MMAG: Multimodal Learning for Mucus Anomaly Grading in Nasal Endoscopy via Semantic Attribute Prompting

Xinpan Yuan[1,3], Mingzhu Huang[1], Liujie Hua[2]*, Jianuo Ju[1], Xu Zhang[1]

[1]School of Computer and Artificial Intelligence, Hunan University of Technology, China,
[2]Hunan Police Academy, Changsha, China,
[3]Hunan Provincial Engineering Technology Research Center for Industrial Data Intelligence, Hunan University of Technology, Zhuzhou, China

Correspondence: liujiehuahnp@163.com

Abstract

Accurate grading of rhinitis severity in nasal endoscopy relies heavily on the characterization of key secretion types, notably clear nasal discharge (CND) and purulent nasal secretion (PUS). However, both exhibit ambiguous appearance and high structural variability, posing challenges to automated grading under weak supervision. To address this, we propose **M**ultimodal Learning for **M**ucus **A**nomaly **G**rading (MMAG), which integrates structured prompts with rank-aware vision-language modeling for joint detection and grading. Attribute prompts are constructed from clinical descriptors (e.g., secretion type, severity, location) and aligned with multi-level visual features via a dual-branch encoder. During inference, the model localizes mucus anomalies and maps the input image to severity-specific prompts (e.g., "moderate pus"), projecting them into a rank-aware feature space for progressive similarity scoring. Extensive evaluations on CND and PUS datasets show that our method achieves consistent gains over Baseline, improving AUC by 6.31% and 4.79%, and F1 score by 12.85% and 6.03%, respectively. This framework enables interpretable, annotation-efficient, and semantically grounded assessment of rhinitis severity based on mucus anomalies.

1 Introduction

In nasal endoscopy(Yuan et al., 2025c; Sedaghat et al., 2025; Acharia et al., 2025), the visual patterns of mucus secretions serve as critical indicators for assessing the severity of rhinitis (Gelardi et al., 2025). Among these, clear nasal discharge (CND)(Arslan and Çukurova, 2025) and purulent nasal secretion (PUS)(Usmonov and Jurayev, 2025) are two representative types frequently observed across different pathological stages. These secretions not only reflect underlying disease processes

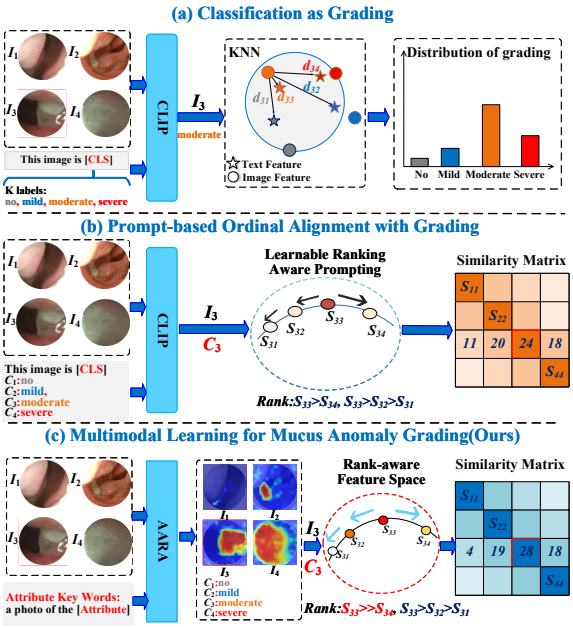


Figure 1: **Motivation of the proposed grading framework.** (a) Basic CLIP-based image-text classification lacks spatial and ordinal modeling. (b) Ranking-aware prompts model severity progression but ignore regional ambiguity. (c) Our method introduces attribute-guided localization and region-level alignment to enable interpretable, progression-consistent grading.

but also play a direct role in clinical decisions such as treatment planning and monitoring. As such, reliable detection and grading of mucus anomalies is a key component of intelligent decision-support systems in ENT diagnostics.

However, automated analysis of mucus in endoscopic imagery presents two major challenges. First, these secretions exhibit highly ambiguous visual properties—they(Gan et al., 2025) are often transparent, reflective, and conform to complex anatomical surfaces—making them difficult to segment or classify using conventional pixel-based or global methods. Second, the clinical progression of rhinitis is inherently continuous, evolving along semantic axes such as mucus type (*CND*

* Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21088–21098
November 4-9, 2025

→ *PUS*) and discharge volume (*no* → *severe*). Yet, existing methods typically formulate severity grading as a fixed-class classification task(Yang et al., 2022), disregarding the ordinal and progressive nature of the underlying pathology, thus limiting interpretability and generalization.

