TubeMLLM: A Foundation Model for Topology Knowledge Exploration in Vessel-like Anatomy

Yaoyu Liu [*1,2], Minghui Zhang [*3,1], Kefan Wang[1,2], Xin You[1,2], Hanxiao Zhang[1], Qingbiao Li[4], Yirong Chen[3], Yun Gu[1,2 (✉)], and Xinglin Zhang[5 (✉)]

[1] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
[2] School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[4] Faculty of Science and Technology, University of Macau, Macau, China
[5] Medical Image Insights, Shanghai, China
geron762@sjtu.edu.cn; xinglin.zhang@imagecore.com.cn

Abstract. Modeling medical vessel-like anatomy is challenging due to its intricate topology and sensitivity to dataset shifts. Consequently, task-specific models often suffer from topological inconsistencies, including artificial disconnections and spurious merges. Motivated by the promise of multimodal large language models (MLLMs) for zero-shot generalization, we propose TubeMLLM, a unified foundation model that couples structured understanding with controllable generation for medical vessel-like anatomy. By integrating topological priors through explicit natural language prompting and aligning them with visual representations in a shared-attention architecture, TubeMLLM significantly enhances topology-aware perception. Furthermore, we construct TubeMData, a pioneer multimodal benchmark comprising comprehensive topology-centric tasks, and introduce an adaptive loss weighting strategy to emphasize topology-critical regions during training. Extensive experiments demonstrate that, even without task-specific finetuning, TubeMLLM exhibits superior generalization on out-of-distribution vessel datasets. It significantly enhances both topological fidelity and segmentation accuracy, effectively reducing topological errors while maintaining robustness against degradations such as blur, noise, and low resolution. Code is available at https://github.com/EndoluminalSurgicalVision-IMR/TubeMLLM.

Keywords: Multimodal Large Language Models · Topology Learning · Vessel Anatomy.

1 INTRODUCTION

Modeling multi-modality medical vessel-like anatomy (e.g., color fundus retinal vasculature and X-ray coronary angiograms) with topology preserved is fundamental to downstream clinical analysis, including vascular quantification [25],

* Yaoyu Liu and Minghui Zhang contributed equally to this work.

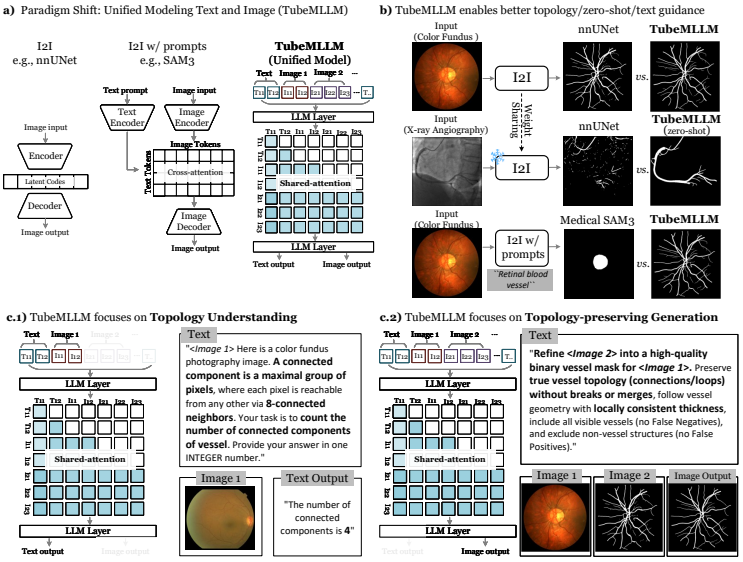

Fig. 1. Overview of TubeMLLM. (a) Unlike task-specific I2I and promptable I2I baselines, TubeMLLM unifies text and image tokens via shared-attention, supporting both text and image outputs. (b) Qualitative comparison showing superior topology preservation, zero-shot cross-modality transfer, and text guidance. (c) Task examples: (c.1) topology-aware visual understanding; (c.2) topology-preserving generation. 8

