CATA: Clinical-Aware Topological Adaptation for Generative Medical VQA

Minh Quang Nguyen*[1,2] and Binh T. Nguyen†[1,2]

[1]University of Science, VNU-HCM
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper presents CATA, a generative medical VQA system for gastrointestinal endoscopy images in MediaEval Medico VQA 2026. CATA combines a frozen ViT/timm visual encoder, Qwen2.5-3B-Instruct with QLoRA, Visual TDA Fusion, and Gated TDA Adapters. Patch-level cubical TDA descriptors are injected into both the visual branch and the final decoder layers to provide structural cues for morphology-aware reasoning. On the 15,955-sample Task 1 full test set, CATA with test adaptation achieves BLEU 0.4771, ROUGE-L 0.6914, METEOR 0.6954, and BERTScore-F1 0.9582. For Task 2, the system outputs structured explanations with self-probes, TDA-derived heatmaps, evidence JSON, and textual rationales. Official evaluation with the Qwen3.6-27B LLM-as-a-judge on the 1,500-sample validation set yields an impressive Correctness of 9.21 and an Overall score of 7.81 on a 0–10 scale.

1 Introduction and Related Work

Medical Visual Question Answering (Medical VQA) requires answering natural-language questions from medical images. Gastrointestinal endoscopy is challenging because relevant evidence can be small, blurry, affected by specular reflections, and embedded in complex mucosal structures. Transformer-based LLMs [1] and Vision Transformers [2] provide strong backbones, but RGB embeddings alone may not consistently emphasize abnormal morphology.

We propose CATA, a generative VQA system that injects patch-level Topological Data Analysis (TDA) descriptors [6, 7] into both the visual branch and the final decoder layers of Qwen2.5-3B-Instruct. The main contribution is a compact Medico VQA pipeline combining ViT/timm, QLoRA-adapted Qwen, Visual TDA Fusion, Gated TDA Adapters, and structured Task 2 explanations with heatmaps, evidence JSON, and reliability signals.

2 Method

2.1 System Overview

Given an image I and a question q, CATA generates an answer y using a frozen ViT/timm visual encoder and Qwen2.5-3B-Instruct adapted with QLoRA [4]. The final system uses patch-level cubical TDA descriptors

$$T \in \mathbb{R}^{14\times14\times C}, \quad C = 12. \tag{1}$$

These descriptors are computed on the resized image grid and used by both Visual TDA Fusion and the decoder-side TDA Adapter.

*23110203@student.hcmus.edu.vn.
†Corresponding author: ngtbinh@hcmus.edu.vn.

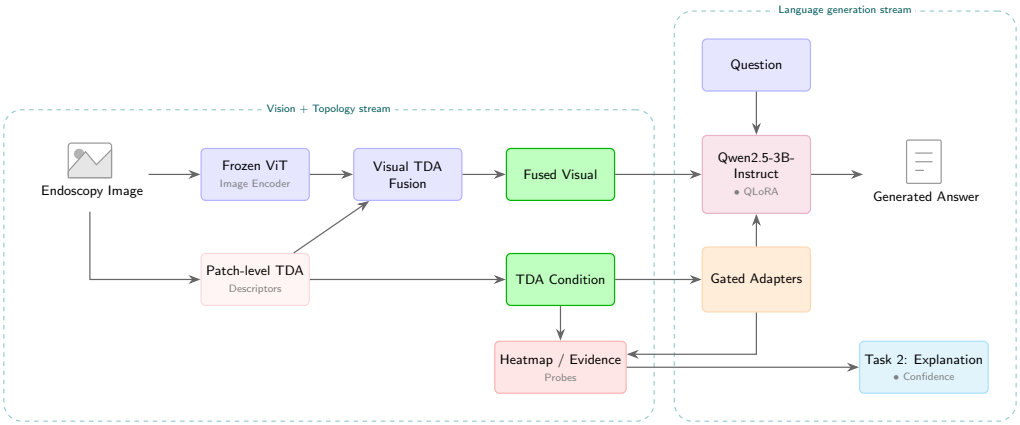

Figure 1: CATA architecture: ViT/timm encodes the image, patch-level cubical TDA descriptors enter Visual TDA Fusion, and aggregated TDA statistics condition Gated TDA Adapters in Qwen2.5.