To address the challenges of grading mucus anomalies in nasal endoscopy, we analyze two representative vision-language modeling strategies, as illustrated in Figure 1: (a) *Classification as Grading*: This approach treats severity estimation as a global image-to-prompt matching task based on predefined category prompts. However, it lacks spatial awareness and cannot capture the ordinal nature of disease progression, often resulting in confusion between adjacent severity levels such as mild and moderate. (b) *Prompt-based Ordinal Alignment with Grading*: Building on (a), this strategy introduces learnable ranking-aware prompts (Yu et al., 2024) to represent ordered severity levels (e.g., *mild* → *moderate* → *severe*). While this improves ordinal modeling, it still relies on whole-image inference, making it less effective in localizing subtle anomalies in low-contrast or weak-structure scenarios. In contrast, we propose a new design (c) *Multimodal Learning for Mucus Anomaly Grading (MMAG)*. We propose a new framework that first performs anomaly localization using structured attribute prompts (e.g., *the secretion type is PUS*) and then aligns the localized features with severity-specific ranking prompts. This detection-to-grading paradigm enables spatially grounded and semantically consistent predictions, particularly under weak supervision.

We propose the first nasal endoscopic mucus grading framework that innovatively integrates attribute-guided localization with rank-aware modeling. Our key contributions include:

- Clinical attributes (e.g., secretion type, color, volume, anatomical site) are structured into semantic prompts to facilitate fine-grained vision-language alignment.
- Multi-level visual feature adapters and a dual-branch inference mechanism are employed to improve detection and localization robustness under weak supervision.
- An ordinal-consistent representation space is established, with a ranking-aware loss $\mathcal{L}_{\text{rank}}$ enforcing severity-aware alignment for interpretable, continuous grading.

We validate our framework on two clinically annotated datasets (CND and PUS), demonstrating significant performance gains in anomaly detection and severity grading tasks, along with strong semantic consistency and clinical applicability.

2 Related Works

2.1 Limitations of Medical Image Segmentation in Fluidic Regions

Medical image segmentation has long been a cornerstone for delineating anatomical or pathological structures. Classic CNN-based models (Long et al., 2015; Ronneberger et al., 2015), have evolved into multi-scale and Transformer-based variants (Cao et al., 2022; Zhong et al., 2024b,a). However, most of these methods are optimized for well-bounded, high-contrast regions.

In contrast, fluidic targets like mucus present low boundary saliency, optical ambiguity, and dynamic morphology. While prior works explored liquid regions in gastrointestinal imaging, they often rely on motion cues or dense labels, limiting their applicability to nasal endoscopy. Even advanced methods (Fan et al., 2020; Ma et al., 2024) struggle with vague boundaries and lack semantic grounding in low-contrast mucus scenarios.

2.2 Zero/Few-Shot Anomaly Detection with Vision-Language Models

Anomaly detection (AD) offers a more flexible and label-efficient alternative to dense segmentation, especially for fluidic patterns like nasal mucus. However, most AD methods (Hua et al., 2024, 2025; Yuan et al., 2025a) focus on rigid structures (e.g., brain (Baid et al., 2021), chest (Wang et al., 2017)), neglecting the ambiguity and dynamics of mucosal regions.

Unsupervised AD methods (Roth et al., 2022; Gudovskiy et al., 2022) rely only on normal data and perform poorly on diffuse or low-contrast anomalies. Few-shot AD methods (Ding et al., 2022; Yao et al., 2023) use limited abnormal labels and improve robustness via contrastive learning, but still face challenges with data imbalance and fluidic variability. Zero-shot AD leverages vision-language models (VLMs) like AnomalyCLIP (Zhou et al., 2023) and WinCLIP (Jeong et al., 2023) distinguish anomalies using handcrafted or learned prompts.