pathology screening [33], and intervention planning [29]. However, vessel-like structures are inherently difficult because they are thin and elongated, with branching and cyclic connectivity where small local mistakes can cause global topological failures. This challenge is further amplified by dataset shift and modality variations [6]. Consequently, existing methods often fail to achieve both high topological fidelity and strong cross-modality transferability. Specifically, task-specific I2I segmentation models (e.g., nnUNet [9]), illustrated on the left of Fig. 1(a), predict masks solely from visual features. Despite substantial efforts on topology-preserving loss functions [13, 22, 24, 36], these approaches still exhibit limited generalization under distribution shifts and modality variations (top and middle rows of Fig. 1(b)). Recent promptable image-to-image models (I2I w/ prompts) [10] introduce an independent text encoder to inject language guidance during segmentation, as illustrated in the center of Fig. 1(a). However, the input language prompts are typically limited to short phrases such as clinical concepts (e.g., "retinal vessels" in MedicalSAM3 [10]), which is insufficient to encode complex topology priors such as the definition of connectivity or loops. For instance, the bottom row of Fig. 1(b) shows a failure case where the optic disc is misclassified as retinal vessels.

To address these limitations, we move beyond the conventional rigid mapping between images, fixed text labels, and pixel-level masks. Inspired by recent VLM works that models medical or topological knowledge via language [15,32] as

well as MLLM works [4,30] that unifies image generation and understanding, we propose TubeMLLM that couples structured understanding with controllable generation for medical vessel-like anatomy. Instead of relying on simple conceptual mappings, TubeMLLM injects explicit topological knowledge through rich, descriptive text prompts (Fig. 1(c.1) and (c.2)). As illustrated on the right of Fig. 1(a), it accepts interleaved image and text tokens, projects them into a shared feature space, and aligns them through shared-attention within LLM layers, thereby internalizing tubular topological priors and associating them with visual features to strengthen topology-aware perception.

Building on this unified modeling paradigm, we introduce **TubeMData**, a pioneer benchmark for topology-aware multimodal medical anatomy learning. As shown in Fig. 2.(b), TubeMData includes two synergistic tasks: (i) topological understanding, which analyzes structural properties like connectivity and loops, formally characterized by the topological invariants β_0 and β_1, and is formulated as visual question answering; and (ii) topology-preserving generation, which refines or generates vessel-like anatomy with topological consistency and is formulated as image generation. As illustrated in Fig. 1(c.1) and Fig. 1(c.2), unlike promptable image-to-image models, TubeMLLM produces both pixel-level mask and language outputs, and replaces short phrases with longer prompts that explicitly encode imaging modalities and topology definitions, thereby providing sufficient topological guidance. Further, inspired by recent works that perform additional region-level supervision in image generation [8] we incorporate an adaptive loss weighting strategy during training. Specifically, TubeMLLM first establishes spatial correspondence between output pixels and visual tokens, then assigns adaptive weights to each token based on the discrepancy between decoded predictions and ground truth, thereby emphasizing topology-critical and error-prone regions for improved generation performance. TubeMLLM is evaluated on fifteen vessel-like datasets spanning two imaging modalities. Experimental results demonstrate the significant improvement in both topological fidelity and segmentation accuracy, effectively reducing topological errors while maintaining robustness against degradations.

2 Methodology

2.1 Problem Formulation and Overall Framework

The overall architecture of TubeMLLM is illustrated in Fig. 2(a). TubeMLLM takes a multimodal input $\mathcal{X} = \big(\{I_k\}_{k=1}^{K}, T\big)$, where $\{I_k\}_{k=1}^{K}$ denotes the input image set and T is the text prompt, and produces both an image output and a text output, $\mathcal{Y} = \big(Y_{\text{img}}, Y_{\text{text}}\big)$. Specifically, Y_{img} is synthesized in a VAE latent space and decoded as $Y_{\text{img}} = \mathcal{D}(\hat{\mathbf{z}}_0)$, where $\mathcal{D}$ is a frozen VAE decoder and $\hat{\mathbf{z}}_0$ is the predicted clean latent. The text output Y_{text} is generated autoregressively conditioned on $\mathcal{X}$. To support both generation and understanding within a unified framework, TubeMLLM adopts a Mixture-of-Transformers design [4] with two coupled branches, $\mathcal{M} = \{\mathcal{G}_\theta, \mathcal{P}_\phi\}$. The two branches share joint attention at each layer to enable cross-branch information exchange.

