

# Learning Surgical Robotic Manipulation with 3D Spatial Priors

## Supplementary Material

### A. Readme.

We provide detailed videos in the supplementary zip file to help reviewers better understand our work:

- **Visualization.mp4:** A detailed visualization of the surgical scene reconstructed by our finetuned geometry transformer throughout task execution, as well as the surgical manipulation performed by our visuomotor policy.
- **Hardware.mp4:** An overview of the hardware setup used in our experiments, including the surgical robot system, computing devices, and wrist-mounted cameras.

### B. Quantitative Comparison of 3D Reconstruction.

Since there is no accessible 3D ground-truth for real surgical scenes, and the primary objective of our method is to learn robust 3D latent embeddings for policy learning rather than achieving superior geometric reconstruction, we do not include the quantitative comparison in the main manuscript.

#### B.1. Datasets and Metrics

To evaluate reconstruction quality, we construct a synthetic test set in NVIDIA Omniverse. We place open-source surgical assets and the PSM instruments used in our setup at various arbitrary poses, and render the corresponding endoscopic images with ground-truth depth. The test set consists of 300 pairs of endoscopic stereo images with accurate depth annotations.

We evaluate four methods: VGGT [44], DUS3R [46], MAST3R [18], and our finetuned MAST3R. Following previous work [18, 44, 46], reconstruction quality is assessed from two perspectives: (1) depth estimation and (2) multi-view 3D reconstruction.

For depth estimation, we compare the predicted depth maps of the left-view images. For VGGT, we directly use the output from its depth estimation head, while for the other methods we derive depth from the  $z$ -coordinates of their predicted point maps. We report the Absolute Relative Error (Rel), computed as the average of  $|d_{\text{pred}} - d_{\text{gt}}|/d_{\text{gt}}$  over all pixels, and the  $\delta$  metric, which measures the percentage of pixels for which the depth prediction falls within a threshold of 1.25 relative difference from the ground truth, and the Root Mean Squared Error (RMSE). All metrics are averaged over the full test set.

For multi-view 3D reconstruction, we use the predicted point maps from each method and evaluate averaged accuracy, averaged completeness, and overall error. Accuracy is defined as the minimum Euclidean distance from each predicted point to the ground-truth surface, while completeness

Methods	Depth Estimation		
	AbsRel. ↓	$\delta_{1.25}$ ↑	RMSE ↓
DUS3R	0.2776	0.5176	0.0781
MASt3R	0.3946	0.3726	0.0974
VGGT	<u>0.2313</u>	<u>0.5779</u>	<u>0.0673</u>
MASt3R (finetuned)	<b>0.1987</b>	<b>0.7161</b>	<b>0.0558</b>

Table 4. **Quantitative Comparison of Depth Estimation.** The best results for each setting are in **bold**, and the second best results are in underline.

Methods	Multi-view Reconstruction		
	Acc. ↓	Comp. ↓	Overall ↓
DUS3R	0.0111	0.0151	0.0131
MASt3R	0.0111	0.0140	0.0126
VGGT	<u>0.0097</u>	<u>0.0098</u>	<u>0.0098</u>
MASt3R (finetuned)	<b>0.0048</b>	<b>0.064</b>	<b>0.0056</b>

Table 5. **Quantitative Comparison of Multi-view Reconstruction.** The best results for each setting are in **bold**, and the second best results are in underline.

is defined as the minimum Euclidean distance from each ground-truth point to the reconstructed surface. The overall error is computed as the mean of them.

#### B.2. Results and Analysis

As shown in Tab. 4 and Tab. 5, finetuning MAST3R on the Surgical3D dataset significantly improves reconstruction quality in surgical scenes, and achieves the best performance across all evaluated methods. Our finetuned model outperforms MAST3R (without finetuning) by **88.5%** in depth accuracy and **125%** in multi-view reconstruction across all metrics. These results highlight the necessity of our proposed Surgical3D dataset in addressing the large domain gap between generic scenes and surgical environments. The multi-view reconstruction errors appear relatively small because surgical scenes are highly compact, with very short distances between the instruments and surrounding tissues. This also reflects that such extremely small working distances exceed the effective perception range of most existing 3D sensors. **The “Visualization.mp4” in the supplementary material** provides detailed reconstruction visualizations throughout the entire surgical task execution, as well as visualizations of our surgical robot performing the learned policies to complete each task.