Despite progress in natural image tasks, VLMs struggle with medical images due to limited paired

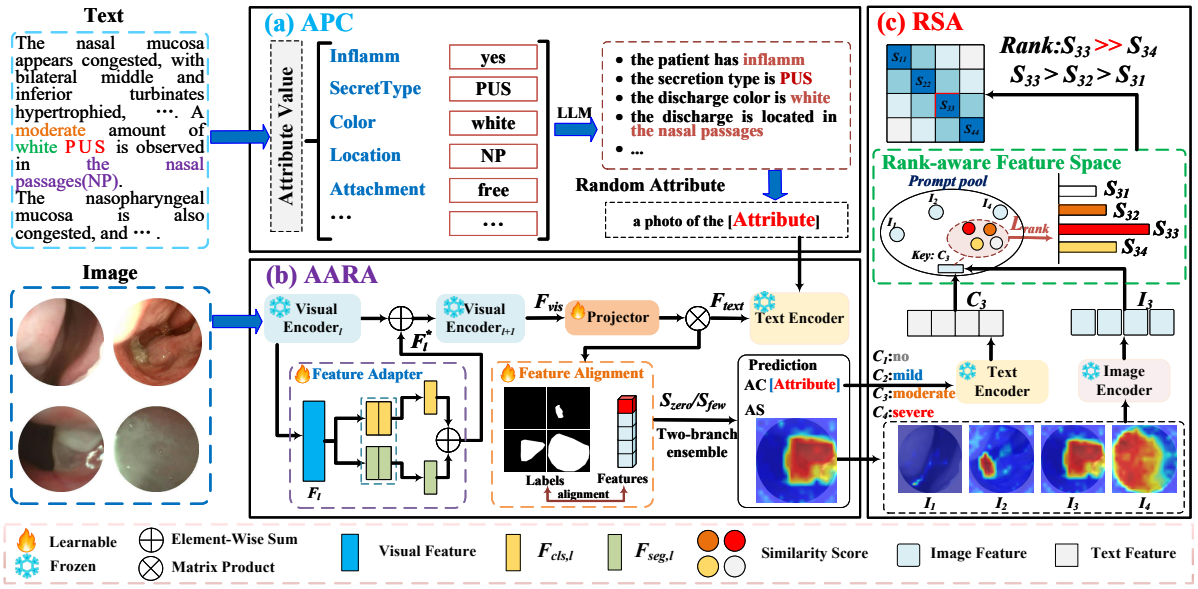


Figure 2: **Framework overview.** (a) **Attribute Prompt Construction(APC):** Structured clinical attributes are converted into textual prompts for visual-semantic alignment. (b) **Anomaly-Aware Region Alignment(AARA):** Visual features are aligned with prompts via adapters and a projector, supporting zero-/few-shot prediction. (c) **Rank-aware Severity Assessment(RSA):** Severity levels are modeled with prompt ranking, guided by rank loss $\mathcal{L}_{\text{rank}}$ for ordinal consistency.

data, noisy labels, and weak spatial alignment. Existing static (Bozinis et al., 2021) and dynamic (Wu et al., 2025) prompt strategies focus on rigid object reasoning, limiting region-aware detection in deformable, low-contrast mucus areas. This motivates a spatially grounded, attribute-aware VLM framework for zero/few-shot medical AD.

2.3 Ordinal Modeling for Medical Image Grading

Medical image grading often involves ordinal labels indicating disease progression (e.g., *mild, moderate, severe in diabetic retinopathy*)(Yang et al., 2022; Che et al., 2023). Traditional multiclass methods use one-hot encoding and cross-entropy loss(Wang et al., 2023), ignoring ordinal structure and treating all misclassifications equally.

To address this, CLIP-DR(Yu et al., 2024) reformulates DR grading as an image-text matching task with ranking-aware prompts and KL-divergence to align image features with ordered text semantics. Its Similarity Matrix Smoothing (SMS) mitigates long-tail imbalance. Other methods explore ordinal regression or domain generalization(Wang et al., 2025), but typically depend on clean fundus images, limiting generalization to ambiguous endoscopic scenarios. Spatial structure modeling has also been explored. For instance, OF-AR(Yuan et al., 2025b) leverages an inverse area relationship between lesion and background regions in monocular endoscopy, using segmentation and contrastive learning to improve grading robustness.

Inspired by prior work, we extend ordinal modeling to fluidic nasal endoscopy by proposing a prompt-guided framework that aligns localized anomalies with severity-ordered prompts, enabling fine-grained and progression-aware predictions in complex scenes.