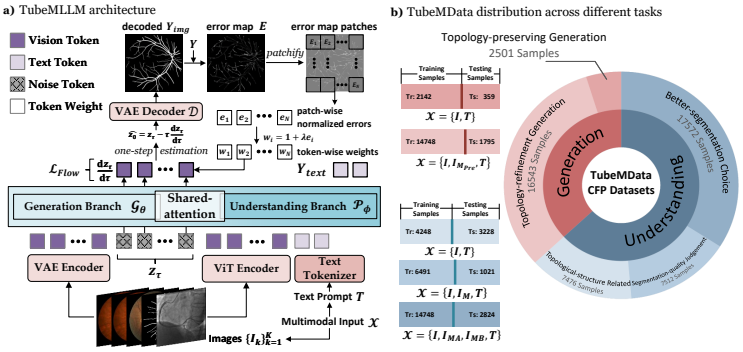

Fig. 2. Detailed TubeMLLM architecture and TubeMData.

Generation branch. $\mathcal{G}_\theta$ operates on tokenized VAE latents and generates images via rectified flow [14]. Image synthesis is formulated as velocity prediction:

$$\frac{\mathrm{d}\mathbf{z}_\tau}{\mathrm{d}\tau} = \mathcal{G}_\theta(\mathbf{z}_\tau, \tau, \mathcal{X}), \tag{1}$$

where $\mathbf{z}_\tau$ denotes the VAE latent at time $\tau \in [0, 1]$. **Understanding branch.** $\mathcal{P}_\phi$ processes visual tokens extracted by a ViT [28] together with text tokens, and models the conditional distribution of the text output as: $p(Y_{\text{text}}|\mathcal{X}) = \mathcal{P}_\phi(\mathcal{X})$.

2.2 Adaptive Loss Weighting

The generative branch $\mathcal{G}_\theta$ is trained using flow matching [16], where the predicted velocity $\mathcal{G}_\theta(\mathbf{z}_\tau, \tau, \mathcal{X})$ is aligned with the target velocity $\mathbf{v}_{\text{target}}$ via a mean-squared error (MSE) loss in latent space. To emphasize topology-critical and error-prone regions, we further derive adaptive loss weights from pixel space, as shown in Fig. 2(a). Specifically, the predicted clean latent $\hat{\mathbf{z}}_0$ is calculated as:

$$\hat{\mathbf{z}_0} = \boldsymbol{\epsilon} - \tau\mathcal{G}_\theta(\mathbf{z}_\tau, \tau, \mathcal{X}), \tag{2}$$

where $\boldsymbol{\epsilon}$ denotes the noise added to $\mathbf{z}_0$ at time τ. The predicted clean latent $\hat{\mathbf{z}}_0$ is then decoded into Y_{img} by the frozen VAE decoder $\mathcal{D}$ as $Y_{\text{img}} = \mathcal{D}(\hat{\mathbf{z}_0})$. Given the ground truth image Y, the pixel-wise error map is derived as $E = Y - Y_{\text{img}}$. As illustrated in Fig. 2(a), E is patchified based on the spatial correspondence between visual tokens and image patches in Y_{img}. For the i-th visual token, we compute within its corresponding patch E_i for the normalized mean error intensity e_i. Its weight w_i is thus defined as $w_i = 1 + \lambda e_i$. λ is set to 10 in all experiments. With token-level weight $w = \{w_i\}_{i=1}^N$, the flow-matching loss for generative branch training is formulated as

$$\mathcal{L}_{\text{Flow}} = \mathbb{E}_{\tau, \mathbf{z}_\tau, \boldsymbol{\epsilon}}\left[\left\| w \odot \left(\mathcal{G}_\theta(\mathbf{z}_\tau, \tau, \mathcal{X}) - \mathbf{v}_{\text{target}}\right)\right\|^2\right]. \tag{3}$$

This adaptive loss weight formulation emphasizes visual tokens associated with pixel errors and further enhances generation quality with flow matching.

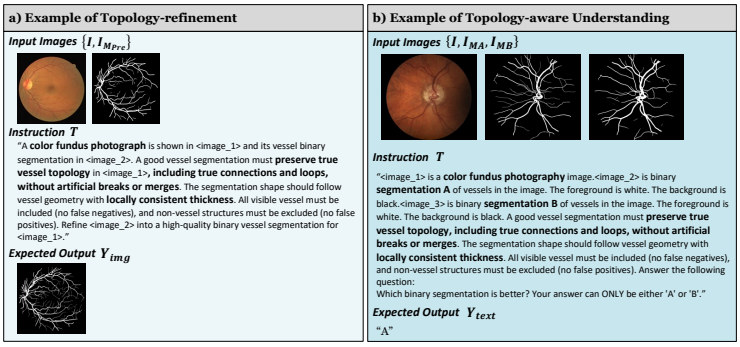

Fig. 3. Topology-centric tasks in TubeMData. (a) Topology-refinement generation. (b) Topology-aware mask selection.

