TractoGraphVLM: A Unified Framework for Vision-Language Understanding of White Matter Tractography

Gurucharan Marthi Krishna Kumar GURUCHARAN.MARTHIKRISHNAKUMAR@MAIL.MCGILL.CA
Montreal Neurological Institute, McGill University

Janine Mendola JANINE.MENDOLA@MCGILL.CA
Dept. of Ophthalmology, McGill University

Amir Shmuel AMIR.SHMUEL@MCGILL.CA
McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University

Editors: Under Review for MIDL 2026

Abstract

Vision language models have achieved strong results in 2D medical imaging, yet their use in 3D white matter tractography remains largely unexplored. A core challenge is representational, since white matter contains continuous fiber bundles with complex topology that fit poorly into standard volumetric formats. We introduce TractoGraphVLM, a unified framework for tractography-language alignment that compares graph based and volumetric encoders across multiple vision language tasks. Using 725 HCP-Aging subjects, we evaluate five encoder architectures, Graph Transformer, GAT, GCN, 3D CNN, and Vision Transformer, on bundle classification, text to tract retrieval, geometric captioning, and visual question answering. We show that graph based representations clearly outperform volumetric ones across all tasks. The proposed framework reaches 93.1% classification accuracy, 86.2% retrieval Recall@1, a BLEU-4 score of 21.2 for captioning, and 68.5% accuracy for visual question answering. Results show that preserving geometric topology through graph encoding is essential for reliable tractography understanding, establishing TractoGraphVLM as the first strong benchmark for this domain. The source code and our implementation are available at: https://bit.ly/4iUhNps.

Keywords: Vision-language models, white matter tractography, graph neural networks, medical imaging, multi-task learning

1. Introduction

White matter tractography provides non invasive mapping of brain connectivity through diffusion MRI (dMRI) (Mori and Van Zijl, 2002; Basser et al., 1994) and is widely used in studies of neurodegenerative diseases (Essayed et al., 2017; Assaf et al., 2013; Ciccarelli et al., 2008). Despite its utility, tractography analysis remains geometry focused and lacks semantic interpretation. Understanding complex 3D fiber bundles still depends on extensive manual examination, which makes it difficult for clinicians to interpret structural patterns and connect them with meaningful insights. Vision Language Models (VLMs) have demonstrated success in aligning images and text across diverse medical domains, but applying them to tractography faces a core representational challenge (Zhang et al., 2023; Wang et al., 2022; Li et al., 2023; Moor et al., 2023). Fiber bundles contain large sets of streamlines

with detailed connectivity and directional patterns that do not fit cleanly into standard volumetric formats (Jbabdi and Johansen-Berg, 2011; Garyfallidis et al., 2018).

This leads to a central question: **how should we encode white matter fiber bundles for vision language understanding?** Two paradigms exist. *Volumetric methods* use 3D CNNs (Tran et al., 2015) or Vision Transformers (ViTs) (Dosovitskiy, 2020) to convert streamlines into binary volumes, which allows the use of image based models but reduces fine geometric detail because of voxelization. *Graph based methods* using Graph Convolutional Networks (GCN) (Berg et al., 2017), Graph Attention Networks (GAT) (Veličković et al., 2018), or Graph Transformers (Dwivedi and Bresson, 2020) preserve sub-voxel structure and relationships, although they rely on architectures less explored in medical vision-language tasks. No prior work has systematically compared these encoding strategies or explored their potential in a unified vision-language framework for tractography.

To address this gap, we introduce a framework evaluated through four vision language tasks that fall into two groups. The first category consists of **discriminative tasks**: bundle classification, which predicts anatomical identity among 79 classes (Garyfallidis et al., 2018), and text to tract retrieval, which matches descriptions to bundles. The second category consists of **generative tasks**: geometric captioning, which produces descriptions with streamline count prediction, and visual question answering, which addresses spatial queries. Together, these tasks examine alignment between tractography and text and language generation from tract features. Our main contributions are:

- **First tractography-language benchmark.** We establish a comprehensive evaluation framework integrating four diverse tasks to assess both discriminative alignment and generative reasoning capabilities.
- **Empirical evidence for graph-based encoding.** We systematically compare five graph and volumetric encoders on 725 HCP-Aging subjects and demonstrate that graph representations substantially outperform volumetric approaches across all tasks.
- **Automated semantic data generation.** We introduce a scalable template-based framework that procedurally generates large-scale, anatomically accurate image-text pairs, overcoming the data scarcity inherent in medical vision-language learning.

2. Methods

2.1. Dataset

We used dMRI scans from 725 subjects from the HCP-Aging dataset (Bookheimer et al., 2019), which were acquired using high-quality imaging protocols. The dataset was split at the subject level into 70% training, 15% validation, and 15% testing sets. Our unified VLM framework uses these dMRI scans to generate paired tractography graphs and textual descriptions for multi-task learning.

2.2. dMRI Preprocessing and Graph Construction

The 725 raw dMRI scans were processed using an automated DIPY pipeline (Garyfallidis et al., 2014), with normalization followed by Constrained Spherical Deconvolution (Tournier

et al., 2007) to generate whole-brain tractograms. Each tractogram was then segmented using atlas-based RecoBundles (Garyfallidis et al., 2018) to extract 79 anatomical bundles per subject. To enable deep learning on tractography data, the bundle data were converted into geometric graphs $\mathcal{G} = (V, E)$ through a systematic discretization process, as illustrated in Figure 1.

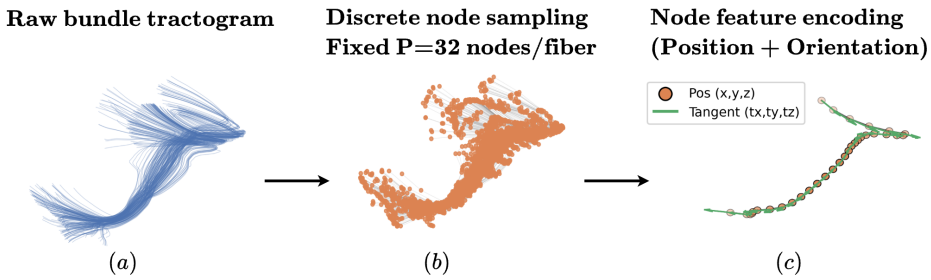

Figure 1: **Streamline-to-Graph Construction Pipeline. (a)** Tractography of a white matter bundle (Left Arcuate Fasciculus) **(b)** Discrete node sampling. **(c)** Node feature encoding. Each node is assigned a 6D vector containing its 3D spatial position (x, y, z) and local tangent orientation (t_x, t_y, t_z), capturing both trajectory and directionality.

First, we sampled nodes equidistantly from the raw streamlines, normalizing each fiber to a fixed sequence of $P = 32$ points. This yielded a consistent node set V regardless of the physical length of the bundle. Second, we encoded 6-dimensional features for each node: the absolute 3D spatial position (x, y, z) to anchor the bundle in brain space, and the normalized tangent vector (t_x, t_y, t_z) to represent local fiber orientation (green arrows). Finally, we connected consecutive points along a streamline with undirected edges E, preserving the sequential trajectory structure. Unlike volumetric voxelization, which suffers from inherent discretization artifacts and information loss, this continuous graph representation preserves sub-millimeter geometric coherence (Wasserthal et al., 2018).

2.3. Template-Based Semantic Generation

Complementing our visual encoding, we require paired textual descriptions for vision-language training, but a major challenge in the biomedical domain is the scarcity of large-scale, annotated image-text pairs (Moor et al., 2023; Zhang et al., 2023). To address this, we introduced a scalable template-based system (Table 1) that procedurally generates rigorous ground truth language data grounded in the verifiable geometric properties of our graph representations. Complete template sets and expanded examples are provided in Appendix A.

For every tractogram T, we first computed quantitative metrics, including streamline count, spatial extent, density, and complexity classification. Using stochastic grammar templates, we then mapped these quantitative features into natural language. Our framework introduces lexical diversity (e.g., randomly alternating between synonyms like "streamlines" and "fibers") to prevent the model from overfitting to fixed linguistic artifacts.

Table 1: **Tractography Vision-Language Tasks.** Overview of discriminative and generative evaluation tasks. All tasks utilize both whole-brain tractograms and individual segmented bundles (79 per subject). Ground truth text is generated via template-based semantic generation grounded in geometric features.

