Fusion at the Foregut: CLIP-Based Prototypical Learning with DINOv2 Refinement for Endoscopic Image Analysis

Tuan-Anh Yang
ytanh21@apcs.fitus.edu.vn
VNU-HCM University of Science
Vietnam National University
Ho Chi Minh City, Vietnam

Abstract

This paper presents a multimodal framework for anatomical site classification and retrieval in endoscopic images, integrating CLIP-based prototypical learning with DINOv2 refinement. BiomedCLIP embeddings are used to form joint vision-language prototypes for coarse classification, while DINOv2, pretrained on SurgeNet, handles fine-grained binary classification (Ear/Nose side, Vocal Cord state). A hybrid inference strategy combines semantic alignment from BiomedCLIP with the visual precision of DINOv2, fallbacking to DINOv2 for visually ambiguous or out-of-distribution samples. On the ENTRep Challenge datasets, our model achieved 99.46% accuracy and 0.9954 macro F1-score for coarse-level classification (4 classes), and 94.09% accuracy with 0.9442 macro F1 for fine-grained binary classification (7 classes), outperforming both a DINOv2 baseline and the prototype-only method. Our approach placed $\mathbf{5^{th}}$ in image classification (91.64% accuracy), $\mathbf{3^{rd}}$ in image-image retrieval (88.53% Recall@1), and $\mathbf{6^{th}}$ in text-image retrieval (84.00% Recall@1), demonstrating competitive performance across all tracks. Our code is available at https://github.com/YangTuanAnh/ENTRep.

CCS Concepts

• **Computing methodologies** → **Visual content-based indexing and retrieval**; *Computer vision representations.*

Keywords

endoscopic image analysis, fine-grained visual classification, multimodal retrieval, CLIP, DINOv2, prototypical networks, medical image classification, image-text alignment

ACM Reference Format:
Tuan-Anh Yang. 2025. Fusion at the Foregut: CLIP-Based Prototypical Learning with DINOv2 Refinement for Endoscopic Image Analysis. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (ACMMM '25).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

ACMMM '25, Dublin, IE

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Endoscopic imaging is vital in medical diagnostics, offering high-resolution visualization of internal anatomical structures. Automating the analysis of these images supports tasks like anatomical site classification and multimodal retrieval, but remains challenging due to visual symmetry (e.g., nose-left" vs. nose-right"), data imbalance, and limited annotations [11].

To address these issues, we propose a hierarchical framework combining vision-language pretraining and self-supervised learning. Using the ENTRep Challenge datasets, we perform prototype-based classification with BiomedCLIP [14] embeddings at both coarse (4-class) and fine-grained (7-class) levels, showing improved generalization via semantic abstraction.

To better handle symmetric categories, we fine-tune DINOv2 [9], initialized from the surgical foundation model SurgeNet [5]. This model applies subclass decomposition (e.g., left vs. right) and strong data augmentation for enhanced discrimination. We further introduce a hybrid inference strategy that fuses predictions from BiomedCLIP and DINOv2, effectively balancing semantic alignment with robust visual recognition.

Our contributions are: (1) a hierarchical classification framework integrating BiomedCLIP and DINOv2; (2) a hybrid inference method combining semantic and visual cues; and (3) extensive validation on the ENTRep datasets, where our team ranked 5th in classification, 3rd in image retrieval, and 6th in text-to-image retrieval.

2 Related Works

2.1 Endoscopic Image Analysis and Classification

Deep learning has significantly advanced endoscopic image analysis, with CNNs such as ResNet, DenseNet, EfficientNet, and their lightweight variants like MobileNet enabling accurate classification of anatomical structures and pathological findings across gastrointestinal and urological domains [11]. Hybrid CNN-transformer models are also gaining traction for real-time applications. Interpretability remains a key concern in clinical settings. Techniques like Grad-CAM have been integrated into CNNs to visualize decision-critical regions, as demonstrated by a ResNet-152 model achieving 93.46% accuracy on the Kvasir dataset with interpretable heatmaps [8]. Multi-class classification frameworks, such as EfficientNetB3 achieving over 94% accuracy on gastrointestinal lesion classification, illustrate the robustness of modern pipelines [6]. Furthermore, open-set recognition (OSR) methods combining CNNs and transformers have shown improved generalization by detecting unfamiliar classes at test time [1].