2.3 Topology-centric Task Design

To equip TubeMLLM with topology-preserving generation and topology-aware understanding capabilities, we design a set of topology-centric tasks that explicitly focus on tubular structures. As illustrated in Fig. 3, these tasks target key topological properties of vessel-like anatomy, including connected components, cycles, and the overall topological fidelity of a predicted mask.

Topology-aware Understanding. For topology-aware understanding, we design VQA tasks and encode topology priors through query prompts T, as illustrated in Fig. 3(b). Rather than serving as a stand-alone function for reporting topological quantities to end users, these understanding tasks force the model to explicitly reason about connectivity and loops, thereby injecting topological priors into the shared representation via shared attention, which in turn strengthens the topology-preserving generation branch. These tasks operate at two input granularities. In the first setting, the input is either $\mathcal{X} = \{I, I_{\mathrm{MA}}, I_{\mathrm{MB}}, T\}$ with two candidate masks $I_{\mathrm{MA}}, I_{\mathrm{MB}}$, or $\mathcal{X} = \{I, I_{\mathrm{M}}, T\}$ with a single mask I_{M}. The model produces Y_{text} to (i) select the mask with better topology, or (ii) judge whether the mask satisfies topology requirements. In the second setting, $\mathcal{X} = \{I, T\}$ contains only the image I, and the model $\mathcal{M}$ predicts in Y_{text} the existence or the number of certain topological structures (e.g., connected components, loops).

Topology-preserving Generation. For topology-preserving generation, we introduce topology-refinement task, as shown in Fig. 3(a). $\mathcal{X} = \{I, I_{\mathrm{M_{Pre}}}, T\}$ includes the original image I and a preliminary prediction $I_{\mathrm{M_{Pre}}}$ with imperfect topology. Conditioned on the topology constraints specified in the instruction T, the model $\mathcal{M}$ generates a refined mask Y_{img} that better preserves topological consistency. These topology-centric tasks leverage shared-attention over interleaved image-text inputs, providing dual supervision from Y_{img} and Y_{text} to strengthen topology-aware representations that directly benefit the generation branch.

Table 1. Results on CFP OOD datasets. T-T denotes topology-centric tasks.

Methods	Dice↑	clDice↑	β_0 Num↓	β_0 Mat↓
nnUNet [9]	74.44	78.48	37.42	36.56
nnUNet w/ clDice [24]	75.19	79.83	25.98	27.62
nnUNet w/ SR [13]	74.76	79.68	34.49	30.52
nnUNet w/ cbDice [22]	74.94	79.34	25.46	30.11
nnUNet w/ CAL [36]	72.63	77.85	56.91	29.79
FR-UNet [17]	71.59	76.76	40.13	27.94
AFN [23]	60.35	59.62	24.29	34.65
SAM3 [2]	0.33	0.27	11.08	29.11
MedicalSAM3 [10]	12.14	10.46	10.33	30.74
Bagel [4]	30.68	38.01	307.14	178.09
TubeMLLM w/o T-T	75.73	80.43	**8.47**	26.57
TubeMLLM	**76.09**	**80.59**	8.58	**18.99**