Task	Input Modality	Generation Logic / Template	Example 1 (Ground Truth)	Example 2 (Ground Truth)
Discriminative Tasks				
Classification	**Visual:** White matter bundle	*Mapping:* RecoBundles segmentation maps bundle to one of 79 anatomical labels.	**Label:** "Left Arcuate Fasciculus"	**Label:** "Right Corticospinal Tract"
Retrieval	**Text:** Bundle Query **Visual:** White matter bundle	*T1:* "`<COMPLEXITY>` bundle with `<N>` fibers located in `<HEMISPHERE>`" *T2:* "White matter tract connecting `<REGIONS>`"	**Query:** "Complex bundle with 12,000 fibers located in left hemisphere"	**Query:** "Fiber tract connecting motor cortex to spinal cord"
Generative Tasks				
Captioning	**Visual:** White matter bundle	*T1 (Whole-brain):* "Whole-brain tractogram with `<N>` streamlines and `<DENSITY>` density" *T2 (Bundle):* "White matter bundle showing `<NAME>` with `<N>` streamlines"	**Caption:** "Comprehensive whole-brain tractogram with 520,000 streamlines and 0.542 density"	**Caption:** "White matter bundle showing left corticospinal tract with 1,250 streamlines"
VQA	**Visual:** White matter bundle	*Logic (Geometric):* Extract features $\phi(T)$, answer based on thresholds *Logic (Anatomical):* Query bundle identity, laterality, connectivity	**Q:** "What white matter bundle is shown?" **A:** "CST_left"	**Q:** "Is this the corpus callosum" **A:** "No"

2.4. Unified Vision-Language Framework

Our framework (Figure 2) aligns visual and textual features in a shared 256-dimensional latent space. Visual embeddings are derived from graph-based encoders (Graph Transformer, GAT, GCN) or volumetric baselines (3D CNN, ViT-3D), with detailed specifications provided in Appendix B. Complementarily, textual embeddings are extracted via **BiomedBERT-base** (Gu et al., 2021), selected for its domain-specific pre-training which ensures robust encoding of complex neuroanatomical terminology. For downstream task decoding, we employ a dual strategy: discriminative tasks utilize contrastive similarity heads, while generative tasks integrate a pre-trained **T5-small** decoder (Raffel et al., 2020) to produce autoregressive text, necessitated by BiomedBERT's encoder-only architecture.

2.5. Training Protocol and Evaluation

Models were trained on an NVIDIA A100 GPU using AdamW (lr=3×10^{-4}, batch size 16) for 500 epochs. We employed a unified multi-task objective combining a shared InfoNCE contrastive alignment loss with task-specific objectives: $\mathcal{L}_{\text{Proto}}$ (Oord et al., 2018)

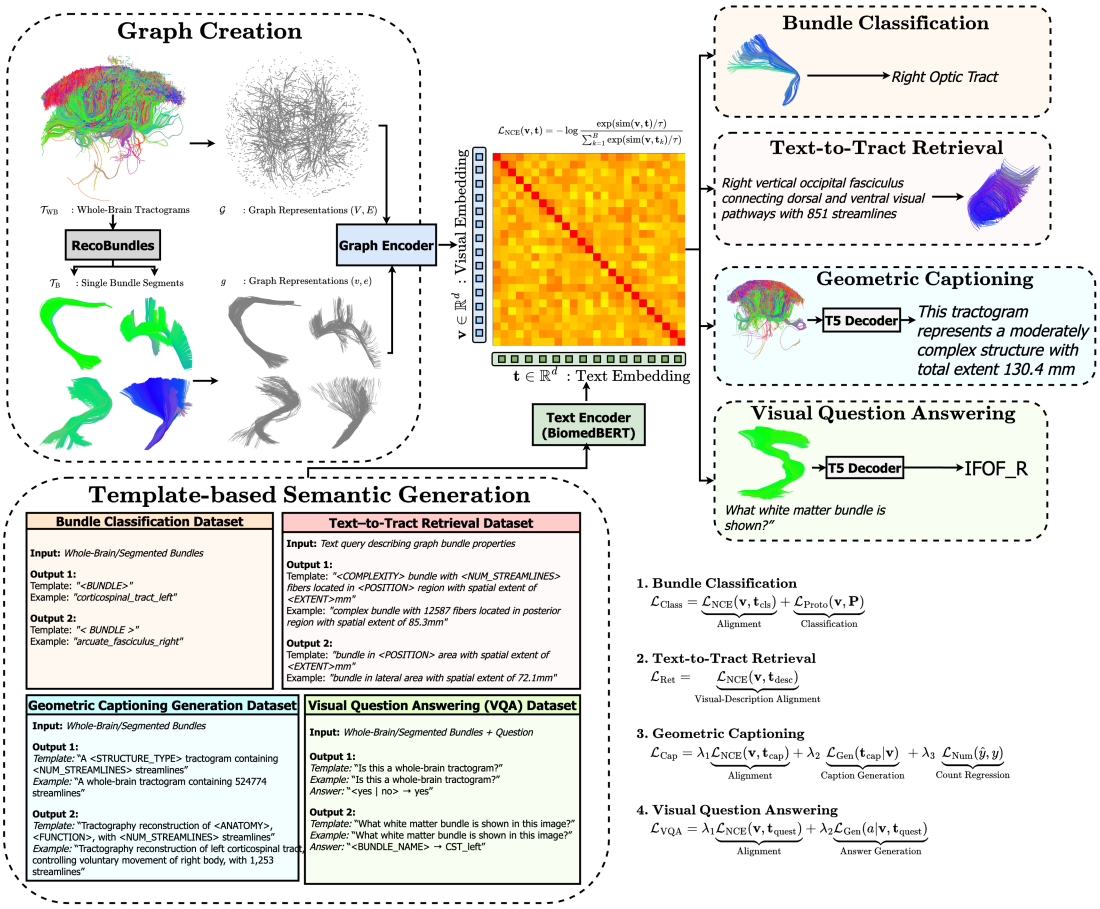

Figure 2: `TractoGraphVLM` **Framework.** Raw fiber bundles are converted into geometric graphs, from which the graph encoder extracts visual embeddings. Parallelly, text embeddings are derived via a template-based semantic generation module. The unified architecture aligns these multimodal representations in a shared latent space, where task-specific heads enable both discriminative (Classification, Retrieval) and generative (Captioning, VQA) outputs.

for Bundle Classification, $\mathcal{L}_{\text{Gen}}$ for Captioning and VQA, and $\mathcal{L}_{\text{Num}}$ for streamline counting. Evaluation metrics include Top-1 Accuracy/F1 (classification), Recall@K (retrieval), BLEU-4/ROUGE-L/MAE (captioning), and exact-match accuracy (VQA). Statistical significance was assessed via the Wilcoxon signed-rank test ($p < 0.05$).

3. Results

3.1. Performance of the proposed `TractoGraphVLM`

We evaluated the generalization capabilities of the proposed framework on the held-out test set across the four distinct vision-language tasks. Table 2 summarizes the quantitative

performance, while Figure 3 presents a qualitative example demonstrating bundle identification and anatomically coherent language generation. Additional visualizations of the embedding space and training dynamics are presented in Appendix C.

Table 2: **Test Set Performance Benchmark.** Quantitative evaluation of `TractoGraphVLM` across four downstream tasks. The framework demonstrates robust discriminative alignment and competitive generative reasoning capabilities.