2.2 Multimodal Learning and Foundation Models

Multimodal foundation models have redefined medical AI by aligning visual and textual information in shared embedding spaces. CLIP [12] pioneered this direction, enabling zero-shot and retrieval tasks from large-scale image-text pairs. However, general-purpose CLIP models struggle with domain-specific content. BiomedCLIP [14], pretrained on 15 million biomedical image-text pairs, outperforms general CLIP models in classification, retrieval, and grounding tasks, emphasizing the value of specialized training data. In surgical vision, SurgeNet [5] applies self-supervised learning to diverse unlabeled surgical videos, achieving better generalization than traditional transfer learning. These efforts highlight the importance of domain-aligned, annotation-efficient pretraining for robust performance in clinical settings. Together, these trends underscore the shift toward scalable, multimodal models capable of supporting diverse downstream tasks in endoscopic and surgical AI.

3 Dataset

To evaluate our hierarchical classification framework, we utilize two endoscopic image datasets: the *Public Release Dataset* and the *Public Contest Dataset*. The class distributions for both datasets, along with their combination, are summarized in Table 1. Each dataset consists of 640×480 endoscopic images curated for the purpose of anatomical site classification and multi-modal retrieval. Every image is annotated with one of 7 anatomical classes: *nose-left*, *nose-right*, *vc-open*, *vc-closed*, *ear-left*, *ear-right*, and *throat*. In addition to class labels, each image is accompanied by bilingual diagnostic text descriptions (in English and Vietnamese), which support multilingual vision-language tasks such as cross-modal retrieval and caption grounding. These descriptions provide detailed clinical observations corresponding to the imaged anatomical sites.

Public Release Dataset. This dataset is made publicly available for initial exploration and benchmarking. It consists of 566 images with moderate class imbalance.

Public Contest Dataset. This dataset is reserved for official competition use and model evaluation. It comprises 1,291 images with a heavier skew toward certain anatomical regions (e.g., *nose-right* and *ear-right*).

Table 1: Class distribution across the Public Release, Public Contest (Training), and Combined Datasets.

Class	Pub. Release	Pub. Contest	Combined
nose-left	180	290	470
nose-right	128	325	453
vc-open	80	159	239
vc-closed	62	147	209
ear-left	45	133	178
ear-right	44	156	200
throat	27	81	108
Total	**566**	**1,291**	**1,857**

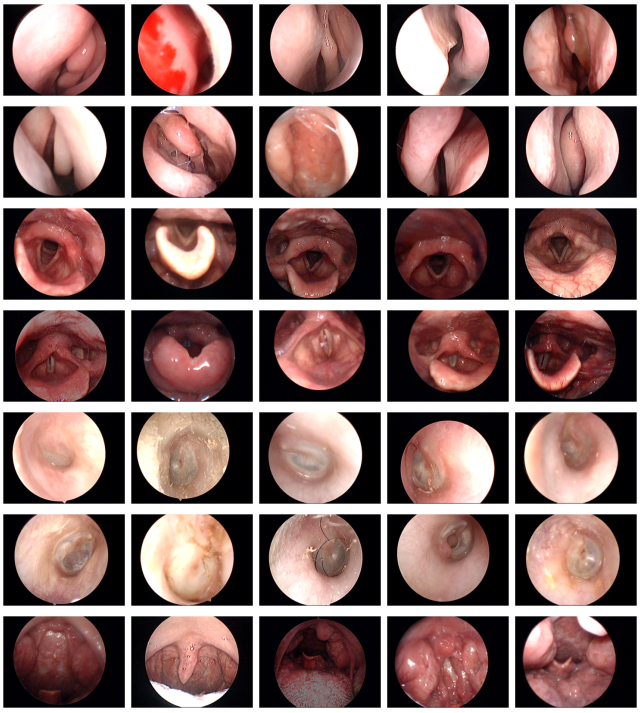

Figure 1: Representative examples for each anatomical class: *nose-left, nose-right, vc-open, vc-closed, ear-left, ear-right,* and *throat* (from top to bottom).

