Background and Related Works

Multimodal large language models have demonstrated unprecedented capabilities in bridging visual and textual modalities through sophisticated attention mechanisms. The transformer architecture with cross-modal attention can be formulated as $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V$, where M represents a learned modality-specific bias matrix that enables effective fusion of heterogeneous information sources. However, the application of MLLMs to medical domains requires careful consideration of domain-specific knowledge structures and reasoning patterns that differ fundamentally from general-purpose vision-language tasks. Medical reasoning involves hierarchical knowledge structures, causal relationships, and uncertainty quantification that are not naturally captured by standard transformer architectures. Traditional approaches to this problem have relied on rule-based systems or simple neural mappings that fail to capture the nuanced relationships between linguistic descriptions and spatial configurations. Recent work in neural implicit representations suggests promising directions, with learned mappings of the form $\boldsymbol{\theta}_{geom} = f_\phi(\text{encode}(s))$ where s represents semantic descriptions and f_ϕ is a neural network that maps semantic embeddings to geometric parameters. However, the medical domain introduces additional complexity through the need to maintain anatomical plausibility, incorporate physiological constraints, and ensure clinical consistency across diverse patient populations and pathological conditions.

Bayesian neural networks provide a principled framework for uncertainty estimation through variational inference, with the posterior distribution over model parameters approximated as $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2))$, leading to predictive uncertainty estimates of the form $p(y|x, \mathcal{D}) = \int p(y|x, \boldsymbol{\theta})q(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$. However, the application of uncertainty quantification to multimodal medical systems requires novel approaches that can propagate uncertainty across different modalities and reasoning stages while maintaining computational tractability for real-time applications. Graph neural networks operating on medical ontologies can be formulated as $h_v^{(l+1)} = \sigma\left(W^{(l)}h_v^{(l)} + \sum_{u \in \mathcal{N}(v)} \alpha_{uv} W^{(l)} h_u^{(l)}\right)$, where $h_v^{(l)}$ represents node embeddings at layer l and α_{uv} denotes attention weights that capture the strength of medical relationships between concepts. The challenge lies in designing graph structures that accurately reflect medical knowledge hierarchies while enabling efficient inference and gradient-based optimization for end-to-end learning.

Despite these advances, current approaches suffer from fundamental limitations that prevent their effective deployment in complex clinical scenarios. The lack of semantic understanding in geometric approaches, the absence of spatial reasoning in language models, and the difficulty of bridging symbolic and subsymbolic representations create persistent gaps that compromise both accuracy and clinical utility. Our work addresses these limitations through a novel framework that treats semantic understanding as the primary organizing principle for interventional guidance, enabling more robust and clinically meaningful assistance systems.

Experiments

Implementation Details

Our semantic-guided cross-dimensional synthesis framework was implemented using PyTorch 2.0 with CUDA 11.8 support, deployed across a distributed computing cluster consisting of 8 NVIDIA A100 GPUs with 80GB memory each. The implementation leverages mixed-precision training with automatic loss scaling to optimize memory utilization while maintaining numerical stability for the complex multimodal computations.

Medical Scene Understanding Module Configuration The domain-adapted multimodal large language model $\mathcal{M}_{\text{med}}$ is built upon a transformer architecture with $L = 24$ layers, embedding dimension $d_{\text{model}} = 1024$, and 16 attention heads. The visual encoder ϕ_v employs a Vision Transformer (ViT-Large) backbone with patch size 16×16, producing feature representations of dimension $d_v = 1024$. The fluoroscopic sequences are preprocessed with adaptive histogram equalization and contrast-limited adaptive histogram equalization (CLAHE) with clip limit $\lambda_{\text{clip}} = 2.0$ to enhance vessel visibility under varying contrast conditions.

The medical knowledge graph $\mathcal{G}_{\text{med}}$ contains 15,247 anatomical concepts and 42,386 relationships extracted from the Unified Medical Language System (UMLS), Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), and domain-specific cardiovascular ontologies. Graph embeddings are learned using a Graph Attention Network with 4 layers, hidden dimension 512, and dropout rate 0.15. The structured state-space model parameters are initialized with $\mathbf{A} \in \mathbb{R}^{512 \times 512}$ following the HiPPO initialization scheme for optimal long-range dependency modeling.