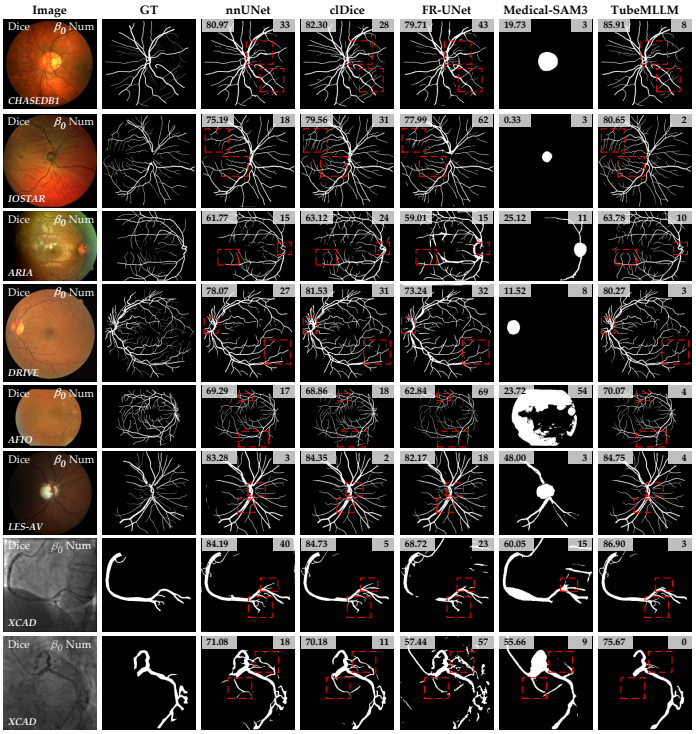

Fig. 4. Qualitative results on CFP and XRA OOD test datasets. β_0 Num denotes the β_0 number error.

3 Experimental Settings and Results

TubeMData. Fifteen vessel-like datasets spanning two imaging modalities are included. Specifically, ten public color fundus photography (CFP) datasets as

Table 2. Comprehensive quantitative results including refinement, cross-dataset, and low quality input settings. ZS denotes zero-shot setting, FS denotes from scratch setting. GB, GN, LR denotes Gaussian Blur, Gaussian Noise, and Low Resolution.

Task	Method	Dice↑	clDice↑	β_0 Num↓	β_0 Mat↓
Task: Single Step Refinement(SSR)					
SSR	nnUNet [9]	74.44	78.48	37.42	36.56
	w/ TubeMLLM	**76.05**	**80.43**	**8.49**	**26.51**
Task: Cross Dataset					
ZS	nnUNet [9]	9.07	11.74	238.26	5.29
	TubeMLLM	**67.50**	**69.17**	**1.21**	**7.69**
FS	nnUNet [9]	77.08	75.43	12.06	0.81
	TubeMLLM	**77.51**	**78.67**	**1.31**	**0.69**
Task: Low Quality Image Input Setting					
GB	nnUNet [9]	78.04	81.98	26.57	21.44
	TubeMLLM	**81.13**	**84.91**	**1.93**	**12.34**
GN	nnUNet [9]	74.77	71.82	18.36	24.10
	TubeMLLM	**77.30**	**73.51**	**1.68**	**13.90**
LR	nnUNet [9]	77.92	81.17	23.61	21.43
	TubeMLLM	**80.25**	**83.53**	**1.68**	**12.32**

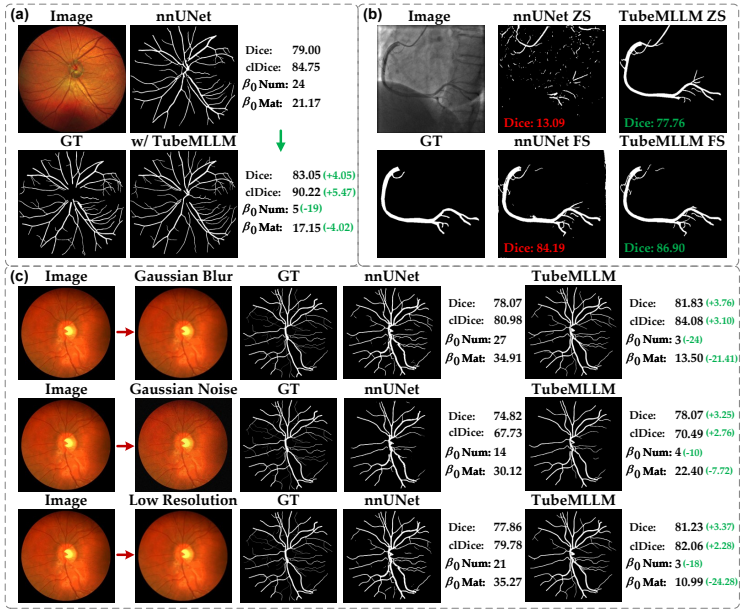

Fig. 5. Qualitative performance on (a) topology-refinement, (b) zero-shot transfer to XRA dataset and (c) degraded input of CFP datasets.