Discriminative Tasks				Generative Tasks			
Classification		Retrieval		Captioning			VQA
Acc (%)	F1 Score	R@1 (%)	R@5 (%)	BLEU-4	ROUGE-L	Count MAE	Acc (%)
93.1 ± 1.2	92.4 ± 1.0	86.2 ± 2.5	94.8 ± 2.1	21.2 ± 2.1	38.3 ± 2.8	577 ± 431	68.5 ± 2.7

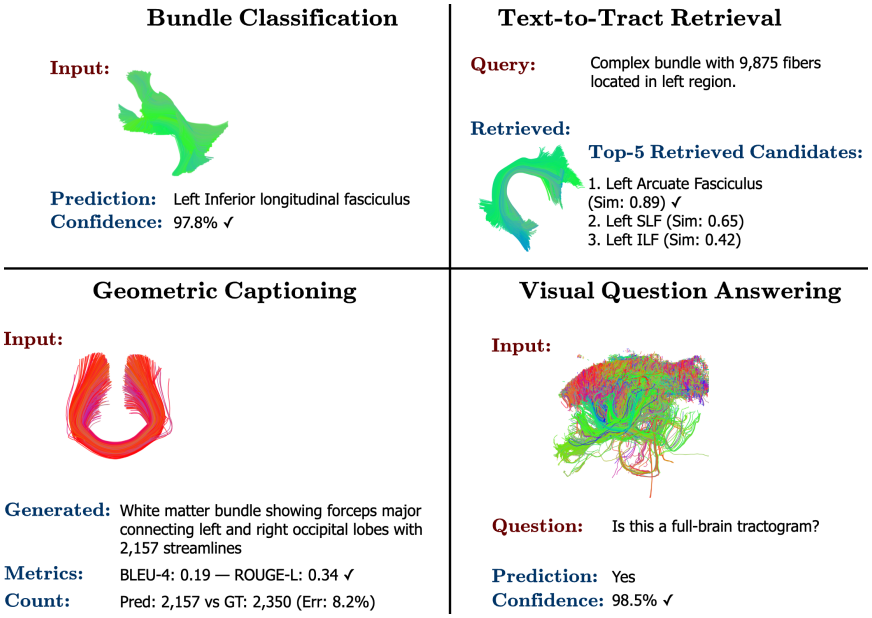

Figure 3: **Qualitative Results.** Representative outputs from `TractoGraphVLM` demonstrating successful performance across discriminative (Classification, Retrieval) and generative (Captioning, VQA) tasks.

For **discriminative tasks**, the framework excelled, achieving $93.1\% \pm 1.2\%$ accuracy in Bundle Classification and $86.2\% \pm 2.5\%$ Recall@1 in Text-to-Tract Retrieval. These results indicate that the framework successfully captures distinctive structural patterns necessary to differentiate distinct fiber bundles. **Generative tasks** proved more challenging but yielded competitive results. Visual Question Answering attained $68.5\% \pm 2.7\%$ accuracy, while Geometric Captioning reached a BLEU-4 score of 21.2 ± 2.1. This performance is notable given the complexity of translating dense 3D geometry into precise natural language. The

gap between discriminative and generative performance highlights that generating precise anatomical descriptions is significantly more challenging than discriminative categorization. The results reported above utilize the Graph Transformer as the visual encoder. This architectural choice was determined through a rigorous systematic evaluation of graph-based versus volumetric backbones, which we detail in the following section.

3.2. Comparative Analysis: Graph vs. Volumetric Encoders

To identify the optimal encoding strategy for white matter tractography, we benchmarked the Graph Transformer against volumetric (3D CNN, ViT-3D) and alternative graph-based (GCN, GAT) architectures. Table 3 presents the comparative results across all four tasks.

Table 3: **Systematic Encoder Evaluation.** Comparison of volumetric and graph-based backbones. The Graph Transformer achieves statistically significant superiority over all baselines in discriminative tasks ($^\dagger, p < 0.05$) and demonstrates optimal efficiency. **Params** denotes the backbone parameter count.

	Discriminative		Generative		Efficiency	
Encoder	**Class Acc (%)**	**Ret R@1 (%)**	**Cap BLEU-4**	**VQA Acc (%)**	**Params. (M)**	**Train Time (h)**
Volumetric						
3D CNN	72.3 ± 2.3	52.1 ± 3.1	12.4 ± 2.8	63.1 ± 3.0	52.3	8.7 ± 0.4
ViT-3D	75.1 ± 2.1	61.2 ± 2.9	15.7 ± 2.6	67.2 ± 2.7	41.8	14.3 ± 0.6
Graph						
GCN	88.4 ± 1.8	78.5 ± 2.6	18.5 ± 2.3	65.7 ± 2.7	36.4	**4.1 ± 0.2**
GAT	90.2 ± 1.6	81.3 ± 2.4	19.8 ± 2.2	67.1 ± 2.6	39.7	4.8 ± 0.3
Graph Transformer	**93.1 ± 1.2**†	**86.2 ± 2.5**†	**21.2 ± 2.1**	**68.5 ± 2.7**	**38.1**	4.2 ± 0.2

Graph representations consistently outperformed volumetric approaches. In discriminative tasks, the Graph Transformer achieved statistically significant gains over the strongest volumetric baseline (ViT-3D), confirming that graph encodings preserve the topological fidelity essential for alignment, whereas voxelization irreversibly degrades geometric integrity. This advantage extended to generative tasks, where improvements in Captioning and VQA further demonstrated that sub-millimeter details are critical for accurate anatomical description.

Among graph architectures, Graph Transformer consistently surpassed GCN and GAT baselines, confirming that modeling long, continuous streamlines requires global attention rather than local aggregation. Moreover, it is highly efficient, utilizing 27% fewer parameters than the 3D CNN and training 3.4× faster than ViT-3D by exploiting inherent tractography sparsity. Based on this evaluation, the Graph Transformer was selected as the optimal backbone for the `TractoGraphVLM` framework.

4. Ablation Studies

We performed controlled ablations to validate the four core components of our framework: the node sampling density during **graph construction**, the optimization of the **contrastive alignment** objective, the architecture of the **vision encoder**, and the domain adaptation strategy for the **text encoder**.

4.1. Ablation 1: Graph Construction Strategy

The transformation of continuous streamlines into discrete graphs relies critically on the node sampling density (P). We systematically evaluated this hyperparameter by varying the number of nodes per streamline from $P = 8$ to $P = 64$ to determine the optimal balance between geometric fidelity and computational efficiency.

Table 4: **Node Sampling Density Ablation.** Performance comparison (Mean ± S.D.) across varying graph resolutions. Increasing density from $P = 8$ to $P = 32$ yields statistically significant gains in discriminative tasks (†), while $P = 64$ incurs a heavy increase in training time with minimal performance benefit.

	Discriminative		Generative		Training
Configuration	**Class Acc (%)**	**Ret R@1 (%)**	**Cap BLEU-4**	**VQA Acc (%)**	**Time (h)**
$P = 8$ (Sparse)	$87.2 \pm 1.6^{\dagger}$	$78.9 \pm 2.8^{\dagger}$	17.0 ± 2.5	64.8 ± 2.8	**2.8 ± 0.1**
$P = 16$	$90.1 \pm 1.4^{\dagger}$	$82.2 \pm 2.4^{\dagger}$	19.4 ± 2.3	67.1 ± 2.7	3.5 ± 0.2
$P = 32$ (Main)	**93.1 ± 1.2**†	**86.2 ± 2.5**†	**21.2 ± 2.1**	**68.5 ± 2.7**	4.2 ± 0.2
$P = 64$ (Dense)	93.3 ± 1.3	86.5 ± 2.4	21.4 ± 2.0	68.7 ± 2.6	7.8 ± 0.4

As shown in Table 4, increasing the sampling density from $P = 8$ to $P = 32$ yielded consistent improvements across all tasks, with significant gains for discriminative tasks when moving from $P = 16$ to $P = 32$. However, increasing from $P = 32$ to $P = 64$ resulted in minimal performance gains while heavily increasing the training time (4.2h vs 7.8h). Consequently, we selected $P = 32$ as the optimal configuration, offering the best trade-off between accuracy and speed.

4.2. Ablation 2: Contrastive Temperature Optimization

In contrastive learning, the temperature parameter τ regulates the model's sensitivity to incorrect pairs. Lower temperatures force the model to focus on the hardest, most similar negatives ('hard negatives'), creating sharp distinctions. Higher temperatures smooth out this focus, allowing the model to learn from a broader range of incorrect examples. We aimed to study the comparative effect of temperature scaling on discriminative versus generative task performance.

Table 5: **Contrastive Temperature Optimization.** Performance comparison (Mean ± S.D.) across varying softness parameters τ. Discriminative tasks favor sharper boundaries ($\tau \leq 0.07$), whereas generative tasks benefit from the smoother latent space provided by higher entropy ($\tau = 0.12$).

