1**	**48.8**	0.390	0.487
MultiMAE	**35.2**	**52.5**	40.0	**28.2**	17.7	47.1	**0.590**	**0.517**

Table 11: **End-to-End Fine-tuning Evaluation.** Comparison of maximum model capacity when all weights are updated. Best results are **bolded**, second best are underlined.

Model	**Object Detection** (mAP ↑)		**Segmentation** (mIoU ↑)		**Pose Est.** (mAP ↑)		**Depth Est.** (δ_1 ↑)	
	CholecTrack20	**m2cai16**	**CholecInst**	**EndoVis18**	**PhaKIR**	**SurgPose**	**DV**	**SCARED**
MAE (IN-1k)	30.1	43.4	55.9	21.1	28.5	43.6	0.521	0.577
MAE	30.1	44.9	51.7	23.5	30.9	48.1	0.524	0.526
DINOv2 (LVD)	22.0	33.0	60.1	25.8	30.0	47.9	0.513	0.523
DINOv2-RGB	30.1	45.5	62.8	26.2	32.5	51.4	0.525	0.561
DINOv2-RGBD	32.4	46.6	63.5	26.9	33.0	49.4	0.526	0.545
Mask3D	18.8	6.6	67.5	40.6	36.9	46.3	0.709	0.618
MultiMAE (IN-1k)	35.4	51.0	64.2	39.4	37.5	51.4	0.662	0.614
MultiMAE	**37.2**	**53.9**	**68.0**	**43.9**	**39.5**	**69.7**	**0.758**	**0.624**

Table 12: **Data Efficiency Analysis.** Comparison of model performance across four downstream tasks when trained with restricted training data (25%, 50%, 75%) versus the full dataset (100%). Best results are **bolded**.

(a) Object Detection (CholecTrack20) – mAP ↑

Model	**25%**	**50%**	**75%**	**100%**
MAE (IN-1k)	20.0	23.8	26.2	25.4
MAE (Surg)	20.5	24.1	26.3	29.6
MultiMAE (IN-1k)	22.8	26.7	29.5	35.4
MultiMAE (Surg)	**26.7**	**30.2**	**34.2**	**37.2**

(b) Instance Segmentation (EndoVis18) – mIoU ↑

Model	**25%**	**50%**	**75%**	**100%**
MAE (IN-1k)	15.3	15.7	16.6	21.1
MAE (Surg)	16.1	19.4	21.0	23.5
MultiMAE (IN-1k)	30.4	32.3	30.7	39.4
MultiMAE (Surg)	**34.2**	**38.5**	**37.2**	**43.9**

(c) Pose Estimation (SurgPose) – mAP ↑

Model	**25%**	**50%**	**75%**	**100%**
MAE (IN-1k)	38.6	39.4	42.6	43.6
MAE (Surg)	38.6	41.1	42.1	48.1
MultiMAE (IN-1k)	47.0	49.4	50.3	51.4
MultiMAE (Surg)	**62.3**	**66.4**	**64.3**	**67.6**

(d) Depth Estimation (DV) – δ_1 ↑

Model	**25%**	**50%**	**75%**	**100%**
MAE (IN-1k)	0.453	0.479	0.503	0.521
MAE (Surg)	0.457	0.481	0.508	0.524
MultiMAE (IN-1k)	0.606	0.638	0.653	0.662
MultiMAE (Surg)	**0.736**	**0.748**	**0.754**	**0.758**