well as five public X-ray angiography (XRA) datasets are included[§]. The ag-

[§] **CFP**: FIVES [11], RETA [18], RVD [12], STARE [7] for training; AFIO [1], ARIA [20], CHASEDB1 [5], DRIVE [26], IOSTAR [35], LES-AV [21] for testing. **XRA**: DCA1 [3], Dr-SAM [37], FS-CAD [34], XCAV [31] for training; XCAD [19] for testing.

gregated corpus contains approximately 3.5K images with heterogeneous spatial resolutions and domain distributions. Topological-imperfect predictions used in Sec 2.3 are generated using nnUNet [9] models trained with different loss functions [13,22,24,36]. For the understanding tasks in Sec. 2.3, ground-truth answers are automatically derived via rule-based verification of topological structures in both manual labels and imperfect masks. Overall, TubeMData contains approximately 52K samples. To assess cross-domain generalization, the test split is strictly out-of-distribution from the training split.

Implementation Details. We initialized TubeMLLM from pretrained checkpoints [4] and finetuned it on TubeMData. The model is trained on 4 NVIDIA H200 GPUs for 90K steps, using AdamW optimizer with learning rate of 2e-5, $\beta_1 = 0.9, \beta_2 = 0.95$. For zero-shot evaluation on XRA modality and low-quality input images, we separately trained TubeMLLM using only the CFP subset in TubeMData. For fair comparison, the nnUNet [9] models, FR-UNet [17] and AFN [23] are trained on the same dataset as TubeMLLM, whereas foundation models (SAM3 [2], MedicalSAM3 [10] and Bagel [4]) are evaluated using their public weights.

Experimental Results Analysis. Generation performance is measured by β_0 number error (β_0 **Num**) and β_0 matching error [27] (β_0 **Mat**) for global and local topological discrepancies, respectively; understanding performance is measured by mean absolute error for counting tasks and classification accuracy for the rest. Tab. 1 evaluates segmentation performance across three distinct paradigms: topology-enhanced nnUNet variants [9, 13, 22, 24, 36], specialized vessel architectures [17,23], and promptable foundation models [2,4,10]. It can be observed that TubeMLLM outperforms all baselines in Dice, clDice, and topological metrics. Specifically, TubeMLLM reduces the β_0 number error to 8.58 and the β_0 matching error to 18.99, a substantial improvement over the $\sim$30 error range typical of nnUNet-based variants. This gap underscores the efficacy of our topology-explicit prompting in enforcing structural consistency. While foundation models like SAM3 and MedicalSAM3 achieve competitive β_0 counts, their markedly lower Dice and clDice indicate that they fail to capture the overall structure of vessel-like anatomy. Fig. 4 provides qualitative visual comparisons where TubeMLLM generates fewer topological errors as highlighted in red boxes. In addition, we evaluated TubeMLLM on topology-refinement, cross-dataset transfer, and zero-shot robustness under low-quality inputs, as summarized in Tab. 2. The results show that TubeMLLM substantially enhances the topological fidelity of imperfect segmentation masks, reducing the β_0 number error from 37.42 to 8.49. Notably, TubeMLLM also exhibits strong zero-shot generalization to the XRA modality, achieving a Dice score of 67.50 while decreasing the β_0 number error from 238.26 to 1.21. Fig. 5(b) further illustrates this improvement. The zero-shot performance of TubeMLLM on XRA successfully captures the coronary vessel structure whereas nnUNet [9] fails. Under low-quality input settings, Tab. 2 indicates that TubeMLLM remains robust in terms of topological fidelity, outperforming nnUNet by approximately 3% in Dice and reducing the β_0 number error by more than 20 across all three degradation scenarios. In Fig. 5(c), we

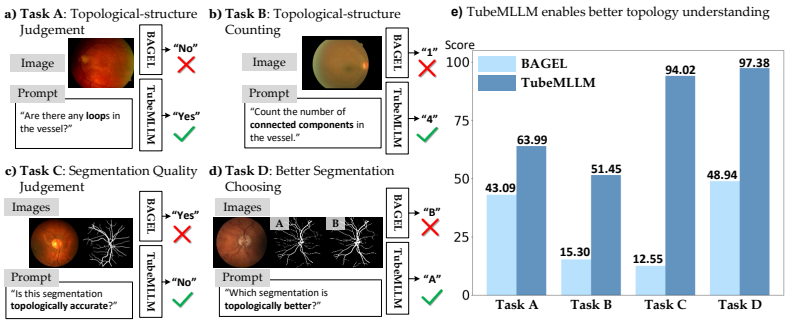

Fig. 6. Qualitative and quantitative performance on four understanding tasks.

