Surgical Procedural Planning as 3D World Modelling: Towards Automated Pulmonary Resection

Supplementary Material

6. Detailed Experimental Setup and Hyperparameters

This section provides comprehensive details for the implementation of our 3D medical world model to ensure full reproducibility.

6.1. 3D MedGemma Architecture and Training

3D Vision Encoder: We replaced the vision encoder in MedGemma 4B [16] with a custom 3D Vision Encoder. The encoder processes input CT volumes of size $192 \times 192 \times 192$ voxels. It first uses a 3D Patch Embedding layer (kernel size $16 \times 16 \times 16$, stride 16) to divide the volume into non-overlapping 3D patches, resulting in a sequence of patch embeddings. A 3D sine-cosine positional embedding is added to these patch embeddings to provide spatial information. To standardize the output sequence length for compatibility with the downstream SigLIP encoder (which expects a fixed number of image tokens), the sequence of patch embeddings is transposed and passed through an AvgPool layer, reducing the number of patches to a fixed target of 256. The resulting sequence of 256 embeddings is then passed to the original SigLIP-based Transformer encoder component of MedGemma via the inputs_embeds parameter. This design allows the 3D Vision Transformer to act as a specialized patch embedding and initial feature processing module, feeding its output into the established SigLIP encoder architecture.

Text Encoder: The text encoder is initialized from the MedGemma language model and is frozen during both pre-training and fine-tuning, as per our LoRA strategy.

Multi-modal Projector: A 2-layer MLP with hidden dimension 2048 and GeLU activation connects the processed 3D visual embeddings (output from the 3D Vision Encoder) and the text embeddings to a shared 2560-dimension latent space.

LoRA Adaptation: For fine-tuning on the Surgical History Dataset, we applied LoRA [8] to the attention weights of the Gemma decoder. The low-rank factor r was set to 16 for all attention layers. Only the LoRA matrices A and B.

Pre-training: Conducted on the Chest CT-Report Dataset (1,000 pairs) using the contrastive loss L_{cont} (Eq. 1 in main text). Optimizer: AdamW, Learning Rate: 5×10^{-5}, Batch Size: 32, Temperature τ: 0.07, Epochs: 200.

Fine-tuning: Conducted on the Surgical History Dataset (150 cases for training and 50 cases for test, formatted as VQA). Optimizer: AdamW, Learning Rate: 1×10^{-4} (for LoRA/projector), Batch Size: 8, Epochs: 100.

Table S1. Summary of 3D MedGemma architecture and training.

Component	Specification
3D Vision Encoder	Patch size 16^3, stride 16; Sine-cosine pos. embed; AvgPool to 256 tokens; Feeds SigLIP Transformer
Text Encoder	Frozen MedGemma LLM
Multi-modal Projector	2-layer MLP, 2048 hidden, GeLU; Output dim: 2560
LoRA Adaptation	Applied to Gemma decoder attn.; Rank $r = 16$, only A/B trained
Pre-training	1K CT-Report pairs; L_{cont}, $\tau = 0.07$; AdamW, LR 5×10^{-5}, B=32, 200 epochs
Fine-tuning	150 surgical cases (VQA); AdamW, LR 1×10^{-4}, B=8, 100 epochs

The configuration of the 3D Vision Encoder is detailed in Table S1.

6.2. 3D Conditional Diffusion Model

Architecture: Based on a 3D U-Net with 4 downsampling and 4 upsampling blocks. The base number of channels is 32, doubling after each downsampling. Each block contains two 3×3×3 convolutions with GroupNorm and SiLU activation, incorporating time and action embeddings via linear projections added to feature maps.

Anatomical Encoder (f_{enc}): A lightweight 3D CNN with 4 convolutional blocks (kernel $3 \times 3 \times 3$, stride 2) to encode the preoperative image into multi-scale feature maps at resolutions [D, H, W], [D/2, H/2, W/2], [D/4, H/4, W/4], and [D/8, H/8, W/8], which are fused into the U-Net's encoder layers through concatenation.

Action Embedder ($MLP(e_{action})$): Uses sinusoidal position embeddings followed by a 2-layer MLP to project discrete surgical actions into 256-dimensional vectors.

Diffusion Process: Employs a linear noise schedule from $\alpha_1 = 0.999$ to $\alpha_T = 0.0001$ over $T = 1000$ timesteps. Time steps are embedded using 256-dimensional sinusoidal embeddings.

Training: Optimized using the standard denoising objec-

Algorithm S1 The Perceive–Simulate–Decide Loop for Surgical Planning

Require: Preoperative CT volume X, Text prompt p, Initial state s_0, Feasibility threshold τ_{acc}, Safety threshold τ_{ris}, Beam width k, Max depth T
Ensure: Optimal surgical plan $\pi^* = [a_1^*, a_2^*, ..., a_L^*]$