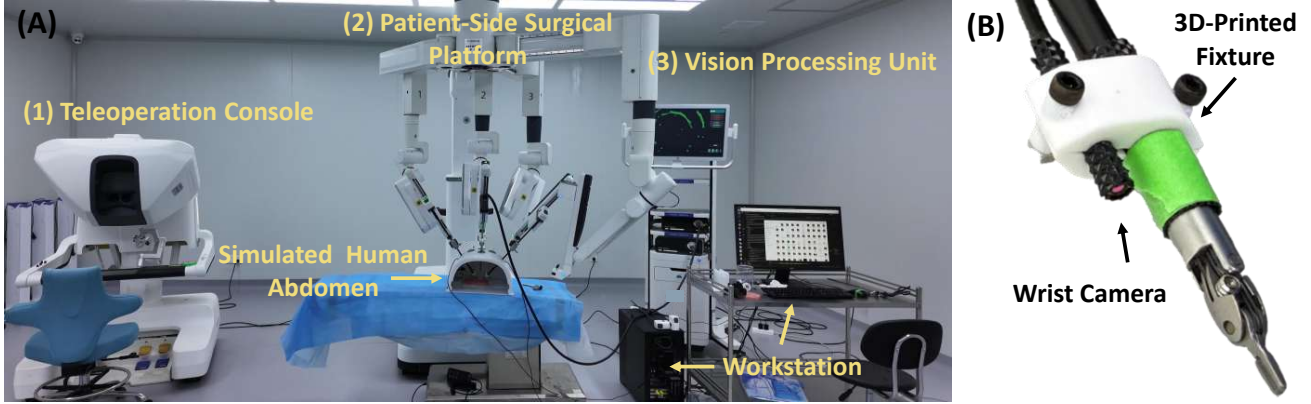


Figure 7. **Details of Hardware Used in Our Paper.** (A) The robotic experimental platform used for demonstration collection and method evaluation. (B) The wrist-mounted camera configuration we used for SRT baseline.

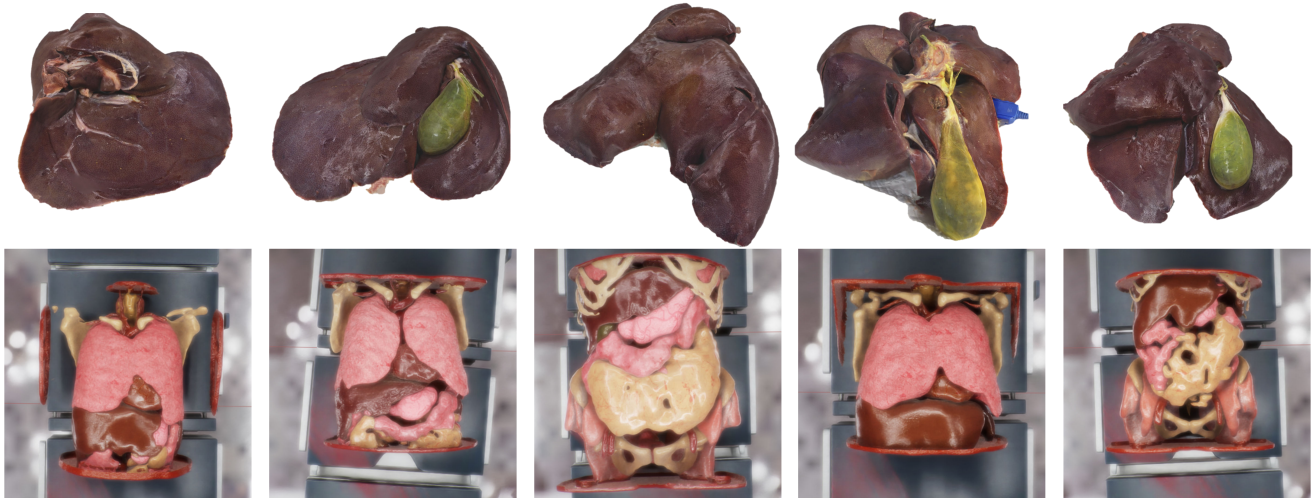


Figure 8. **Examples of 3D Assets Used for Constructing The Surgical3D Dataset.**