can see that TubeMLLM successfully preserves the vessel topology for input images under Gaussian blur, Gaussian noise and low-resolution conditions, while nnUNet fails to do so. Fig. 6(a-d) presents qualitative results of topology understanding tasks. Fig. 6(e) further shows that TubeMLLM improves the ability to count and classify topological structures. Moreover, TubeMLLM achieves 97.38% accuracy in distinguishing good from poor topological quality, a substantial gain over the baseline (48.94%), demonstrating its effectiveness in assessing overall topological quality. Importantly, these understanding tasks are not intended as a stand-alone tool for reporting topological quantities to end users; rather, their gains reflect an improved topology-aware shared representation that, through shared attention, directly benefits the generation branch analyzed above.

4 Conclusion

We propose TubeMLLM, a unified foundation model that integrates vessel-like anatomical priors through language and aligns them with visual representations to enhance topology-preserving perception. Trained on the comprehensive TubeMData with a topology-centric task design and adaptive loss weighting, TubeMLLM substantially improves both topological fidelity and segmentation accuracy, while demonstrating robust generalization to out-of-distribution datasets. Overall, TubeMLLM opens new possibilities for modeling vessel-like anatomy with topological fidelity in a unified multimodal framework.

Acknowledgments. This work is supported by National Key R&D Program of China (2022ZD0212400), Science and Technology Commission of Shanghai Municipality (No.25511102602), Natural Science Foundation of China (62373243), Open Project of the Shanghai Key Laboratory of Flexible Medical Robotics (Grant No. SKLFMR-0211), and Shanghai Tongren Hospital Medical-Engineering Collaboration Project (lhyjzx2024-xm03).

Disclosure of Interests. The authors declare no competing interests.

References

1. Akbar, S., Hassan, T., Akram, M.U., Yasin, U.U., Basit, I.: Avrdb: annotated dataset for vessel segmentation and calculation of arteriovenous ratio. In: IPCV. pp. 129–134 (2017)
2. Carion, N., Gustafson, L., Hu, Y.T., et al.: Sam 3: Segment anything with concepts. arXiv:2511.16719 (2025)
3. Cervantes-Sanchez, F., Cruz-Aceves, I., Hernandez-Aguirre, A., Hernandez-Gonzalez, M.A., Solorio-Meza, S.E.: Automatic segmentation of coronary arteries in x-ray angiograms using multiscale analysis and artificial neural networks. Appl. Sci. **9**(24), 5507 (2019)
4. Deng, C., Zhu, D., Li, K., et al.: Emerging properties in unified multimodal pretraining. arXiv:2505.14683 (2025)
5. Fraz, M.M., Remagnino, P., Hoppe, A., et al.: An ensemble classification-based approach applied to retinal blood vessel segmentation. IEEE Trans. Biomed. Eng. **59**(9), 2538–2548 (2012)
6. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. **69**(3), 1173–1185 (2021)
7. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans. Med. Imaging **19**(3), 203–210 (2000)
8. Huang, X., Zhou, H., Yang, Q., et al.: Jova: Unified multimodal learning for joint video-audio generation. arXiv:2512.13677 (2025)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
10. Jiang, C., Ding, T., Song, C., et al.: Medical sam3: A foundation model for universal prompt-driven medical image segmentation. arXiv:2601.10880 (2026)
11. Jin, K., Huang, X., Zhou, J., et al.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Sci. Data **9**(1), 475 (2022)
12. Khan, M.W., Sheng, H., Zhang, H., et al.: Rvd: a handheld device-based fundus video dataset for retinal vessel segmentation. Adv. Neural Inf. Process. Syst. **36**, 18203–18224 (2023)
13. Kirchhoff, Y., Rokuss, M.R., Roy, S., et al.: Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures. In: ECCV. pp. 218–234. Springer (2024)
14. Labs, B.F., Batifol, S., Blattmann, A., et al.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv:2506.15742 (2025)
15. Li, C., Lux, L., Berger, A.H., Menten, M.J., Sabuncu, M.R., Paetzold, J.C.: Fine-tuning vision language models with graph-based knowledge for explainable medical image analysis. In: MICCAI. pp. 198–207. Springer (2025)
16. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv:2210.02747 (2022)
17. Liu, W., Yang, H., Tian, T., et al.: Full-resolution network and dual-threshold iteration for retinal vessel and coronary angiograph segmentation. IEEE J. Biomed. Health Inform. **26**(9), 4623–4634 (2022)
18. Lyu, X., Cheng, L., Zhang, S.: The reta benchmark for retinal vascular tree analysis. Sci. Data **9**(1), 397 (2022)
19. Ma, Y., Hua, Y., Deng, H., et al.: Self-supervised vessel segmentation via adversarial learning. In: ICCV. pp. 7536–7545 (2021)