Semantic-to-Geometry Translation Engine Configuration The hierarchical semantic encoder operates at $L = 6$ hierarchy levels, with level-specific embedding dimensions $\{128, 256, 512, 512, 256, 128\}$ to capture multi-scale semantic information. The neural field network $\mathcal{N}_\theta$ consists of 8 fully-connected layers with 256 hidden units each, employing spectral normalization with spectral radius constraint $\rho_{\max} = 0.95$ for training stability.

Positional encoding $\gamma(\mathbf{p})$ uses 10 frequency bands with logarithmic sampling: $\gamma(\mathbf{p}) = [\sin(2^k\pi\mathbf{p}), \cos(2^k\pi\mathbf{p})]_{k=0}^{9}$, resulting in 60-dimensional encoded positions. The anatomical plausibility constraints are enforced through a differentiable optimization process with tolerance $\epsilon_{\text{plaus}} = 0.01$ and maximum 50 iterations using the L-BFGS optimizer.

Adaptive Neural Rendering System Configuration The neural radiance field $\mathcal{F}_{\text{med}}$ employs a hierarchical sampling strategy with 64 coarse samples and 128 fine samples along each ray. The base MLP consists of 8 layers with 256 hidden units, while the density and color MLPs use 2 layers with 128 units each. Skip connections are inserted at the 4th layer to preserve gradient flow through the deep architecture.

Volume rendering is performed using hierarchical sampling with importance-based ray selection, where rays are prioritized based on semantic importance scores $\omega_{\text{sem}}(s_i)$ computed from medical knowledge graph traversal. The rendering resolution is set to 512×512 pixels with anti-aliasing through 4x supersampling and bicubic downscaling.

Uncertainty quantification employs Monte Carlo Dropout with 20 forward passes during inference, combined with learned variance estimation through a dedicated uncertainty head. The variational parameters are optimized using the reparameterization trick with β-VAE objective and $\beta = 0.5$ for balanced reconstruction-regularization trade-off.

Training Configuration and Optimization The model is trained end-to-end using the AdamW optimizer with learning rate $\eta = 3 \times 10^{-4}$, weight decay $\lambda_{\text{wd}} = 1 \times 10^{-2}$, and $\beta_1 = 0.9, \beta_2 = 0.999$. Learning rate scheduling follows a cosine annealing strategy with warm-up period of 1,000 iterations and minimum learning rate $\eta_{\min} = 1 \times 10^{-6}$.

The loss function weights are set as $\lambda_{\text{recon}} = 1.0$, $\lambda_{\text{sem}} = 0.5$, $\lambda_{\text{med}} = 0.3$, and $\lambda_{\text{reg}} = 0.1$ based on extensive hyperparameter optimization using Bayesian optimization with Gaussian Process surrogate models. Gradient clipping is applied with maximum norm $\|\nabla\|_{\max} = 1.0$ to prevent gradient explosion in the complex multimodal architecture.

Training data augmentation includes random spatial transformations (rotation ±15, translation ±10%, scaling $0.9 - 1.1$), intensity variations (brightness ±0.2, contrast $0.8 - 1.2$), and temporal perturbations (frame dropping with probability 0.1, temporal jittering ±2 frames).

Baselines

We compare our proposed method against state-of-the-art approaches across three categories: traditional geometric methods, deep learning-based segmentation systems, and recent multimodal fusion techniques. All baseline methods are implemented with identical preprocessing pipelines and evaluated on the same datasets to ensure fair comparison.

Traditional Geometric Approaches **Centerline Extraction with Tubular Tracking (CETT):** This physics-based approach employs the Cosserat rod model for guidewire dynamics simulation combined with multi-scale Hessian-based vessel enhancement. The method uses eigenvalue analysis of the Hessian matrix $\mathbf{H}$ at multiple scales $\sigma \in \{0.5, 1.0, 1.5, 2.0\}$ to identify tubular structures through vesselness measures $V(\mathbf{x}) = \max_\sigma V_\sigma(\mathbf{x})$ where V_σ is computed from Hessian eigenvalues.