2.2 Visual TDA Fusion

Visual TDA Fusion combines each ViT token v_i with the corresponding TDA descriptor t_i through residual projection:

$$\bar{v}_i = \mathrm{LN}_v(v_i), \tag{2}$$
$$\tilde{t}_i = \mathrm{MLP}_t(\mathrm{LN}_t(t_i)), \tag{3}$$
$$\hat{v}_i = \mathrm{LN}_{out}\left(\bar{v}_i + \beta \tilde{t}_i\right). \tag{4}$$

The learnable residual scale lets the model use TDA as a local correction while preserving pretrained visual representations.

2.3 Gated TDA Adapter

The Gated TDA Adapter introduces TDA into the LLM. Patch descriptors are summarized by mean, standard deviation, and max statistics:

$$\mathbf{z} = [\mu(T), \sigma(T), \max(T)] \in \mathbb{R}^{36}. \tag{5}$$

The adapter is inserted into the final 8 decoder layers of Qwen. Given hidden state h_l, it computes:

$$\mathbf{t} = W_2\,\mathrm{SiLU}(W_1\mathbf{z}), \tag{6}$$
$$\mathbf{g} = \sigma(W_g\mathbf{t}), \tag{7}$$
$$\Delta\mathbf{h}_l = W_{up}\left(W_{down}\mathbf{h}_l \odot \mathbf{g}\right), \tag{8}$$
$$\mathbf{h}'_l = \mathbf{h}_l + \Delta\mathbf{h}_l. \tag{9}$$

The sigmoid gate controls the residual TDA update, and W_{up} is zero-initialized so the adapter starts near identity.

2.4 Training

CATA optimizes autoregressive cross-entropy with gate regularization:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_{gate}\mathcal{L}_{gate}. \tag{10}$$

Gate regularization discourages persistently ambiguous gates:

$$\mathcal{L}_{gate} = \frac{1}{N}\sum_{i=1}^{N} g_i(1 - g_i). \tag{11}$$

The model uses QLoRA rank 16, LoRA alpha 32, LoRA dropout 0.05, a frozen pretrained ViT/timm encoder, Visual TDA Fusion, and a TDA Adapter bottleneck dimension of 32. We report both train-only and test-adapt checkpoints for transparency.

3 Explanation for Task 2

For Task 2, CATA generates a structured explanation package containing the answer, targeted self-probes, a TDA-derived heatmap, evidence JSON, a short textual explanation, and a reliability-style confidence score. The explanation summarizes the predicted answer, visual evidence, probe agreement, and possible uncertainty. The confidence score is used only as a review signal, not as a calibrated clinical probability.

4 Experiments

4.1 Dataset, Metrics, and Model Setup

We evaluate CATA on Kvasir-VQA-x1 [14], including the 15,955-sample full test set, the 1,500-sample Public Leaderboard setting, and the final Private Challenge Set. Metrics include BLEU [9], ROUGE [10], METEOR [11], chrF++ [13], and BERTScore-F1 [12] when references are available. CATA uses Qwen2.5-3B-Instruct, QLoRA rank 16, LoRA alpha 32/dropout 0.05, a frozen ViT/timm encoder, Visual TDA Fusion, and Gated TDA Adapters in the final 8 decoder layers.

4.2 Ablation Study on TDA Modules

Table 1: Ablation of TDA modules on full-test.

Model	BLEU	R-1	R-2	R-L	METEOR	chrF++	BERT-F1
Baseline	0.354	0.613	0.410	0.570	0.586	0.565	0.942
Visual TDA	0.373	0.646	0.442	0.604	0.620	0.587	0.947
TDA Adapter	0.447	0.701	0.513	0.673	0.683	0.647	0.956
Visual TDA + TDA Adapter	**0.451**	**0.706**	**0.518**	**0.677**	**0.687**	**0.651**	**0.956**

The ablation indicates that TDA is not merely an auxiliary signal. Visual TDA alone improves the image-only baseline from BLEU 0.354 to 0.373 and ROUGE-L 0.570 to 0.604, showing that local topological descriptors help the visual branch represent endoscopic structures. The TDA Adapter gives the largest gain, reaching BLEU 0.447 and ROUGE-L 0.673, while the combined system reaches the best BLEU, ROUGE-L, METEOR, chrF++, and BERTScore-F1. This suggests that patch-level fusion and decoder-side conditioning are complementary.