## C. Hardware Details.

### C.1. Details of Surgical Robot System.

The complete robotic surgical system used in this work is illustrated in Fig. 7 (A). It consists of three core components: (1) a teleoperation console that allows the surgeon to command robot actions, (2) a patient-side surgical platform that physically executes the manipulation, and (3) a vision processing unit that handles the visual perception of the endoscopic camera manipulator (ECM). Conventional robotic surgery follows a master–slave paradigm: the surgeon perceives the intra-abdominal scene through the console and continuously plans and executes motions, while the patient-side platform mirrors every commanded action. The surgeon remains fully in the control loop to correct motion and ensure surgical precision. To be precise, our Spatial Surgical Transformer (SST) is designed to offload a subset of low-level manipulation skills from the surgeon, enabling

the patient-side surgical platform to autonomously execute well-defined subtasks under algorithmic control.

### C.2. Details of Experiments Setting.

We use a workstation equipped with an RTX-3090 GPU is connected to the system via a local network and is responsible for demonstration collection and policy inference deployment. During the experiments, the initial poses of the patient-side manipulators (PSMs) and the endoscopic camera manipulator (ECM) are kept approximately same at the configuration shown in Fig. 7 (A), while a dome simulating the human abdomen is positioned following the setup used in SRT [16]. To support the SRT baseline, we attach compact wrist-mounted cameras to the PSMs using a 3D-printed fixture, as illustrated in Fig. 7 (B). The “Hardware.mp4” in supplementary provides a clear demonstration of the practical limitations that prevent wrist-mounted cameras from being widely adopted in clinical settings, thereby of-

fering a more intuitive understanding of the motivation behind our work.

## D. Model Hyperparameters

**Spatial Surgical Transformer (SST).** We initialize the geometry transformer in SST using MAST3R [18] weights. We finetune all parameters on the Surgical3D dataset using an 8×A100-40GB server with a batch size of 2 and a learning rate of  $1 \times 10^{-5}$ . During the policy training stage, we freeze the geometry transformer and train the policy components on an 8×RTX 4090 server with a batch size of 48 and a learning rate of  $1 \times 10^{-4}$ . We adopt the Adam optimizer with a weight decay of  $1 \times 10^{-4}$ . The input image resolution is set to  $288 \times 512$  to match the geometry transformer’s training configuration.

**SRT and ACT.** SRT [16] adopts a model architecture similar to ACT [56], where a ResNet18 network serves as the visual encoder. The policy head consists of a 4-layer transformer encoder followed by 7 transformer decoder layers. The learning rates for both the vision backbone and the policy head are set to  $1 \times 10^{-5}$ , and we use Adam with a weight decay of  $1 \times 10^{-4}$ . The input image resolution is also  $288 \times 512$  to ensure fair comparison with SST.

**Diffusion Policy (DP).** For the diffusion policy [7], we use a ResNet18 network as the visual backbone and a U-Net architecture as the noise predictor. We adopt the Adam optimizer with a learning rate of  $1 \times 10^{-4}$  and a weight decay of  $1 \times 10^{-6}$ . The denoising process follows the DDPM scheduler. The input image resolution is set to  $288 \times 512$  for consistency with the other methods.

## E. More Details of Surgical3D.

Fig. 8 illustrates the assets used for constructing the Surgical3D dataset. The assets in the top row are reconstructed from real pig organs, including the liver and gallbladder, captured using an iPad. These anatomical samples come from different pigs, resulting in variations in size and geometry. They are placed in diverse poses to ensure sufficient variability within the dataset. The bottom row presents assets obtained from open-source anatomical models. We remove the outer skin layer to expose internal organs and further enhance scene diversity by randomly omitting organs or modifying body poses.