3 Methods

3.1 Overview

As illustrated in Figure 2, the framework performs visual-language reasoning for anomaly localization and severity grading in nasal endoscopy, comprising three key components: **Attribute Prompt Construction(APC).** Clinical attributes (e.g., *secretion type, color, location, inflammation*) are parsed into natural language prompts (e.g., *the secretion type is PUS*) to guide cross-modal alignment. **Anomaly-Aware Region Alignment(AARA).** Images are encoded by a frozen visual backbone with lightweight adapters and a projector. Cross-modal similarity aligns image features with prompts. A dual-branch design combines zero-shot and few-shot inference for classification and localization. **Rank-aware Severity Assessment(RSA).** Ordinal prompts represent severity levels, and the ranking-

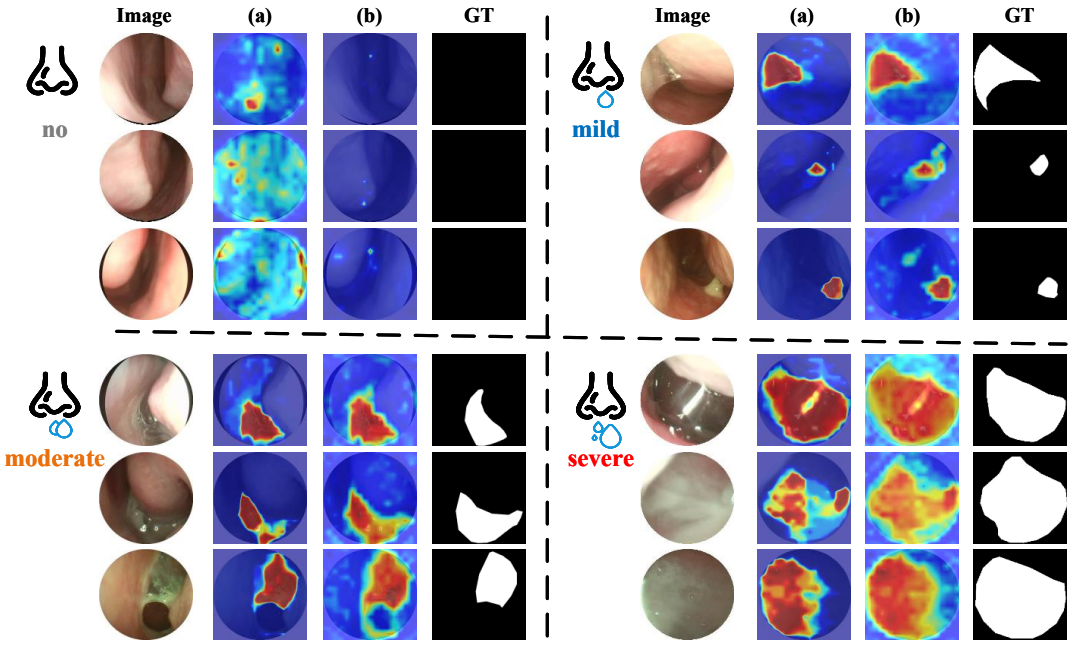


Figure 3: **Qualitative comparison of detection-guided grading.** (a) Zero-shot baseline detection. (b) Two-Branch Ensemble prediction. Compared to the baseline, our method produces more accurate and concentrated anomaly heatmaps across severity levels on both CND and PUS datasets, with improved robustness under challenging conditions such as blurred boundaries and specular reflections.

aware loss $\mathcal{L}_{\text{rank}}$ enforces progressive alignment for fine-grained, interpretable grading.

3.2 APC

We construct lightweight prompts from structured attributes to align image features with clinical semantics. Given an attribute space $\mathbf{A}$(e.g., *secretion type, color, location*), attribute values are converted into short sentences (e.g., *the secretion type is PUS*) and embedded into the template *a photo of the [Attribute]*. This enables fine-grained alignment, supervised by two objectives: Image-Attribute Contrastive (IAC) and Matching (IAM).

Attribute	Representation in words
Inflammation	"yes", "no"
Secretion Type	"PUS", "CND", "none"
Color	"white", "yellow", "clear",...
Volume	"no", "mild", "moderate", "severe"
Attachment	"free", "adherent to mucosa",...
Location	"nasal passages", "nasopharynx",...