Structure Tensor-based Vessel Segmentation (STVS): This method combines structure tensor analysis with level-set evolution for vessel boundary delineation. The structure tensor $\mathbf{T} = \nabla I \nabla I^T * G_\sigma$ captures local image orientation, where $*$ denotes convolution with Gaussian kernel G_σ. Evolution is governed by the geometric active contour model with curvature regularization term $\kappa = \text{div}(\frac{\nabla\phi}{|\nabla\phi|})$.

Multi-View Geometric Reconstruction (MVGR): This approach reconstructs 3D vessel geometry from multiple fluoroscopic views using epipolar constraints and bundle adjustment optimization. The method minimizes reprojection error $\sum_{i,j} \|\mathbf{p}_{i,j} - \Pi_j(\mathbf{X}_i)\|^2$ where $\mathbf{p}_{i,j}$ are 2D observations and Π_j represents camera projection matrices.

Deep Learning-based Segmentation Systems **Attention U-Net with Dense Connections (AU-Dense):** This architecture enhances the standard U-Net with attention gates and dense connectivity patterns. Attention gates compute attention coefficients $\alpha_{i,j} = \sigma_2(\mathbf{W}_\alpha^T[\sigma_1(\mathbf{W}_g^T\mathbf{g}_i + \mathbf{W}_x^T\mathbf{x}_j + \mathbf{b})] + \mathbf{b}_\alpha)$ to suppress irrelevant features while highlighting vessel regions.

DeepLab-v3+ with Atrous Spatial Pyramid Pooling (DeepLab-ASPP): This semantic segmentation approach employs atrous convolution with multiple dilation rates $\{6, 12, 18\}$ to capture multi-scale contextual information. The ASPP module aggregates features through parallel atrous convolutions: $\mathbf{y} = \sum_{r \in R} \mathbf{W}_r *_r \mathbf{x}$ where $*_r$ denotes atrous convolution with rate r.

Swin Transformer for Medical Segmentation (Swin-MedSeg): This method adapts the Swin Transformer architecture for medical image segmentation with shifted window attention mechanisms. The architecture employs hierarchical feature extraction with window sizes $\{7 \times 7, 14 \times 14\}$ and shifting strategies to model long-range dependencies in fluoroscopic images.

Recent Multimodal Fusion Techniques **Cross-Modal Attention Fusion Network (CMAFN):** This approach combines visual and textual features through bidirectional cross-attention mechanisms. The fusion process employs scaled dot-product attention: $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}$ where queries come from one modality and keys/values from another.

Neural Radiance Fields for Medical Imaging (NeRF-Med): This baseline adapts standard NeRF to medical scenarios with density-based volume rendering $\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt$ where $T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds)$ represents transmittance along ray $\mathbf{r}$.

Vision-Language Pre-trained Model for Medical Analysis (VL-MedAnalysis): This method fine-tunes CLIP-based architectures on medical image-text pairs using contrastive learning objectives $\mathcal{L} = -\log \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_i)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau)}$ where sim denotes cosine similarity and τ is the temperature parameter.

Case Study: Complex Interventional Scenarios

To demonstrate the clinical utility and robustness of our semantic-guided framework, we present three representative case studies that highlight challenging scenarios commonly encountered in interventional cardiology. These cases were selected from our multi-institutional dataset to showcase the system's ability to handle complex anatomical variations, pathological conditions, and procedural contexts that often challenge existing guidance systems.

Case 1: Chronic Total Occlusion with Extensive Calcification Our first case involves a 68-year-old patient with chronic total occlusion (CTO) of the right coronary artery complicated by extensive calcification and collateral circulation. The fluoroscopic sequence exhibited severe contrast limitations due to heavy calcification, with Hounsfield unit variations exceeding 400 HU within the target vessel region. Traditional geometric approaches (CETT, STVS) failed to maintain consistent vessel tracking, with centerline deviation errors reaching 8.7 ± 2.3 mm in calcified segments. Deep learning methods (Swin-MedSeg, DeepLab-ASPP) showed improved performance but struggled to differentiate between calcified vessel walls and contrast-filled lumens, resulting in anatomical consistency scores below 0.65.