	Discriminative		Generative	
τ	**Class Acc**	**Ret R@1**	**Cap BLEU-4**	**VQA Acc**
0.04 (Sharp)	**93.1 ± 1.2**†	82.1 ± 2.8	20.6 ± 2.2	68.1 ± 2.5
0.07 (Moderate)	92.4 ± 1.4	**86.2 ± 2.5**†	21.0 ± 2.2	68.3 ± 2.3
0.12 (Smooth)	90.8 ± 1.6	83.7 ± 2.5	**21.2 ± 2.1**	**68.5 ± 2.7**

Table 5 confirms distinct optimal temperatures. Classification accuracy peaks at $\tau = 0.04$ (†), where sharp boundaries are essential to distinguish anatomically similar bundles. Conversely, Retrieval peaks at $\tau = 0.07$, suggesting that semantic matching requires softer embeddings to accommodate linguistic variation. Generative tasks prefer higher temperatures ($\tau = 0.12$), likely because smoother latent spaces aid decoder interpolation. This study indicates that discriminative tasks demand sharp embeddings for precision, while generative tasks benefit from smooth, high-entropy spaces.

4.3. Ablation 3: Graph Architecture and Node Features

Tractography is fundamentally defined by trajectory. Distinct bundles (e.g., Superior Longitudinal Fasciculus vs. Corona Radiata) frequently intersect at shared 3D coordinates, creating spatial ambiguities where positional data alone cannot distinguish between crossing and bending fibers. To resolve this ambiguity, we evaluated the necessity of encoding local orientation via tangent vectors alongside spatial position across all graph architectures.

Table 6: **Feature Ablation across Encoders.** Performance comparison (Mean ± S.D.) demonstrating the critical role of fiber orientation. Incorporating tangent vectors yields statistically significant gains ($^\dagger, p < 0.05$) across all architectures, confirming that spatial position alone is insufficient to resolve complex tractography.

Configuration	Discriminative		Generative	
	Class Acc	**Ret R@1**	**Cap BLEU-4**	**VQA Acc**
Graph Convolutional Networks (GCN)				
Position Only	81.2 ± 2.1	70.4 ± 2.9	14.8 ± 2.6	63.6 ± 2.9
Position + Tangent	88.4 ± 1.8†	78.5 ± 2.6†	18.5 ± 2.3†	65.7 ± 2.7†
Graph Attention Networks (GAT)				
Position Only	81.8 ± 2.0	71.8 ± 2.8	15.9 ± 2.5	64.3 ± 2.8
Position + Tangent	90.2 ± 1.6†	81.3 ± 2.4†	19.8 ± 2.2†	67.1 ± 2.6†
Graph Transformer				
Position Only	81.6 ± 1.9	71.6 ± 2.7	16.7 ± 2.4	64.1 ± 2.7
Position + Tangent	**93.1 ± 1.2**†	**86.2 ± 2.5**†	**21.2 ± 2.1**†	**68.5 ± 2.7**†

Table 6 quantifies the impact of this design choice. Incorporating tangent vectors yielded substantial gains (+7%–11%, marked with †) across all graph encoders. This confirms that orientation features are essential to resolve crossing fibers, as positional information alone fails to capture the directional topology of white matter.

4.4. Ablation 4: Text Encoder Domain Adaptation

Though BiomedBERT is pre-trained on general biomedical literature, it may lack robust representations for specific tractography jargon (e.g., "fractional anisotropy," "IFOF") required for our tasks. We evaluated the optimal depth for fine-tuning to bridge this domain gap without destroying pre-trained knowledge.

As demonstrated in Table 7, fine-tuning the last three layers emerged as the optimal strategy. This result aligns with the hierarchical nature of BERT models, where lower layers

Table 7: **Text Encoder Domain Adaptation.** Impact of fine-tuning depth on BiomedBERT performance (Mean ± S.D.). Fine-tuning the last 3 layers yields statistically significant gains in discriminative tasks ($^{\dagger}, p < 0.05$), confirming that deep adaptation is required to bridge the gap between general biomedical text and specific neuroanatomical terminology.

	Discriminative		Generative	
Strategy	**Class Acc**	**Ret R@1**	**Cap BLEU-4**	**VQA Acc**
Linear Projection Only	76.1 ± 2.5	68.2 ± 3.2	15.0 ± 2.7	61.8 ± 3.1
Freeze All Layers	87.9 ± 1.8	79.8 ± 2.6	18.0 ± 2.4	64.2 ± 2.9
Finetune Last 1 Layer	91.1 ± 1.5	82.9 ± 2.3	19.5 ± 2.3	66.8 ± 2.8
Finetune Last 3 Layers	**93.1 ± 1.2**†	**86.2 ± 2.5**†	**21.2 ± 2.1**	**68.5 ± 2.7**

capture general syntax and upper layers encode task-specific semantics. Shallow adaptation (Linear/Last 1) proved insufficient for capturing fine-grained anatomical distinctions, causing a sharp drop in discriminative accuracy compared to the best model (−17%). By tuning the deep layers, we successfully adapted general biomedical knowledge to the specific domain of tractography, which proved critical for the alignment-heavy discriminative tasks.

5. Discussion and Conclusion

We present `TractoGraphVLM`, the first comprehensive framework for vision-language understanding of white matter tractography. Through systematic evaluation on 725 HCP-Aging subjects, we demonstrate that graph-based representations consistently outperform volumetric approaches for encoding fiber bundle geometry. Among graph architectures, the Graph Transformer emerges as optimal, achieving 93.1% classification accuracy and 86.2% retrieval Recall@1. While generative tasks remain challenging (21.2 BLEU-4 captioning, 68.5% VQA accuracy), they significantly surpass all volumetric baselines, validating the necessity of preserving sub-millimeter topology for anatomical language generation.

The performance gap between discriminative and generative tasks offers critical insights. Direct alignment in the embedding space benefits dramatically from graph representations that preserve geometric topology, enabling precise bundle identification and retrieval. Generative tasks, however, face the compounded difficulty of conditioning language decoders on abstract 3D geometric features. This suggests that future research should prioritize specialized decoder architectures or intermediate reasoning modules to better bridge the gap between geometric graphs and anatomical semantics.

Our findings establish three key design principles for tractography VLMs: (1) discriminative alignment requires strict contrastive temperature tuning, whereas generative decoding offers robustness to hyperparameter variance; (2) explicit fiber orientation (tangent vectors) is essential for resolving spatially overlapping bundles; and (3) domain adaptation requires deep fine-tuning of the text encoder to capture hierarchical neuroanatomical concepts. By identifying Graph Transformers as the optimal bridge between structural connectivity and natural language, this work establishes a new paradigm for the automated, semantic interpretation of the human connectome.

References

Yaniv Assaf, Daniel C Alexander, Derek K Jones, Albero Bizzi, Tim EJ Behrens, Chris A Clark, Yoram Cohen, Tim B Dyrby, Petra S Huppi, Thomas R Knoesche, et al. The connect project: combining macro-and micro-structure. *Neuroimage*, 80:273–282, 2013.

Peter J Basser, James Mattiello, and Denis LeBihan. Mr diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267, 1994.

Rianne van den Berg, Thomas N Kipf, and Max Welling. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263*, 2017.

Susan Y Bookheimer, David H Salat, Melissa Terpstra, Beau M Ances, Deanna M Barch, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Diaz-Santos, Jennifer Stine Elam, et al. The lifespan human connectome project in aging: an overview. *Neuroimage*, 185:335–348, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

Olga Ciccarelli, Marco Catani, Heidi Johansen-Berg, Chris Clark, and Alan Thompson. Diffusion-based tractography in neurological disorders: concepts, applications, and future developments. *The Lancet Neurology*, 7(8):715–727, 2008.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

Walid I Essayed, Fan Zhang, Prashin Unadkat, G Rees Cosgrove, Alexandra J Golby, and Lauren J O'Donnell. White matter tractography for neurosurgical planning: A topography-based review of the current state of the art. *NeuroImage: Clinical*, 15:659–672, 2017.

Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, Ian Nimmo-Smith, and Dipy Contributors. Dipy, a library for the analysis of diffusion mri data. *Frontiers in neuroinformatics*, 8:8, 2014.