Table 1: Core Attributes for Endoscopic Mucus Analysis

IAC. We construct image-attribute prompt pairs for cross-modal alignment: For each image I, descriptive sentences (e.g., *secretion is PUS*) are embedded into the template *a photo of the [Attribute]* to form positive prompt T_a, while negative prompt $\bar{T}_a$ is generated by replacing key terms (e.g., *PUS* $\rightarrow$ *CND*). The cosine similarities are defined as $s^+ = s(F_I, F_{T_a})$ and $s^- = s(F_I, F_{\bar{T}_a})$, where F_I denotes the visual `[CLS]` feature and $F_{T_a}/F_{\bar{T}_a}$ represent the text features of corresponding prompts. The temperature parameter τ is introduced to adjust the probability distribution sharpness:

$$S_{\text{i2a}}(I) = \frac{\exp(s^+/\tau)}{\exp(s^+/\tau) + \exp(s^-/\tau)} \quad (1)$$

In each training batch, we sample all matched pairs to form the set B_a. The loss function is:

$$\mathcal{L}_{iac} = -\frac{1}{|B_a|} \sum_{(I,T_a)\in B_a} \log S_{i2a}(I) \quad (2)$$

$\mathcal{L}_{iac}$ encourages the model to align with correct prompts and suppress mismatches, enhancing fine-grained attribute discrimination.

IAM. IAM models the image-attribute prompt relationship. For each batch, we sample $5|B|$ pairs (I, T_a), including matched (*from image attributes*) and mismatched pairs. Each is encoded to extract the c^{cls} representation from `[CLS]`, with matching probability computed as: $p^{match}(I, T_a) =$

Sigmoid $\left(\text{MLP}(c^{cls})\right)$. The cross-entropy loss is defined as:

$$\mathcal{L}_{iam} = -\frac{1}{|B_a|}\sum_{(I,T_a)\in B_a}\Big(y_a^{match}\log p^{match}(I,T_a) + (1-y_a^{match})\log\Big(1-p^{match}(I,T_a)\Big)\Big) \quad (3)$$

where $y_a^{match} = 1$ indicates a matching pair (i.e., the prompt is consistent with the image's attributes); otherwise, it is 0.

3.3 AARA

To enhance localization in low-contrast, complex regions, we propose the AARA module. It uses the Multi-level Visual Feature Adapter (MVFA) (Huang et al., 2024) to align encoder features across stages, which are then fused with attribute prompts via a cross-modal projector. The aligned features support dual-branch inference for attribute classification (AC) and anomaly scoring (AS).

MVFA. The image is encoded by a pre-trained CLIP visual encoder to extract multi-level features (S1–S3), where each stage outputs $F_l \in \mathbb{R}^{G\times d}$, with $l \in \{1,2,3\}$, G denoting spatial grids, and d the feature dimension. To improve adaptability and prevent overfitting, lightweight adapters $A_l(\cdot)$ are applied at each stage as learnable linear transformations. The adaptation is defined as:

$$F_l^* = \gamma A_l(F_l)^T + (1-\gamma)F_l \quad (4)$$

here, $A_l(\cdot)$ is the adapter for the l-th layer, and γ is a residual weight (default 0.1) controlling adaptation strength. This enables multi-level feature adjustment without altering the CLIP encoder.

Feature Alignment. For feature alignment, we optimize the alignment between image features F_l^* and text prompts F_{text} based on their similarity. We first compute the similarity score $S_{align}(F_l^*, F_{text})$ between the image and text, using cosine similarity as the metric:

$$S_{align}(F_l^*, F_{text}) = \frac{F_l^* \cdot F_{text}}{\|F_l^*\|\|F_{text}\|} \quad (5)$$

Then, to enhance image-text alignment, we maximize similarity scores and define an alignment loss $\mathcal{L}_{\text{align}}(F_l^*, F_{\text{text}})$ for each layer l to measure visual-textual consistency:

$$\mathcal{L}_{align} = -\sum_{l=1}^{3} S_{align}(F_l^*, F_{text}) \quad (6)$$

$\mathcal{L}_{\text{align}}$ encourages multi-level semantic alignment between image and text, improving the accuracy of anomaly localization.

Two-Branch Ensemble. We employ a dual-branch combining zero-shot and few-shot paths. The zero-shot branch requires no labeled data, predicting anomalies via image-text similarity using pretrained features and multi-level adapters.

For each layer's classification feature $F_{cls,l}$ and text prompt feature F_{text}, we compute their similarity and apply softmax to obtain layer-wise anomaly classification scores:

$$C_{zero} = \frac{1}{4}\sum_{l=1}^{4}\max_{G}\text{softmax}(F_{cls,l}\cdot F_{text}) \quad (7)$$

where softmax measures image-text similarity across layers to identify potential anomaly regions.

For each segmentation feature $F_{seg,l}$, we compute its similarity with the text feature, apply softmax for anomaly scoring, and use bilinear interpolation (BI) to restore the original resolution:

$$S_{zero} = \frac{1}{4}\sum_{l=1}^{4}\text{BI}\left(\text{softmax}(F_{seg,l}\cdot F_{text})\right) \quad (8)$$

where BI$(\cdot)$ denotes bilinear interpolation used to upsample the anomaly map to image resolution.