4.3 Evaluation of Convergence and Test Adaptation

Table 2: Comparison of CATA performance by evaluation set and checkpoint. Public and Private report only the Epoch 3 + Test Adaptation configuration.

Setting	BLEU	R-1	R-2	R-L	METEOR	chrF++	BERT-F1
Public Leaderboard (1,500 samples)							
Epoch 3 + T est Adaptation	0.478	0.719	0.536	0.692	0.698	–	–
Kvasir-VQA-x1 Full-test (15,955 samples)							
Epoch 2	0.472	0.715	0.531	0.688	0.692	0.657	0.958
Epoch 3	0.473	0.715	0.532	0.688	0.693	0.658	0.958
Epoch 3 + Test Adaptation	**0.477**	**0.718**	**0.536**	**0.691**	**0.695**	**0.661**	**0.958**
Private Challenge Set							
Epoch 3 + Test Adaptation	–	–	–	–	–	–	–

On full-test, Epoch 2 and Epoch 3 are nearly saturated, while test adaptation gives a clearer improvement, reaching BLEU 0.477, ROUGE-L 0.691, and METEOR 0.695. The improvement is modest but consistent across BLEU, ROUGE, METEOR, and chrF++, indicating better alignment with the target distribution rather than a single-metric fluctuation. The Public Leaderboard result is close to the full-test trend, supporting Epoch 3 + Test Adaptation as the final reported configuration while still separating it from train-only checkpoints.

4.4 Evaluation by Complexity Level

We further decompose Task 1 into three complexity levels. CATA Epoch 3 + Test Adaptation performs best on Levels 2 and 3, where questions require more visual context and relation reasoning, while Epoch 3 keeps the highest METEOR on Level 1. This pattern suggests that adaptation is most useful for visually demanding questions, whereas direct recognition questions are already close to saturation before adaptation.

Table 3: Task 1 performance by complexity level on the full-test set.

Model	Level 1		Level 2		Level 3	
	METEOR	ROUGE-L	METEOR	ROUGE-L	METEOR	ROUGE-L
Baseline	0.573377	0.583520	0.559433	0.548706	0.625298	0.577214
Visual TDA	0.632093	0.638241	0.580144	0.574124	0.648000	0.596493
TDA Adapter	0.682265	0.699550	0.648984	0.643863	0.717231	0.674029
Visual TDA + TDA Adapter	0.684252	0.705289	0.654763	0.648470	0.721452	0.677068
CATA Epoch 3	**0.685590**	0.710964	0.662625	0.661651	0.730188	0.691651
CATA Epoch 3 + Test Adaptation	0.685340	**0.711067**	**0.667287**	**0.665878**	**0.734149**	**0.696235**

4.5 Task 2 Evaluation: Multimodal Explanations

For Task 2, we evaluate the generated explanation package using the official Qwen3.6-27B LLM-as-a-judge on the 1,500-sample Validation Set. The judge evaluates five independent criteria on a 0–10 scale: Correctness, Faithfulness, Clinical Relevance, Clarity, and Completeness.

Table 4: Task 2: Official LLM-as-a-Judge Evaluation (Qwen3.6-27B) on the Validation Set.

Model	Correctness	Faithfulness	Clin. Relevance	Clarity	Completeness	Overall
CATA-Final	9.21	7.16	7.26	7.49	7.78	7.81