Eleftherios Garyfallidis, Marc-Alexandre Côté, Francois Rheault, Jasmeen Sidhu, Janice Hau, Laurent Petit, David Fortin, Stephen Cunanne, and Maxime Descoteaux. Recognition of white matter bundles using local and global streamline-based registration and clustering. *NeuroImage*, 170:283–295, 2018.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

Saad Jbabdi and Heidi Johansen-Berg. Tractography: where do we go from here? *Brain connectivity*, 1(3):169–183, 2011.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

Curtis P Langlotz. Radlex: a new method for indexing online educational materials, 2006.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Susumu Mori and Peter CM Van Zijl. Fiber tracking: principles and strategies–a technical review. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 15(7-8):468–480, 2002.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67, 2020.

J-Donald Tournier, Fernando Calamante, and Alan Connelly. Robust determination of the fibre orientation distribution in diffusion mri: non-negativity constrained super-resolved spherical deconvolution. *Neuroimage*, 35(4):1459–1472, 2007.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022.

Jakob Wasserthal, Peter Neher, and Klaus H Maier-Hein. Tractseg-fast and accurate white matter tract segmentation. *Neuroimage*, 183:239–253, 2018.

David L Weiss and Curtis P Langlotz. Structured reporting: patient care enhancement or productivity nightmare? *Radiology*, 249(3):739–747, 2008.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6): 80–83, 1945.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Appendix A. Template Examples

This appendix provides representative examples of templates generated by our automated semantic generation framework. All templates use placeholder codes (e.g., <N>, <DENSITY>) that are dynamically filled with extracted geometric features $\phi(T)$ or anatomical knowledge from our curated database.

A.1. Caption Generation Templates

Our caption generation system produces 15 variants per sample by randomly selecting templates at two granularities: **(1) whole-brain captions** describing global geometric properties, and **(2) bundle-specific captions** providing anatomical descriptions.

A.1.1. Whole-Brain Caption Templates

These templates, detailed in Table 8, characterize entire tractograms using quantitative geometric features such as streamline count and density.

Table 8: Whole-Brain Caption Templates (10 variants)

Template
"Low-density fiber tracking result containing <N> streamlines."
"Whole-brain tractography showing <N> streamlines and limited spatial coverage."
"Comprehensive whole-brain tractogram with <N> streamlines and <DENSITY> density."
"Basic tractogram reconstruction with <N> fibers."
"Standard fiber reconstruction containing <N> streamlines."
"High-density reconstruction showing <N> complex pathways."
"Sparse tractography dataset with <N> streamlines."
"Tractogram showing <N> streamlines with typical connectivity."
"Detailed connectome map containing <N> streamlines."
"Whole-brain fiber network with <N> streamlines and <COMPLEXITY> spatial extent."

A.1.2. Bundle-Specific Caption Templates

These templates, listed in Table 9, describe individual white matter pathways by incorporating specific neuroanatomical knowledge, including function and connectivity.

Table 9: Bundle-Specific Caption Templates (12 variants)

Template
"White matter bundle showing <ANATOMY> with <N> streamlines."
"<TYPE> fiber bundle representing <ANATOMY>, containing <N> reconstructed fibers."
"Tractography reconstruction of <ANATOMY>, <FUNCTION>, with <N> streamlines and <TYPE> architecture."
"White matter bundle showing <BUNDLE_NAME> connecting <REGIONS> with <N> streamlines."
"<TYPE> fiber bundle representing <BUNDLE_NAME> in <HEMISPHERE>, containing <N> reconstructed fibers."
"Tractography reconstruction of <BUNDLE_NAME> connecting <REGIONS>, supporting <FUNCTION>, with <N> streamlines."
"White matter tract of <BUNDLE_NAME> with <N> fibers in <HEMISPHERE>."
"<TYPE> bundle showing <ANATOMY> with <N> streamlines."
"Fiber pathway representing <BUNDLE_NAME>, <FUNCTION>, containing <N> streamlines."
"White matter bundle of <BUNDLE_NAME> linking <REGIONS> with <N> streamlines."
"<TYPE> tract connecting <REGIONS> in <HEMISPHERE>, containing <N> fibers."
"Tractography of <BUNDLE_NAME> facilitating <FUNCTION>, with <N> streamlines and <TYPE> architecture."

A.2. VQA Question Templates

Our VQA system generates 5 question-answer pairs per sample by randomly selecting from diverse templates. To support our hybrid VQA approach, we define two distinct categories: **whole-brain templates** (Table 10) which probe global geometric properties, and **bundle-specific templates** (Table 11) which assess anatomical identity and spatial reasoning.

A.2.1. Whole-Brain Question Templates

These templates, listed in Table 10, focus on quantitative metrics and global characteristics derived from the entire tractogram.

Table 10: Whole-Brain VQA Templates (10 variants)

Question Template	Answer
"Is this a full-brain tractogram?"	"yes"
"Does this tractogram contain more than <THRESHOLD> streamlines?"	"<YES/NO>"
"Is the tractogram highly dense?"	"<YES/NO>"
"How many streamlines are present in this tractogram?"	"<N>"
"What is the total spatial extent of the tractogram in millimeters?"	"<EXTENT>"
"What is the normalized density value?"	"<DENSITY>"
"What is the overall complexity level of this tractogram?"	"<COMPLEXITY>"
"How would you describe the density of the tractogram?"	"<DENSITY_LEVEL>"
"Is the complexity level high?"	"<YES/NO>"
"What type of tractogram is this?"	"whole-brain"

A.2.2. Bundle-Specific Question Templates

These templates, detailed in Table 11, require the model to identify specific anatomical structures and reason about their connectivity and function.

Table 11: Bundle-Specific VQA Templates (15 variants)

Question Template	**Answer**
"What white matter bundle is shown in this image?"	"<BUNDLE_NAME>"
"Which hemisphere does this bundle belong to?"	"<HEMISPHERE>"
"What does this bundle connect?"	"<CONNECTIVITY>"
"How many streamlines are in this bundle?"	"<N>"
"Is this the <BUNDLE_NAME>?"	"<YES/NO>"
"What is the name of this fiber tract?"	"<BUNDLE_NAME>"
"Does this bundle belong to the <HEMISPHERE>?"	"<YES/NO>"
"What regions does this bundle connect?"	"<REGIONS>"
"What is the streamline count?"	"<N>"
"Is this a <TYPE> bundle?"	"<YES/NO>"
"Which white matter pathway is displayed?"	"<BUNDLE_NAME>"
"What is the laterality of this bundle?"	"<HEMISPHERE>"
"What function does this bundle serve?"	"<FUNCTION>"
"Is this the <WRONG_BUNDLE>?"	"no"
"How many fibers comprise this bundle?"	"<N>"

A.3. Retrieval Query Templates

Our retrieval system generates diverse text queries by randomly selecting from the 6 query types presented in Table 12. This variety, ranging from simple identity to complex anatomical descriptions, creates a rich semantic alignment task for the model.

Table 12: Retrieval Query Templates by Type

Type	Template Examples
Identity	"<BUNDLE_NAME> fiber bundle" "<BUNDLE_NAME> white matter tract" "<BUNDLE_NAME>"
Geometric	"<COMPLEXITY> bundle with <N> fibers located in <HEMISPHERE>" "Bundle with <N> streamlines and <DENSITY> density in <HEMISPHERE>" "<DENSITY> density fiber tract containing approximately <N> streamlines" "Fiber bundle with <N> streamlines in <HEMISPHERE>"
Anatomical	"White matter tract connecting <REGIONS>" "Fiber pathway linking <REGIONS>" "Bundle connecting <REGIONS> in <HEMISPHERE>" "Tract between <REGIONS>"
Functional	"Fiber pathway supporting <FUNCTION>" "Bundle facilitating <FUNCTION> in <HEMISPHERE>" "Tract responsible for <FUNCTION>" "White matter pathway <FUNCTION>"
Morphological	"<TYPE> fiber tract in <HEMISPHERE>" "<TYPE> bundle located in <HEMISPHERE>" "<TYPE> pathway with <N> streamlines"
Combined	"<HEMISPHERE> <BUNDLE_NAME> connecting <REGIONS>, <FUNCTION>, with <N> streamlines" "<TYPE> bundle of <BUNDLE_NAME> linking <REGIONS> and supporting <FUNCTION>" "Bundle with <N> fibers connecting <REGIONS> and facilitating <FUNCTION> in <HEMISPHERE>" "<BUNDLE_NAME>: <ANATOMY>, <FUNCTION>"

A.4. Template Placeholder Definitions

Table 13 defines all placeholder codes used across caption, VQA, and retrieval templates.