20. Niemeijer, M., van Ginneken, B., Staal, J., Suttorp-Schulten, M.S.A., Abràmoff, M.D.: Automatic detection of red lesions in digital color fundus photographs. IEEE Trans. Med. Imaging **24**(5), 584–592 (2005)
21. Orlando, J.I., Barbosa Breda, J., Van Keer, K., et al.: Towards a glaucoma risk index based on simulated hemodynamics from fundus images. In: MICCAI. pp. 65–73. Springer (2018)
22. Shi, P., Hu, J., Yang, Y., Gao, Z., Liu, W., Ma, T.: Centerline boundary dice loss for vascular segmentation. In: MICCAI. pp. 46–56. Springer (2024)
23. Shi, T., Ding, X., Zhou, W., et al.: Affinity feature strengthening for accurate, complete and robust vessel segmentation. IEEE J. Biomed. Health Inform. **27**(8), 4006–4017 (2023)
24. Shit, S., Paetzold, J.C., Sekuboyina, A., et al.: cldice-a novel topology-preserving loss function for tubular structure segmentation. In: CVPR. pp. 16560–16569 (2021)
25. Sianos, G., Morel, M.A., Kappetein, A.P., et al.: The syntax score: an angiographic tool grading the complexity of coronary artery disease. EuroIntervention **1**(2), 219–227 (2005)
26. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**(4), 501–509 (2004)
27. Stucki, N., Paetzold, J.C., Shit, S., Menze, B., Bauer, U.: Topologically faithful image segmentation via induced matching of persistence barcodes. In: ICML. pp. 32698–32727. PMLR (2023)
28. Tschannen, M., Gritsenko, A., Wang, X., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv:2502.14786 (2025)
29. Veneziano, F.A., Cocco, N., Veneziano, R.: Artificial intelligence in interventional cardiology: current applications and future clinical integration. Vessel Plus **9**, N–A (2025)
30. Wang, X., Zhang, X., Luo, Z., et al.: Emu3: Next-token prediction is all you need. arXiv:2409.18869 (2024)
31. Wu, C.H., Chen, S.H., Hu, C.Y., et al.: Denver: Deformable neural vessel representations for unsupervised video vessel segmentation. In: CVPR. pp. 15682–15692 (2025)
32. Xu, M., Hu, Q., Hu, X., et al.: Topo-r1: Detecting topological anomalies via vision-language models. arXiv:2603.13054 (2026)
33. Yu, H., Dong, X.: Ensemble-based eye disease detection system utilizing fundus and vascular structures. Sci. Rep. **15**(1), 19298 (2025)
34. Zeng, Y., Liu, H., Hu, J., Zhao, Z., She, Q.: Pretrained subtraction and segmentation model for coronary angiograms. Sci. Rep. **14**(1), 19888 (2024)
35. Zhang, J., Dashtbozorg, B., Bekkers, E., Pluim, J.P.W., Duits, R., ter Haar Romeny, B.M.: Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. IEEE Trans. Med. Imaging **35**(12), 2631–2644 (2016)
36. Zhang, M., Gu, Y.: Towards connectivity-aware pulmonary airway segmentation. IEEE J. Biomed. Health Inform. **28**(1), 321–332 (2023)
37. Zohranyan, V., Navasardyan, V., Navasardyan, H., Borggrefe, J., Navasardyan, S.: Dr-sam: An end-to-end framework for vascular segmentation diameter estimation and anomaly detection on angiography images. In: CVPR Workshops. pp. 5113–5121 (2024)