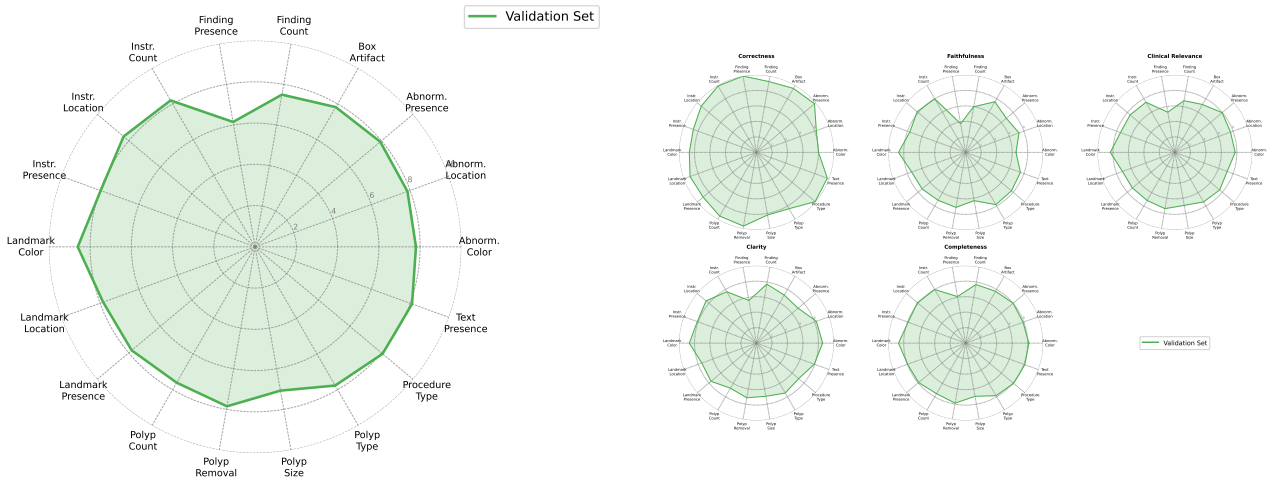

Figure 2: Official Task 2 evaluation results on the Validation Set. Left: Mean LLM Judge score per question class. Right: Breakdown of scores across the five evaluation dimensions.

The results in Table 4 and Figure 2 demonstrate that CATA-Final excels in core factual accuracy, achieving an outstanding Correctness score of 9.21. The radar charts illustrate that the model maintains this high accuracy consistently across diverse question classes, including complex queries like abnormality presence and instrument count. This indicates that despite the generative nature of the LLM, the topological fusion successfully grounds the model to answer the primary clinical questions accurately.

The system also shows strong Completeness (7.78) and Clarity (7.49), reflecting the utility of providing a structured output comprising heatmaps, evidence JSON, and textual reasoning. However, Faithfulness (7.16) and Clinical Relevance (7.26) represent the main areas for improvement. Qualitative analysis of the judge's reasoning reveals that while the primary answers are highly accurate, the model's internal self-probing explanations occasionally exhibit logical contradictions. Addressing this internal reasoning alignment remains a critical direction for future work.

4.6 Clinical Safety Evaluation (Clinical Reliability & Calibration)

Because clinical use requires uncertainty awareness, we also evaluate the reliability-style confidence score with Expected Calibration Error (ECE) [8]. Empirical accuracy is computed by treating a Qwen2.5-72B Correctness score ≥ 3 as correct:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathrm{Acc}(B_m) - \mathrm{Conf}(B_m)|, \quad M = 10. \tag{12}$$

On 300 judged Task 2 samples, CATA obtains ECE = 0.131. This indicates moderate calibration for prioritizing review, but the confidence score should not be interpreted as a clinical diagnostic probability.

5 Discussion and Outlook

CATA illustrates the potential of combining a frozen visual encoder, a QLoRA-adapted decoder, and topological descriptors for specialized endoscopic VQA. The system performs best when questions require localized structural evidence, while more complex multi-clause questions and explanation clarity remain harder. Future work should emphasize stronger decoder-side reasoning, expert-annotated explanation evaluation, stricter calibration, and validation across broader endoscopic distributions.

6 Generative AI Use Declaration

This paper was assisted by generative AI tools, including ChatGPT and Overleaf's autocomplete, for grammar refinement, vocabulary improvement, and readability editing. All experiments, metrics, analyses, and final technical claims were checked and controlled by the authors.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *ICLR*, 2022.

[4] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Advances in Neural Information Processing Systems*, 2023.

[5] Qwen Team, "Qwen2.5 technical report," arXiv preprint, 2025.

[6] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.

[7] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological persistence and simplification," *Discrete & Computational Geometry*, vol. 28, pp. 511–533, 2002.

[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *ICML*, 2017.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *ACL*, 2002.

[10] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004.

[11] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures*, 2005.

[12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *ICLR*, 2020.

[13] M. Popović, "chrF: Character n-gram F-score for automatic MT evaluation," in *WMT*, 2015.

[14] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, M. Taschwer, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *MMSys*, 2017.

[15] S. Gaihre and A. Thapa Magar, "From answers to explanations: Self-probing efficiently fine-tuned vision–language models for medical VQA at Medico 2025," 2025.