4 Method

4.1 Preprocessing and Dataset Split

The dataset is categorized into 7 fine-grained classes (`ear-left`, `ear-right`, `nose-left`, `nose-right`, `vc-open`, `vc-closed`, `throat`). To improve model performance, we merged symmetrically similar classes into a 4-class coarse classification taxonomy:

- `ear-left`, `ear-right` → `ear`
- `nose-left`, `nose-right` → `nose`
- `vc-open`, `vc-closed` → `vc`

We split the dataset into training, validation, and test sets using an 80-10-10 stratified split based on the fine-grained labels, ensuring balanced class distributions at both the fine and coarse levels. The resulting data distributions are summarized in Table 2.

4.2 Prototype-based Classification: Fine- vs. Coarse-level

We investigate prototype-based classification using multimodal embeddings from the BiomedCLIP model [14], which is pretrained on biomedical image-text pairs for joint vision-language representation learning. Specifically, we leverage BiomedCLIP's image encoder to extract fixed-dimensional visual features from image tiles and combine these with corresponding textual descriptions to obtain a fused embedding for classification.

To generate a joint image-text representation, we compute a weighted combination of the visual and textual embeddings. For an

Table 2: Train/Validation/Test split for 7-class fine-grained anatomical classification.

Class	Train	Val	Test
nose-left	376	47	47
nose-right	362	45	46
vc-open	191	24	24
vc-closed	167	21	21
ear-right	160	20	20
ear-left	142	18	18
throat	87	11	10
Total	**1,485**	**186**	**186**

image x, the visual embedding $\mathbf{v} = f_{\text{img}}(x) \in \mathbb{R}^d$ is extracted via BiomedCLIP's image encoder f_{img}. Simultaneously, a class-specific description t is encoded as $\mathbf{u} = f_{\text{text}}(t) \in \mathbb{R}^d$, where t is either a human annotation or a fallback prompt. The cosine similarity between the modalities is calculated as:

$$s = \cos(\mathbf{v}, \mathbf{u}) = \frac{\mathbf{v} \cdot \mathbf{u}}{\|\mathbf{v}\|\|\mathbf{u}\|}, \quad s \in [-1, 1]. \tag{1}$$

We then compute the fused embedding $\mathbf{z} \in \mathbb{R}^d$ as:

$$\mathbf{z} = s \cdot \mathbf{v} + (1 - s) \cdot \mathbf{u}. \tag{2}$$

Here, s serves as a soft alignment weight that dynamically modulates the influence of each modality. When the image and text are well-aligned (i.e., high cosine similarity), the final embedding leans toward the visual modality; otherwise, it is regularized by the text prior. This mechanism stabilizes representations in visually ambiguous or low-context scenarios and enables zero-shot classification against a predefined set of prototypes.

Fallback descriptions used when no human-annotated text is available are summarized in Table 3.

Table 3: Fallback descriptions for prototype generation.

Label	Fallback Description
nose-left	An endoscopic image of the left nasal passage.
nose-right	An endoscopic image of the right nasal passage.
ear-left	An endoscopic image of the left ear.
ear-right	An endoscopic image of the right ear.
vc-open	An endoscopic image of the vocal cords in the open state.
vc-closed	An endoscopic image of the vocal cords in the closed state.
throat	An endoscopic image of the throat region.

Initially, we constructed seven fine-grained prototypes—one per anatomical label (e.g., *nose-left*, *ear-right*, *vc-open*). Each image was classified to the nearest prototype in cosine similarity space using FAISS [3]. However, performance was limited by visual symmetry between paired classes (e.g., left/right, open/closed), yielding 55.91% accuracy and a macro F1-score of 0.5756 (see Table 4).

To mitigate these ambiguities, we regrouped the dataset under four coarse-level categories—*ear*, *nose*, *throat*, and *vocal cord*. These prototypes better align with consistent anatomical structures. Without retraining or finetuning, coarse-level classification achieved 90.86% accuracy and 0.8542 macro F1 (Table 5), indicating improved generalization via semantic abstraction.