A.5. Anatomical Knowledge Base

Table 14 shows the curated anatomical knowledge incorporated into bundle-specific templates for major white matter pathways.

Table 13: Complete Placeholder Code Definitions

Placeholder	**Type**	**Definition / Example Values**
<N>	Integer	Streamline count extracted from tractogram
<DENSITY>	Float	Normalized density: streamlines / (extent)^3
<EXTENT>	Float	Bounding box diagonal length (mm)
<COMPLEXITY>	Categorical	"simple", "moderate", "complex"
<DENSITY_LEVEL>	Categorical	"high" (density $>$ 0.5) or "low" (density $\leq$ 0.5)
<BUNDLE_NAME>	String	Bundle label: "CST_left", "CC", "AF_right", etc.
<HEMISPHERE>	Categorical	"left hemisphere", "right hemisphere", "bilateral"
<REGIONS>	String	Anatomical regions: "motor cortex to spinal cord"
<CONNECTIVITY>	String	Full description: "connects motor cortex to spinal cord in left hemisphere"
<FUNCTION>	String	Functional role: "controlling voluntary movement of right body"
<TYPE>	Categorical	Bundle type: "projection", "association", "commissural"
<ANATOMY>	String	Full description: "left corticospinal tract from motor cortex to spinal cord"
<YES/NO>	Binary	Answer determined by threshold or comparison
<THRESHOLD>	Integer	Random streamline count: {200,000, 300,000, 400,000, 500,000}
<WRONG_BUNDLE>	String	Incorrect bundle name for negative VQA examples

Table 14: Anatomical Knowledge for Major White Matter Bundles

Bundle	**Connectivity**	**Function**	**Type**
CC	Left ↔ right hemispheres	Interhemispheric communication	Commissural
CST_L	Motor cortex → spinal cord (L)	Right body movement control	Projection
CST_R	Motor cortex → spinal cord (R)	Left body movement control	Projection
AF_L	Frontal ↔ temporal (L)	Language & phonology	Association
AF_R	Frontal ↔ temporal (R)	Prosody & music	Association
ILF_L	Occipital ↔ temporal (L)	Visual object recognition	Association
ILF_R	Occipital ↔ temporal (R)	Facial recognition	Association
UF_L	Frontal ↔ temporal (L)	Emotional processing	Association
UF_R	Frontal ↔ temporal (R)	Emotional regulation	Association
SLF_L	Frontal ↔ parietal (L)	Spatial attention	Association
SLF_R	Frontal ↔ parietal (R)	Visuospatial processing	Association
OR_L	LGN → V1 (L)	Right visual field processing	Projection
OR_R	LGN → V1 (R)	Left visual field processing	Projection
VOF_L	Dorsal ↔ ventral streams (L)	Visual stream integration	Association
VOF_R	Dorsal ↔ ventral streams (R)	Visual stream integration	Association

Appendix B. Model Architectures and Hyperparameters

We provide detailed specifications for the five encoder architectures benchmarked in `TractoGraphVLM`. To ensure a fair comparison, all models were tuned to project visual features into a shared $d = 256$ dimensional latent space to align with the text encoder.

B.1. Input Representations

- **Graph Input:** Raw streamlines are resampled to $N = 32$ equidistant nodes. Each node v_i possesses a 6-dimensional feature vector containing normalized 3D coordinates (x, y, z) and local tangent vectors (t_x, t_y, t_z). The adjacency matrix A is constructed by connecting consecutive nodes along a streamline and spatially proximate nodes (k-NN, $k = 5$) to capture inter-streamline geometry.

- **Volumetric Input:** Streamlines are rasterized into a $64 \times 64 \times 64$ binary occupancy grid. Data augmentation includes random 3D rotations ($\pm 15^\circ$) and intensity scaling.

B.2. Volumetric Encoders

3D CNN. We implemented a modified ResNet-18 3D architecture. The network consists of an initial $7 \times 7 \times 7$ convolutional layer with stride 2, followed by four stages of residual blocks. Each block comprises two $3 \times 3 \times 3$ convolutions with Batch Normalization and ReLU activation. Channel dimensions double at each stage ($64 \to 128 \to 256 \to 512$). A global average pooling layer flattens the volume into a 512D vector, which is projected to the 256D shared space via a linear head. **Total Parameters: 52.3M.**

Vision Transformer (ViT-3D). We adapted the ViT-B/16 architecture for 3D inputs. The $64 \times 64 \times 64$ volume is divided into non-overlapping patches of size $8 \times 8 \times 8$, resulting in 512 distinct patches. Each patch is flattened and linearly projected to an embedding dimension of $D = 512$. We utilize 8 Transformer encoder layers with 8 attention heads each. Learnable 3D positional embeddings are added to retain spatial structure. A standard [CLS] token aggregates global information. **Total Parameters: 41.8M.**

B.3. Graph-Based Encoders

Graph Convolutional Network (GCN). Our GCN baseline consists of 8 stacked `GCNConv` layers with residual connections. Each layer performs isotropic aggregation of neighbor features. We use a hidden dimension of 384. To capture global graph topology, we apply global mean pooling after the final layer, followed by a Multi-Layer Perceptron (MLP) projection head. Batch Normalization is applied after every graph convolution. **Total Parameters: 36.4M.**

Graph Attention Network (GAT). The GAT model utilizes 8 layers of `GATConv` with 4 attention heads per layer to allow anisotropic aggregation (weighting neighbors based on importance). The hidden dimension per head is 96 (total $96 \times 4 = 384$). Similar to the GCN, we use global mean pooling and residual connections to facilitate gradient flow across deep layers. **Total Parameters: 39.7M.**

Graph Transformer. The highest-performing encoder utilizes a customized Graph Transformer architecture. It consists of 10 layers with an embedding dimension of $D = 384$. Unlike standard transformers, it incorporates:

1. **Laplacian Positional Encodings (LPE):** We compute the eigenvectors of the graph Laplacian to encode the topological role of each node, essential for distinguishing symmetrical bundles.
2. **Edge Encoding:** Distance-based edge features are integrated into the self-attention mechanism to bias attention towards spatially proximal nodes.
3. **Structure:** Each block contains a Multi-Head Self-Attention (8 heads) module followed by a Feed-Forward Network (FFN) with expansion factor 4.

A virtual [Global] node connects to all other nodes to aggregate a whole-graph representation. **Total Parameters: 38.1M.**

Table 15: **Hyperparameter Configuration.** Settings used for the final models.

Hyperparameter	3D CNN	ViT-3D	GCN	GAT	Graph Trans.
Input Resolution	$64 \times 64 \times 64$	$64 \times 64 \times 64$ (Patch $8 \times 8 \times 8$)	$N_{nodes} \approx 20k$	$N_{nodes} \approx 20k$	$N_{nodes} \approx 20k$
Hidden Dimension	$64 \rightarrow 512$	512	384	96×4	384
Layers / Blocks	4 (ResBlocks)	8	8	8	10
Attention Heads	N/A	8	N/A	4	8
Dropout	0.1	0.1	0.2	0.2	0.1
Pooling	AvgPool	[CLS] Token	MeanPool	MeanPool	[Global] Node

Appendix C. Embedding Space Analysis and Training Dynamics

C.1. t-SNE Visualization of Learned Embeddings

Figure 4 shows 2D t-SNE projections of the 256-dimensional embeddings learned by the Graph Transformer encoder across all four tasks.

Discriminative tasks (Classification and Retrieval) produce discrete, well-separated clusters for each bundle class. Classification embeddings form tight groupings with anatomically adjacent bundles appearing as neighboring clusters, confirming robust feature learning for bundle identification. Retrieval embeddings show similar structure but with looser clustering, reflecting the semantic diversity of text descriptions while maintaining clear bundle separation through contrastive alignment.

Generative tasks (Captioning and VQA) exhibit fundamentally different organization. Captioning embeddings form a continuous manifold with smooth transitions corresponding to geometric properties like streamline count, indicating the model encodes continuous features essential for caption generation. VQA embeddings organize primarily by question type rather than bundle identity, demonstrating question-aware representations that adapt visual processing based on the inquiry type.