In the Few-Shot branch, Constructs multi-level memory bank $\mathcal{G}$sing limited labeled normal images, enabling anomaly classification/localization via layer-wise feature distance during inference:

$$C_{few} = \frac{1}{4}\sum_{l=1}^{4}\max_{G}\left(\min_{m\in\mathcal{G}}\text{Dist}(F_{\text{cls},l}, m)\right) \quad (9)$$

where Dist$(\cdot)$ denotes cosine distance between the test feature $F_{\text{cls},l}$ and memory feature m.

For each segmentation feature $F_{seg,l}$, we compute the minimum distance to memory features to obtain the anomaly score per layer:

$$S_{\text{few}} = \frac{1}{4}\sum_{l=1}^{4}\text{BI}\left(\min_{m\in\mathcal{G}}\text{Dist}(F_{\text{seg},l}, m)\right) \quad (10)$$

Dist$(\cdot)$ computes the cosine distance between test and memory segmentation features, and BI$(\cdot)$ upsamples the scores to the original resolution.

During training, the Zero-Shot and Few-Shot branches independently produce anomaly classification and region-level predictions, integrated via weighted summation:

$$C_{\text{pred}} = \beta_1 C_{\text{zero}} + \beta_2 C_{\text{few}}, S_{\text{pred}} = \beta_1 S_{\text{zero}} + \beta_2 S_{\text{few}} \quad (11)$$

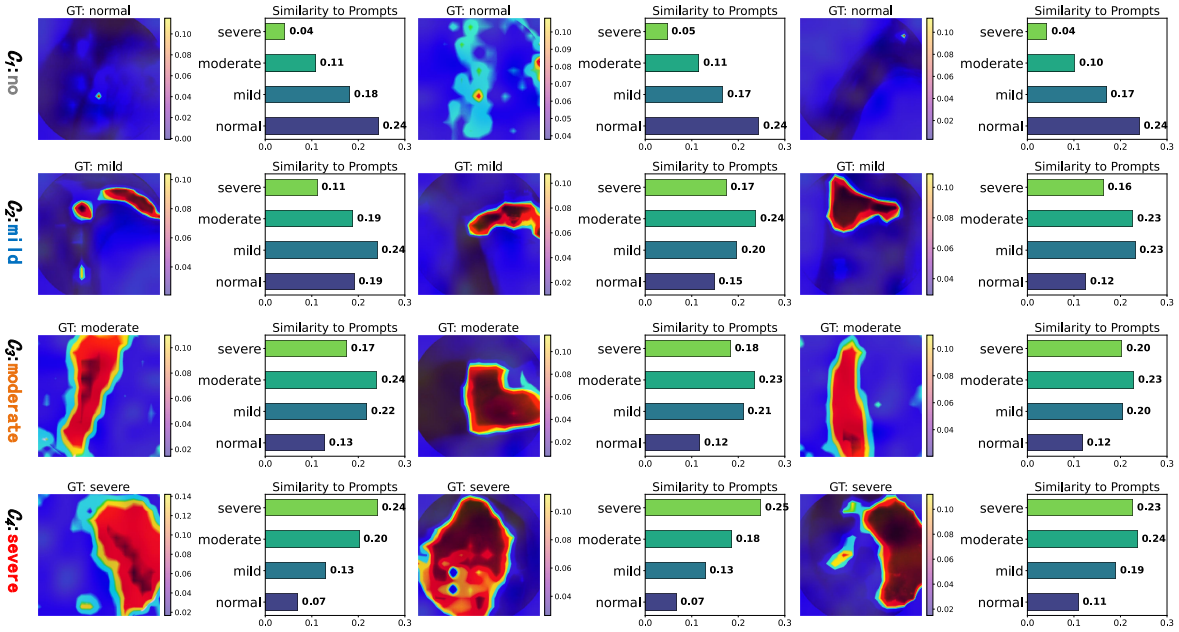


Figure 4: Attribute-prompt grading analysis: Ground-truth prompts show peak similarity with progressively decreasing non-matching scores, demonstrating precise pathology grading (normal/mild/moderate/severe).

where β_1 and β_2 are weighting coefficients (default: $\beta_1 = \beta_2 = 0.5$) controlling the contribution of each branch.