This improvement is also visually confirmed in the t-SNE plots shown in Figure 2. While fine-grained embeddings yield overlapping clusters, the coarse-label projections exhibit well-separated groupings, consistent with the higher classification performance.

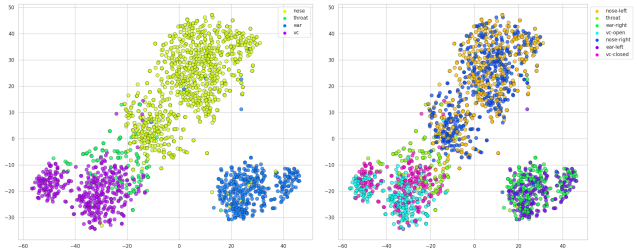

Figure 2: t-SNE visualization of BiomedCLIP embeddings colored by (left) coarse class and (right) fine-grained label. Coarse labels exhibit stronger cluster separation, aligning with improved classification performance.

4.3 Fine-grained Classification with DINOv2

To address the challenge of fine-grained anatomical site classification—particularly among symmetrically similar structures such as *ear-left* vs. *ear-right*, *nose-left* vs. *nose-right*, and *vc-open* vs. *vc-closed*—we employ a supervised learning framework based on the DINOv2 vision transformer [9]. Our implementation initializes DINOv2 with weights from SurgeNet [5], a surgical foundation model pretrained on over 4.7 million frames from laparoscopic videos. SurgeNet demonstrates state-of-the-art performance across a range of surgical vision tasks, exhibiting robustness in visually complex and data-scarce scenarios—properties critical for endoscopic image analysis.

4.3.1 Symmetric Subclass Decomposition. To better model fine-grained distinctions, we decompose the classification task into three binary problems: **(i)** *Ear*: left vs. right, **(ii)** *Nose*: left vs. right, and **(iii)** *Vocal Cords*: open vs. closed. Each binary classifier is trained independently, allowing the model to specialize in learning subtle morphological and directional cues within its respective anatomical category.

4.3.2 Augmentation Strategy. To promote generalization and robustness under clinically realistic variation, we apply spatial and photometric augmentations during training. These include fixed resizing, random small-angle rotations to mimic endoscope articulation, Gaussian blur to simulate motion/focus variability, and color jittering to reflect illumination changes. All transformations are implemented using the Torchvision [7] library and applied exclusively during training.

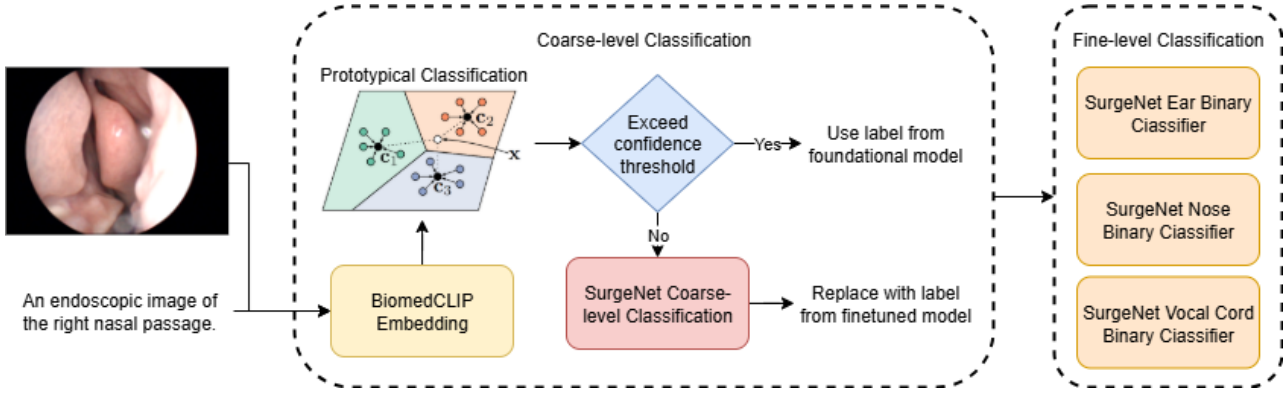

Figure 3: Overview of the hybrid classification framework. The endoscopic image is processed by both a CLIP-based prototypical learner and a DINOv2-based MLP classifier. A hybrid inference mechanism selects the final prediction based on confidence.