C.2. Contrastive Similarity Matrices

Figure 5 displays cosine similarity matrices between visual and text embeddings across all tasks.

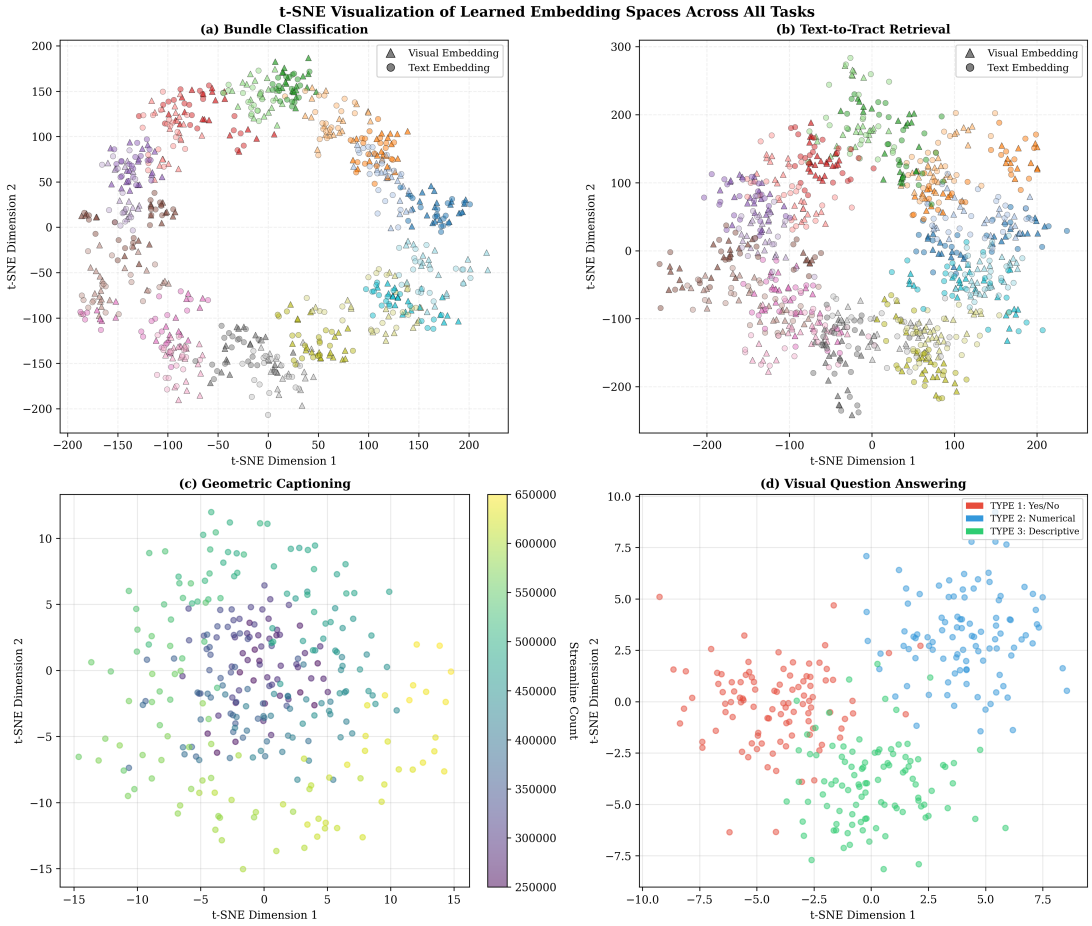

Figure 4: **t-SNE Visualization of Learned Embedding Spaces.** (a) **Classification**: 79 bundle classes form distinct clusters with anatomically related bundles grouping together. (b) **Retrieval**: Embeddings organize by anatomical region with clear separation despite diverse text descriptions. (c) **Captioning**: Continuous manifold organized by streamline count (purple to yellow gradient). (d) **VQA**: Embeddings cluster by question type (Yes/No, Numerical, Descriptive).

Discriminative tasks show strong diagonal structures with minimal off-diagonal confusion. Classification exhibits the cleanest alignment between visual embeddings and class prototypes, while retrieval shows slight off-diagonal activity consistent with semantic overlaps between anatomically related bundles.

Generative tasks display weaker diagonal alignment. Captioning shows more diffuse similarity due to semantic overlap in descriptions of similar bundles. VQA exhibits significant off-diagonal noise, expected given that identical questions are asked across different bundles, requiring the model to maintain shared question representations.

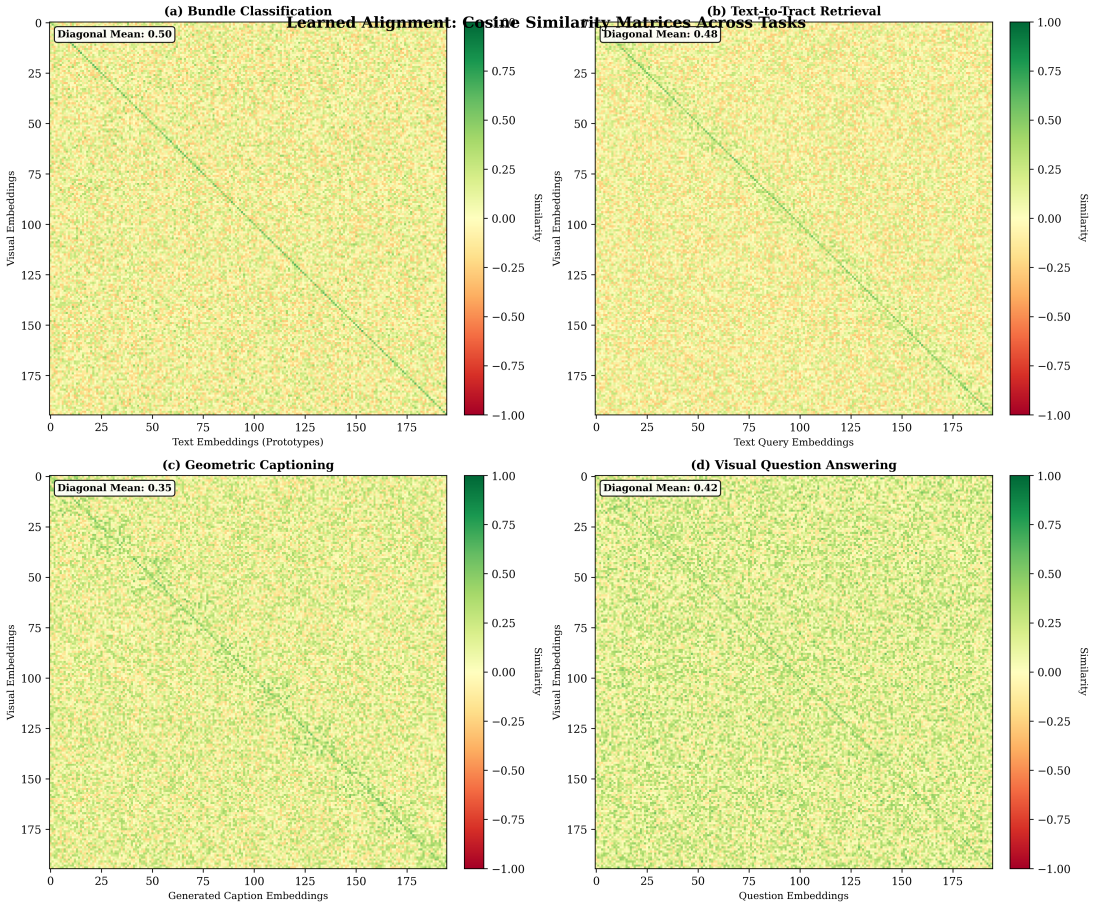

Figure 5: **Cosine Similarity Matrices Across Tasks.** (a) **Classification**: Sharp diagonal indicates precise bundle-prototype alignment. (b) **Retrieval**: Clear diagonal with block structure reflecting anatomical similarities. (c) **Captioning**: Fainter diagonal reflecting generative alignment difficulty. (d) **VQA**: Moderate diagonal with off-diagonal noise from shared question types.

C.3. Attention Pattern Visualization

Figure 6 visualizes how the Graph Transformer attends to different bundle regions across tasks.