Initialize beam: $\mathcal{B}_0 \leftarrow \{(\text{empty\_plan}, s_0)\}$
$\mathcal{P}_{\text{candidates}} \leftarrow \emptyset$ ▷ Candidate complete plans
for $t = 1$ to T **do** ▷ Step-by-step planning
  $\mathcal{B}_t \leftarrow \emptyset$ ▷ Next beam
  for all $(\pi, s) \in \mathcal{B}_{t-1}$ **do** ▷ Expand each partial plan
    Perceive: Generate candidate actions $\mathcal{A}_c \sim$ 3D MedGemma(X, p, s)
    for all $a_c \in \mathcal{A}_c$ **do**
      Simulate: Predict resection target $M_{target} \leftarrow \text{DiffusionModel}(X, a_c)$
      Update State: $s' \leftarrow f_{\text{state}}(s, a_c, M_{target})$ ▷ Apply action to get new state
      Evaluate:
        $\alpha(x) \leftarrow \text{AccessibilityScore}(s', a_c)$ ▷ Eq. 7
        $\sigma \leftarrow \text{SafetyScore}(s', a_c)$ ▷ Eq. 8
      if $\alpha(x) \geq \tau_{acc}$ **and** $\sigma \leq \tau_{ris}$ **then** ▷ Feasible and Safe?
        Append $(\pi + [a_c], s')$ to $\mathcal{B}_t$
  if $\mathcal{B}_t = \emptyset$ **then** ▷ No valid actions at step t
    break ▷ Terminate early
  Keep top-k candidates in $\mathcal{B}_t$ based on cumulative reward or prior probability
  Add any complete plans from $\mathcal{B}_t$ to $\mathcal{P}_{\text{candidates}}$
if $\mathcal{P}_{\text{candidates}} = \emptyset$ **then**
  return Failure or fallback plan
else
  Global Optimization: Score all candidates in $\mathcal{P}_{\text{candidates}}$
  $R(\pi) \leftarrow \lambda_1 R_{res}(\pi) + \lambda_2 R_{fun}(\pi)$ ▷ Eq. 9
  $\pi^* \leftarrow arg\,max_{\pi \in \mathcal{P}_{\text{candidates}}} R(\pi)$
  return π^*

tive L_{diff}. Optimizer: AdamW, Learning Rate: 2×10^{-4}, Batch Size: 4, Epochs: 300.

6.3. Decision Engine and Optimization

Local Feasibility Threshold: $\tau_{acc} = 0.7$. Actions with $\alpha(x) < \tau_{\text{acc}}$ are deemed infeasible and replaced with newly sampled candidates.

Safety Threshold: $\tau_{ris} = 0.2$. Actions with $\sigma > \tau_{\text{ris}}$ are considered unsafe and replaced with newly sampled candidates.

Global Reward Weights: The weights λ_1 and λ_2 in Eq. 9 were calibrated through a retrospective analysis of historical surgical plans and expert consensus. We found $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$ to provide the best balance between oncological safety (tumor clearance) and pulmonary function preservation.

Beam Search: For global plan selection, we used a beam width of 5. The search depth was limited to 10 steps to ensure computational tractability.

Algorithmic Workflow: The complete algorithmic workflow of our decision engine, integrating perception, simulation, and step-wise evaluation within a beam search framework, is summarized in Algorithm S1. This pseudocode provides a concise, formal representation of the end-to-end planning process described in this section.

7. Expert-Derived Clinical Risk Factors and Weights (r_i and w_i)

This section details the clinical risk factors r_i and their corresponding expert-derived weights w_i used in the Safety Scoring function (Eq. 8 in the main text). As shown in Table S2, these weights were derived from a survey of senior thoracic surgeons, who were asked to rank the relative severity of 5 common procedural violations during segmental/subsegmental resection. The weights are normalized for use in the scoring function.

The safety score σ is calculated as $\sigma = 1 - \exp(-\sum_{i=1}^{5} w_i \cdot r_i)$. A risk factor with a higher w_i has a more significant impact on the final safety score. The specific definitions and calculation methods for each r_i are de-

Table S2. Clinical Risk Factors (r_i), Their Calculation Methods, and Expert-Derived Weights (w_i)

ID	Risk Factor (r_i)	Calculation Method	Weight (w_i)
R1	Vascular Injury Risk	If the action targets an artery or vein that is not part of the target segment's supply/drainage, then $r_1 = 1$. If it targets a common trunk (e.g., A4+5) shared by multiple segments, then $r_1 = 0.5$. Otherwise, $r_1 = 0$.	0.5
R2	Sequence Violation Risk	If the action involves transecting a vein before its corresponding artery within the target segment, then $r_2 = 1$. Otherwise, $r_2 = 0$.	0.4
R3	Anatomical Variation Sensitivity	If the target structure belongs to a predefined list of high-variation anatomical regions (e.g., right middle lobe vein, lingual bronchus), then $r_3 = 0.6$. Otherwise, $r_3 = 0$.	0.3
R4	Functional Unit Misresection Risk	If the action leads to the interruption of blood supply or ventilation to a non-target healthy lung segment, then r_4 represents the ratio of the volume of resected non-target tissue to the total volume of the affected segment. Otherwise, $r_4 = 0$.	0.3
R5	Instrument Accessibility Deficiency	If the target structure is in a position difficult for safe instrument manipulation, then $r_5 = 1 - \alpha(x)$, where $\alpha(x)$ is the accessibility score defined in Eq. 7. Otherwise, $r_5 = 0$.	0.2

tailed below.

The values of r_i are computed based on the specific conditions described above for each risk factor. The sum $\sum_{i=1}^{5} w_i \cdot r_i$ represents the total weighted risk penalty for a given surgical action. A higher value indicates a less safe action.