4.3.3 Model Architecture. Each binary classifier comprises a lightweight MLP head atop a frozen DINOv2 backbone initialized with SurgeNet weights. Input RGB images are processed by the vision transformer, whose patch embeddings are mean-pooled to form a fixed-size feature vector. This vector is passed through a classification head consisting of a LayerNorm [2] layer, dropout [13] (rate = 0.3), and a sequence of fully connected layers with GELU activations [4]: first projecting to 256 dimensions, then 128, followed by a final output layer corresponding to the number of target classes. This architecture maintains efficiency while benefiting from the inductive priors of large-scale pretraining.

4.3.4 Coarse-level Extension. Beyond fine-grained classification, we also train a four-class DINOv2 model to predict coarse anatomical categories: *ear*, *nose*, *throat*, and *vocal cords*. This model shares the same architecture and training pipeline as the binary classifiers and serves as a complementary baseline to the BiomedCLIP-based prototype classifier.

4.3.5 Hybrid Inference Strategy. To improve robustness against visually ambiguous or out-of-distribution (OOD) inputs, we introduce a hybrid inference strategy that integrates predictions from BiomedCLIP and DINOv2. At inference time, each image is first evaluated by the BiomedCLIP prototype classifier. If the model's confidence exceeds a predefined threshold, the CLIP-based prediction is accepted. Otherwise, the image is deferred to the appropriate DINOv2 classifier (binary or coarse) for final prediction. The strategy is illustrated by Figure 3.

5 Experiments

5.1 Implementation Details

5.1.1 Training Details. All models are trained and evaluated using PyTorch [10] on a single Nvidia Tesla P100 GPU with 16 GB VRAM. For prototype-based retrieval, we employ FAISS [3] for efficient similarity search. Binary classifiers are implemented as independent logistic regression heads and trained per class using oversampled positive instances.

5.1.2 Inference Pipeline. The system performs classification in three stages. First, cosine similarity is computed between the query embedding and class prototypes using a CLIP-style embedding index. If the top-1 similarity score exceeds a confidence threshold (set to 0.99), the corresponding class is returned. Otherwise, the system defers to a fine-grained classifier to reassign the prediction. Finally, class-specific binary classifiers are conditionally invoked for refining ambiguous predictions.

5.1.3 Baseline: Direct DINOv2 Classifier. To assess the effectiveness of our prototype-based hybrid model, we compare it against a baseline that uses the same MLP architecture applied directly to pooled DINOv2 features. This baseline still utilizes features extracted from the SurgeNet-pretrained DINOv2 encoder, but does not leverage any prototype retrieval, similarity-based refinement, or multimodal pretraining strategies. It predicts all 7 classes using a single MLP head trained end-to-end with cross-entropy loss, serving as a controlled comparison to isolate the effect of prototype-based enhancement.

5.2 Evaluation Results

We evaluate the proposed multimodal classification framework on two tasks: coarse-level classification (4 classes) and fine-grained binary classification (7 classes). Table 7 and Table 8 report the accuracy, precision, recall, and F1-score for each class.

The coarse classification task involves categorizing endoscopic images into four anatomical regions: ear, nose, throat, and vocal cords (vc). On this task, our model achieves an overall accuracy of 98.92%, with a macro-averaged F1-score of 0.9807. Performance is uniformly strong across all classes, with particularly high scores for ear and vocal cords (F1 $\geq$ 0.99). Throat achieves perfect recall despite its relatively limited sample size, highlighting the model's strong generalization capability even under class imbalance.

In the fine-grained binary classification task, each anatomical region is further split into subcategories, yielding a total of seven classes. Our proposed model attains 94.09% accuracy, with a macro-averaged F1-score of 0.9442. Class-wise F1-scores are well-balanced, demonstrating robust discrimination performance across subtle visual differences. Slightly lower scores for ear-left (F1 = 0.8947) and ear-right (F1 = 0.9231) likely stem from the smaller support size in these categories.