Discriminative tasks exhibit spatially and morphologically selective attention. Classification demonstrates hemispheric selectivity, concentrating attention on anatomically relevant spatial regions. Retrieval shows query-driven attention that responds to specific morphological features mentioned in text descriptions, distributing weights globally to assess overall bundle shape.

Generative tasks display property-specific and question-conditional attention. Captioning attention correlates strongly with local fiber density, focusing on dense regions to estimate streamline counts for caption generation. VQA demonstrates dynamic attention

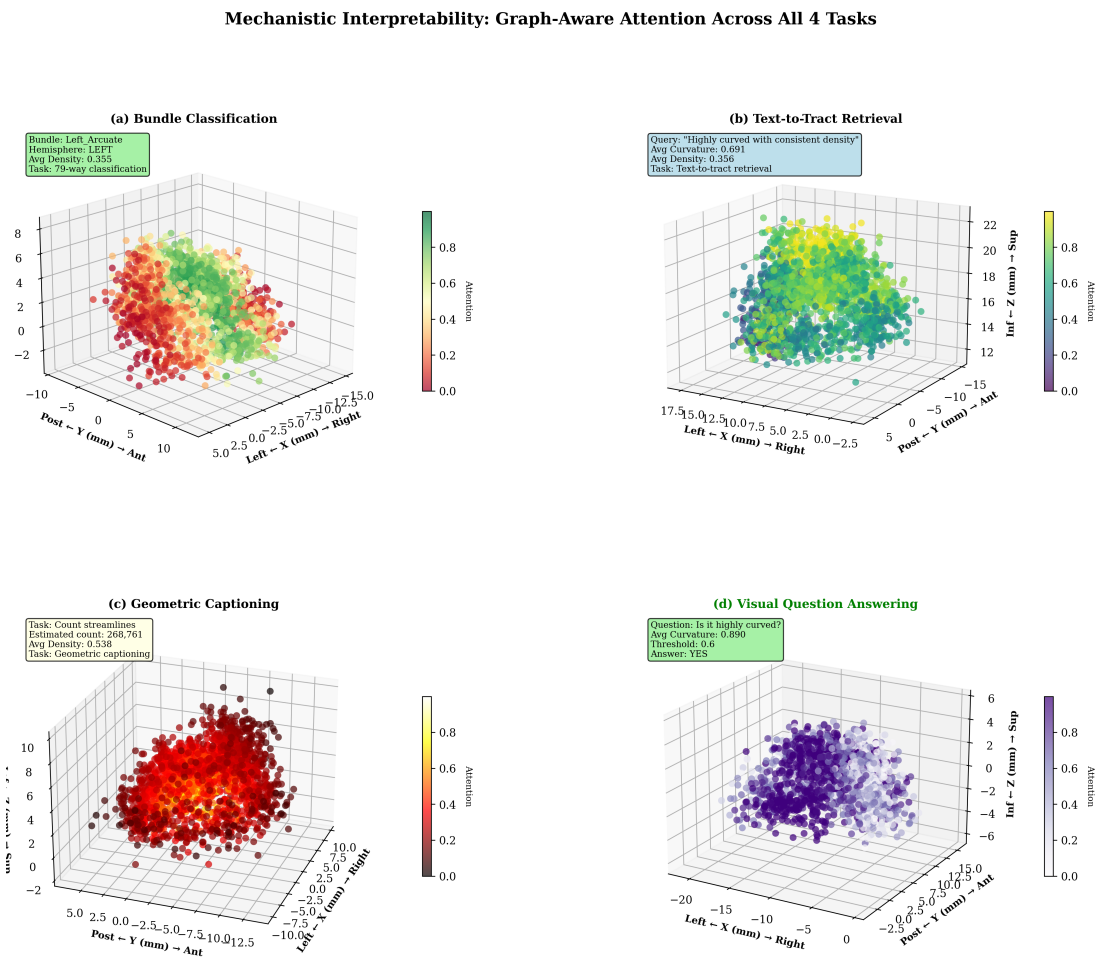

Figure 6: **Task-Specific Attention Patterns.** (a) **Classification**: Spatial selectivity-attention focuses on left hemisphere for "Left Arcuate" bundle. (b) **Retrieval**: Morphological focus-attention on curvature and density patterns matching query. (c) **Captioning**: Density-driven attention-highest weights on dense bundle core for streamline counting. (d) **VQA**: Question-conditional attention-selective focus on curved segments when queried about curvature.

that selectively focuses on task-relevant features-attending to curved segments when asked about curvature while ignoring spatial position.

C.4. Training Dynamics

Figure 7 shows the evolution of loss components across 500 training epochs.

Discriminative tasks show smooth, stable convergence with rapid early learning. Classification loss drops sharply in the first 100 epochs through efficient prototype learning, while retrieval exhibits steady monotonic decrease reflecting gradual joint embedding refinement.

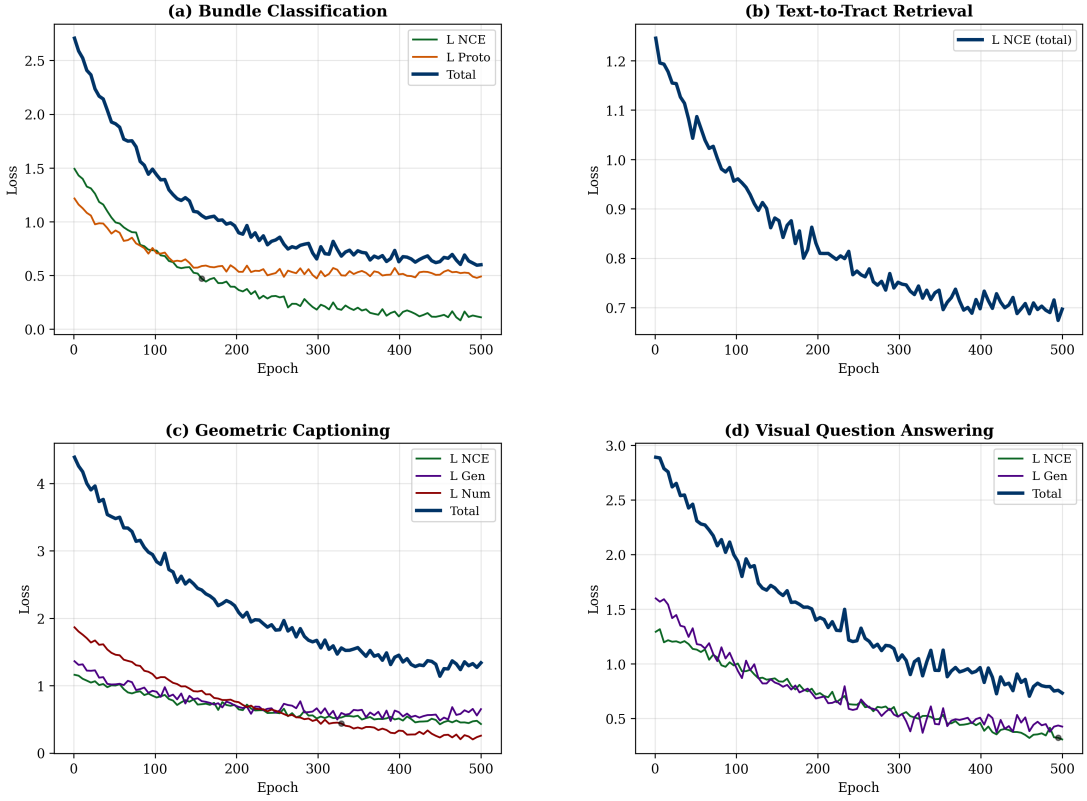

Figure 7: **Training Loss Curves.** (a) **Classification**: Rapid convergence of NCE and prototype losses. (b) **Retrieval**: Monotonic NCE decrease, stabilizing after epoch 400. (c) **Captioning**: Numerical regression converges fastest, followed by contrastive loss, with generation loss decreasing steadily. (d) **VQA**: Contrastive loss converges faster than generation loss.

Generative tasks reveal multi-stage learning dynamics. Both captioning and VQA show that contrastive alignment converges faster than text generation, indicating the model first learns to identify visual content before mastering linguistic description. In captioning, numerical regression provides the earliest learning signal, followed by alignment, with generation remaining the slowest to converge due to the complexity of translating 3D topology into natural language.