3.4 RSA

Rank-Aware Feature Space. This space aligns anomaly localization images with severity-ordered prompts. Let $\tilde{s}_{i,j}$ be the similarity between image i and class-j prompt. To enforce correct severity ordering, we introduce a Rank Loss encouraging $\tilde{s}_{i,j} > \tilde{s}_{i,j+1} > \cdots > \tilde{s}_{i,K}$ and $\tilde{s}_{i,j} > \tilde{s}_{i,j-1} > \cdots > \tilde{s}_{i,1}$, guiding the model to rank severe anomalies higher than mild ones.

Right Loss. Right Loss encourages the model to correctly order the predicted similarity scores across adjacent classes. Specifically, it compares the similarity score of the current category $\tilde{s}_{i,j'}$ with that of its rightward neighbor $\tilde{s}_{i,j'+1}$, encouraging $\tilde{s}_{i,j'} > \tilde{s}_{i,j'+1}$. The loss is defined as:

$$\mathcal{L}^i_{\text{right}} = -\sum_{j'=j}^{k-1} \log \frac{\exp(\tilde{s}_{i,j'}/\tau)}{\exp(\tilde{s}_{i,j'}/\tau) + \exp(\tilde{s}_{i,j'+1}/\tau)} \tag{12}$$

where $\tilde{s}_{i,j'}$ is the predicted similarity between image i and the prompt for class j', τ is a temperature parameter (set to 1 in our experiments), and k is the total number of severity classes.

Left Loss. Left Loss similarly compares the similarity score of the current category $\tilde{s}_{i,j'}$ with its leftward neighbor $\tilde{s}_{i,j'-1}$, enforcing $\tilde{s}_{i,j'} > \tilde{s}_{i,j'-1}$ to preserve the ordinal relationship. The loss is defined as:

$$\mathcal{L}^i_{\text{left}} = -\sum_{j'=2}^{j} \log \frac{\exp(\tilde{s}_{i,j'}/\tau)}{\exp(\tilde{s}_{i,j'}/\tau) + \exp(\tilde{s}_{i,j'-1}/\tau)} \tag{13}$$

where $\tilde{s}_{i,j'}$ is the similarity between image i and the class j' prompt, τ is the temperature parameter, and j is the index of the current ground-truth class.

The rank loss function. The overall rank loss is averaged over all M training samples:

$$\mathcal{L}_{\text{rank}} = \frac{1}{M} \sum_{i=1}^{M} \left(\mathcal{L}^i_{\text{right}} + \mathcal{L}^i_{\text{left}} \right) \tag{14}$$

4 Experiments

4.1 Experimental Setups

Datasets. This study uses a curated nasal endoscopy dataset collected from a tertiary Grade-A hospital, comprising 1,000 high-definition images (1920×1080) selected from 2,880 rhinitis cases. Each image shows clear nasal discharge (CND), purulent nasal secretion (PUS), or normal structures, anonymized and ethically approved, and independently annotated by two ENT specialists. Blurred or ambiguous frames were excluded. The dataset supports grading and detection algorithm development and is split into training, validation, and test sets in a 7:2:1 ratio.

Dataset	Method	Metrics 1		Metrics 2						
		AC	**AS**	**ACC**	**F1**	**AUC**	**no**	**mild**	**moderate**	**severe**
CND	GDRNet (Che et al., 2023)	-	-	48.92	43.27	74.75	98.08	55.71	77.56	74.28
	GDRNet+Ours	97.49	88.56	59.57	51.19	79.02	99.42	75.95	80.81	72.85
	CLIP (Rame et al., 2022)	-	-	40.23	31.52	70.54	83.61	64.88	67.45	64.10
	CLIP+Ours	96.27	84.27	50.37	53.66	75.89	96.11	61.30	65.37	73.86
	OrdinalCLIP (Niu et al., 2016)	-	-	46.33	41.44	72.45	94.43	55.48	64.72	67.61
	OrdinalCLIP+Ours	96.70	85.03	61.71	53.27	79.11	99.27	74.28	82.97	73.21
	CLIP-DR (Yu et al., 2024)	-	-	65.96	56.93	81.83	98.35	76.19	**85.40**	76.78
	CLIP-DR+Ours	**99.97**	**92.34**	**71.97**	**69.78**	**92.99**	**99.78**	**94.70**	82.47	**94.50**
PUS	GDRNet (Che et al., 2023)	-	-	72.36	68.45	90.96	99.47	85.95	86.75	97.50
	GDRNet+Ours	95.55	94.78	80.46	83.03	96.16	99.51	93.25	87.78	97.69
	CLIP (Rame et al., 2022)	-	-	68.06	53.59	88.61	98.85	90.47	86.75	86.75
	CLIP+Ours	92.92	90.64	75.58	78.44	94.90	99.89	93.21	87.06	96.78
	OrdinalCLIP (Niu et al., 2016)	-	-	70.21	63.17	90.69	99.04	85.47	**89.18**	96.07
	OrdinalCLIP+Ours	94.32	93.95	76.80	79.57	95.25	99.77	93.05	85.91	98.67
	CLIP-DR (Yu et al., 2024)	-	-	74.46	69.90	91.71	99.45	86.90	88.10	96.78
	CLIP-DR+Ours	**99.13**	**96.45**	**84.13**	**85.93**	**96.24**	**99.98**	**93.78**	88.79	**99.24**