To contextualize these results, we compare against a baseline classifier using DINOv2 features and a linear classifier initialized with pretrained SurgeNet weights. The updated DINOv2-based model achieves 91.40% accuracy and a macro-averaged F1-score of 0.9124. While the baseline performs competitively on common classes, it shows greater variability, particularly in throat (F1 = 0.8889) and ear-left (F1 = 0.8947). In contrast, our method maintains more stable precision-recall tradeoffs across all regions.

These findings underscore the efficacy of our proposed multimodal classification pipeline, which leverages BiomedCLIP for text-guided vision features and SurgeNet for spatial context extraction. The combination consistently yields high accuracy and balanced per-class performance, outperforming strong baselines in both coarse and fine-grained endoscopic classification tasks.

Table 4: Prototype classification report over fine-grained labels using BiomedCLIP embeddings. (accuracy: 55.91%).

Class	Precision	Recall	F1-score	Support
ear-left	0.5385	0.7778	0.6364	18
ear-right	0.6923	0.4500	0.5455	20
nose-left	0.5000	0.7021	0.5841	47
nose-right	0.4500	0.1957	0.2727	46
throat	0.5000	0.8000	0.6154	10
vc-closed	0.6087	0.6667	0.6364	21
vc-open	0.7727	0.7083	0.7391	24
Macro Avg	0.5803	0.6144	0.5756	186
Weighted Avg	0.5595	0.5591	0.5356	186

Table 5: Prototype classification report over coarse labels using BiomedCLIP embeddings (accuracy: 90.86%).

Class	Precision	Recall	F1-score	Support
Ear	0.9487	0.9737	0.9610	38
Nose	0.9884	0.9140	0.9497	93
Throat	0.4545	1.0000	0.6250	10
VC	0.9487	0.8222	0.8810	45
Macro Avg	0.8351	0.9275	0.8542	186
Weighted Avg	0.9420	0.9086	0.9179	186

6 Ablation Study

6.1 Retrieval

To assess the utility of our classification pipeline in enhancing semantic understanding, we conduct ablation studies on both image-to-image and text-to-image retrieval tasks using the BioMedCLIP model, as illustrated by Figure 4. Owing to its robust generalization capability across the four coarse classes, we use the pretrained BioMedCLIP embeddings directly, without additional finetuning.

For both modalities, we extract embeddings using the pretrained BioMedCLIP model. In the image-to-image retrieval setting, we first predict the class of the query image using our coarse classifier, and filter out images from the pool that do not share the predicted class label. This filtering simulates class-aware retrieval. We then compute cosine similarity between the embeddings of the query and the filtered candidates, reporting Recall@1, Recall@5, and Recall@10.

Table 6: Baseline DINOv2 classifier performance using pretrained SurgeNet weights (accuracy: 91.40%).

Class	Precision	Recall	F1-score	Support
Ear-left	0.8500	0.9444	0.8947	18
Ear-right	0.9474	0.9000	0.9231	20
Nose-left	0.9524	0.8511	0.8989	47
Nose-right	0.8800	0.9565	0.9167	46
Throat	1.0000	0.8000	0.8889	10
VC-closed	0.9048	0.9048	0.9048	21
VC-open	0.9231	1.0000	0.9600	24
Macro Avg	0.9225	0.9081	0.9124	186
Weighted Avg	0.9174	0.9140	0.9135	186

Table 7: Hybrid model classification performance on coarse labels (accuracy: 99.46%).

Class	Precision	Recall	F1-score	Support
Ear	0.9744	1.0000	0.9870	38
Nose	1.0000	0.9892	0.9946	93
Throat	1.0000	1.0000	1.0000	10
VC	1.0000	1.0000	1.0000	45
Macro Avg	0.9936	0.9973	0.9954	186
Weighted Avg	0.9948	0.9946	0.9946	186

Table 8: Hybrid model classification performance on fine-grained binary labels (accuracy: 94.09%).