Our semantic-guided approach successfully interpreted the complex pathological context through natural language descriptions such as *"heavily calcified right coronary artery with chronic total occlusion showing bridging collaterals from left anterior descending system."* The medical knowledge graph effectively encoded relationships between calcification patterns, collateral circulation, and optimal navigation strategies, enabling accurate geometric reconstruction despite poor image quality. The system achieved an ACS of 0.842 and CDE of 2.8 mm in this challenging scenario, representing a 29% improvement over the best-performing baseline. Critically, the semantic understanding module correctly identified the functional occlusion location and predicted optimal entry points for guidewire advancement, information that proved invaluable for procedural planning.

Case 2: Bifurcation Lesion with Dynamic Vessel Movement The second case presents a complex left main coronary artery bifurcation lesion in a 72-year-old patient with significant respiratory motion artifacts. The imaging sequence contained 127 frames with cardiac cycle-induced vessel displacement ranging from 3.2 to 8.9 mm, creating substantial temporal inconsistencies that challenged reconstruction algorithms. Motion artifacts were particularly pronounced during mid-diastolic phases, where vessel boundaries became indistinct due to breathing-related displacement.

Baseline methods exhibited poor temporal consistency, with frame-to-frame reconstruction variations exceeding 15% for geometric approaches and 12% for deep learning methods. The multimodal baselines (VL-MedAnalysis, NeRF-Med) showed improved stability but failed to maintain anatomical plausibility during high-motion phases, resulting in geometrically implausible vessel configurations with unrealistic curvature values ($\kappa > 0.8$ mm^{-1}).

Our approach leveraged semantic context to maintain anatomical consistency across temporal sequences. The system generated descriptions such as *"left main bifurcation with moderate plaque burden at ostial LAD showing preserved TIMI-3 flow with respiratory motion artifacts."* The adaptive neural rendering system successfully incorporated motion priors from the medical knowledge graph, constraining geometric reconstructions to anatomically plausible configurations even during challenging motion phases. The temporal consistency measure reached 0.891, compared to 0.734 for the best baseline, while maintaining an ACS of 0.859 throughout the entire sequence. The semantic understanding enabled the system to distinguish between pathological vessel irregularities and motion-induced apparent deformations, crucial for accurate lesion assessment.

Case 3: Multi-Vessel Disease with Complex Anatomical Variants The third case involves a 59-year-old patient with multi-vessel coronary artery disease featuring anatomical variants including a dominant circumflex system with anomalous right coronary artery origin from the left sinus of Valsalva. This complex anatomy presented significant challenges for automated guidance systems, as standard anatomical models could not accommodate the unusual vessel topology and spatial relationships. Traditional and deep learning approaches failed to recognize the anatomical variants, resulting in systematic reconstruction errors with CDE values exceeding 6.2 mm in variant vessel segments. The methods attempted to force the unusual anatomy into standard templates, creating geometrically inconsistent 3D models that could potentially mislead clinical decision-making. Semantic consistency scores for baselines remained below 0.45, indicating poor alignment between visual observations and anatomical understanding.

Our semantic-guided framework successfully identified and accommodated the anatomical variants through comprehensive medical knowledge integration. The system generated detailed descriptions including *"anomalous RCA origin from left coronary cusp with retroaortic course, dominant circumflex system supplying posterior descending artery, and intermediate vessel with focal stenosis."* The medical knowledge graph contained extensive anatomical variant information, enabling the semantic-to-geometry translation engine to generate appropriate geometric constraints for the unusual vessel configuration. The approach achieved an ACS of 0.876 and SGCI of 0.851 in this challenging case, demonstrating robust adaptation to anatomical variations without requiring retraining or manual parameter adjustment.

Across all three cases, our method consistently outperformed baseline approaches while providing clinically relevant semantic interpretations that enhanced procedural understanding. The semantic descriptions generated by our system closely matched expert cardiologist assessments, with clinical relevance scores averaging 4.2 ± 0.3 on a 5-point scale. These case studies validate the practical utility of semantic-guided medical scene understanding in addressing real-world clinical challenges that extend beyond standard performance metrics.