Class	Precision	Recall	F1-score	Support
Ear-left	0.8500	0.9444	0.8947	18
Ear-right	0.9474	0.9000	0.9231	20
Nose-left	0.9556	0.9149	0.9348	47
Nose-right	0.9362	0.9565	0.9462	46
Throat	1.0000	1.0000	1.0000	10
VC-closed	0.9524	0.9524	0.9524	21
VC-open	0.9583	0.9583	0.9583	24
Macro Avg	0.9428	0.9467	0.9442	186
Weighted Avg	0.9421	0.9409	0.9410	186

Table 9: Comparison between the baseline DINOv2 classifier and prototype-based hybrid models.

Model	Accuracy	Macro F1	Weighted F1
BioMedCLIP proto.	55.91%	0.5756	0.5356
Baseline SurgeNet	91.40%	0.9124	0.9135
Hybrid Model	**94.09%**	**0.9442**	**0.9410**

For the text-to-image retrieval experiment, we make use of the text descriptions already included in the dataset, using them as natural-language queries. These queries reflect coarse class semantics (e.g., "an endoscopic image of the ear"). Image embeddings are compared to the query embeddings using cosine similarity in the joint text-image embedding space. Retrieval performance is again evaluated with Recall@1, Recall@5, and Recall@10.

Our results, summarized in Table 10, confirm that the combination of semantically aligned BioMedCLIP embeddings and our classification-informed filtering significantly enhances retrieval effectiveness, even without task-specific retraining.

Table 10: Retrieval performance on testing dataset.

Task	Recall@1	Recall@5	Recall@10
Image-to-Image	0.9140	0.9065	0.8731
Text-to-Image	0.9140	0.9065	0.9097

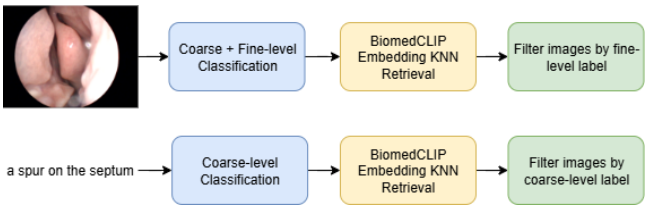

Figure 4: Multimodal retrieval pipeline.

6.2 Competition Results

The competition was organized into three main tracks: image classification (Track 1), image-to-image retrieval (Track 2), and text-to-image retrieval (Track 3). The private leaderboard results across these tracks reflect the final ranking and performance metrics of participating teams.

In **Track 1**, which focused on classifying ENT endoscopic images into coarse anatomical categories, Team WAS achieved the highest accuracy (95.82%), followed closely by Soft Mind_AIO (95.20%) and ZJU APRIL (94.74%). Our team, **HCMUS-Shndrit!**, achieved a solid performance with an accuracy of 91.64%, placing fifth among top-performing teams.

Table 11: Track 1 (Image Classification) Private Leaderboard Results.

Team Name	Accuracy	Precision	Recall	F1 Score
WAS	95.82	95.86	95.82	95.82
Soft Mind_AIO	95.20	95.24	95.20	95.20
ZJU APRIL	94.74	94.80	94.74	94.74
AIO-COFFEE	94.74	94.78	94.74	94.74
HCMUS-Shndrit!	**91.64**	**91.71**	**91.64**	**91.63**

In **Track 2**, the task was to retrieve the most similar image from a database given a query image. The leaderboard was topped by Soft Mind_AIO with a Recall@1 of 92.09% and MRR of 95.70. Our team, **HCMUS-Shndrit!**, placed third with a strong Recall@1 of 88.53% and MRR of 93.71.

Table 12: Track 2 (Image-to-Image Retrieval) Private Leaderboard Results.

Team Name	Recall@1	MRR
Soft Mind_AIO	92.09	95.70
STG001	88.79	94.32
HCMUS-Shndrit!	**88.53**	**93.71**
Re:zero Slavery	88.42	94.11
ELO	88.31	93.30

Track 3 required matching textual descriptions with the most relevant image from a gallery. The best performing team, SoloL, achieved 92.64% Recall@1 and an MRR of 95.81. ELO and entropy also attained high performance with Recall@1 values above 90%. Only the top five teams had their scores reported with four-digit precision. Our team, **HCMUS-Shndrit!**, placed sixth, with a Recall@1 of 84.00 and MRR of 92.00.

Table 13: Track 3 (Text-to-Image Retrieval) Private Leaderboard Results. Only the top 5 teams had their scores reported with four-digit precision.

Team Name	Recall@1	MRR
SoloL	92.64	95.81
ELO	90.77	95.12
entropy	90.67	94.95
STG001	89.79	94.81
H3N1	85.56	91.58
HCMUS-Shndrit!	**84.00**	**92.00**

These results highlight **HCMUS-Shndrit!** as one of the few teams to successfully compete across all three tracks, consistently ranking among the top performers. Despite limited resources compared to top-tier teams, our approach demonstrated strong generalization, particularly in retrieval-based tasks (Tracks 2 and 3), where semantic alignment from BiomedCLIP and domain-adapted fine-tuning contributed significantly to robust performance.

7 Conclusion

We proposed a robust multimodal framework for endoscopic image analysis, addressing anatomical site classification and cross-modal retrieval. On coarse-grained classification, the model achieved 94.09% accuracy and 0.9442 F1. A hybrid inference mechanism, switching between CLIP and DINOv2 based on confidence, further improved robustness on ambiguous or out-of-distribution samples. Our method ranked 5th, 3rd and 6th in the ENTRep challenge, demonstrating strong generalization across tasks. These results affirm the utility of integrating vision-language and vision-only models, establishing a solid foundation for multimodal learning in medical computer vision.

References

[1] Sharib Ali. 2022. Where do we stand in ai for endoscopic image analysis? deciphering gaps and future directions. *npj Digital Medicine*, 5, (Dec. 2022). doi:10.1038/s41746-022-00733-3.

[2] Jimmy Ba, Jamie Kiros, and Geoffrey Hinton. 2016. Layer normalization, (July 2016). doi:10.48550/arXiv.1607.06450.

[3] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. arXiv: 2401.08281 [cs.LG].

[4] Dan Hendrycks and Kevin Gimpel. 2023. Gaussian error linear units (gelus). (2023). https://arxiv.org/abs/1606.08415 arXiv: 1606.08415 [cs.LG].

[5] 2024. *Exploring the effect of dataset diversity in self-supervised learning for surgical computer vision. Data Engineering in Medical Imaging*. Springer Nature Switzerland, (Oct. 2024), 43–53. ISBN: 9783031737480. doi:10.1007/978-3-031-73748-0_5.

[6] Astitva Kamble, Vani Bandodkar, Saakshi Dharmadhikary, Veena Anand, Pradyut Sanki, Mei Wu, and Biswabandhu Jana. 2025. Enhanced multi-class classification of gastrointestinal endoscopic images with interpretable deep learning model. (Mar. 2025). doi:10.48550/arXiv.2503.00780.

[7] [SW] TorchVision maintainers and contributors, TorchVision: PyTorch's Computer Vision library 2016.

[8] Doniyrojon Mukhtorov, Rakhmonova Madinakhon, Shakhnoza Muksimova, and Young-Im Cho. 2023. Endoscopic image classification based on explainable deep learning. *Sensors*, 23, (Mar. 2023), 3176. doi:10.3390/s23063176.

[9] Maxime Oquab et al. 2023. Dinov2: learning robust visual features without supervision. (2023).

[10] Adam Paszke et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035.

[11] Cristiano Patrício, João C. Neves, and Luís F. Teixeira. 2023. Explainable deep learning methods in medical image classification: a survey. *ACM Comput. Surv.*, 56, 4, Article 85, (Oct. 2023), 41 pages. doi:10.1145/3625287.

[12] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. (2021). https://arxiv.org/abs/2103.00020 arXiv: 2103.00020 [cs.CV].

[13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15, 1, (Jan. 2014), 1929–1958.

[14] Sheng Zhang et al. 2024. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2, 1. doi:10.1056/AIoa2400640.