Table 2: Performance comparison of different architectures and methods on CND and PUS datasets

Dataset	**Type**	**AS**	**AC**	**ACC**	**F1**	**AUC**
Hyper-Kvasir	Polyp	97.96	99.98	81.15	78.04	94.75
	Mucus_CND	91.23	99.98	76.90	71.16	87.09
	Mucus_PUS	91.02	99.31	60.54	62.29	77.45
CVC_ColonDB	Polyp	97.77	99.95	79.71	76.68	94.47
CVC_ClinicDB	Polyp	97.80	99.96	78.26	76.91	95.07

Table 3: Transfer performance of our method on Hyper-Kvasir, ColonDB, and ClinicDB, including both polyp and mucus subsets.

CVC_ClinicDB. CVC_ClinicDB (Bernal et al., 2015) contains 612 colonoscopy frames with pixel-wise polyp annotations. It is widely used for benchmarking segmentation methods, with challenges such as blurred boundaries and low-contrast flat polyps. We use it both for polyp segmentation evaluation and for testing generalization from mucus-trained models. We split the dataset into training, validation, and test sets in a 7:2:1 ratio.

CVC_ColonDB. CVC_ColonDB (Tajbakhsh et al., 2015) includes 380 frames from different procedures, featuring greater variation in polyp size, shape, and texture than ClinicDB. It is used to assess model robustness under diverse appearances and as a transfer target in our mucus-to-polyp generalization experiments. We split the dataset into training, validation, and test sets in a 7:2:1 ratio.

Hyper-Kvasir. Hyper-Kvasir (Borgli et al., 2020) is the largest publicly available gastrointestinal endoscopy dataset, containing more than 110,000 images and video frames. In our study, we use its mucus-related subset as an external evaluation domain to examine the generalization ability of models trained on CND and PUS when transferred to gastrointestinal mucus analysis tasks. We split the subset into training, validation, and test sets in a 7:2:1 ratio.

Evaluation Metrics. We adopt a dual evaluation protocol comprising detection metrics (AC-AUC, AS-AUC) and grading metrics (ACC, F1, AUC). Detection metrics assess anomaly identification and localization quality, while grading metrics evaluate classification accuracy across severity levels (*normal/mild/moderate/severe*). All metrics are reported on a 0–100% scale, with higher scores indicating better performance.

Implementation Details. This study employs the CLIP-ViT/L14 architecture, with all experiments conducted on an NVIDIA Tesla V100 GPU (16GB memory) at 240×240 resolution. The AdamW optimizer (learning rate 0.0001) was used for 100-epoch training (batch size=16) enhanced by a lightweight multi-view feature adapter. For endoscopic mucus characterization, we established an annotation system (Table 1) encompassing six clinical attributes: inflammatory status, secretion type, color characteristics, volume grading, attachment properties, and anatomical location, with all annotations strictly complying with endoscopic standards to ensure clinical relevance and visual discriminability.

Methods	CND		PUS	
	F1	AUC	F1	AUC
Baseline	20.52	45.25	23.18	49.21
+Attribute	25.05	58.99	25.53	60.05
+Attr.+Adapter	34.76	64.10	35.23	64.79
+Attr.+Adap.+$\mathcal{L}_{rank}$	**41.01**	**67.83**	**42.40**	**71.45**

Table 4: Ablation study on Model Components under Zero-shot Setting

Setting	CND			PUS		
	AS	F1	AUC	AS	F1	AUC
0-shot	60.76	43.56	74.79	68.31	45.54	75.45
1-shot	79.29	50.37	75.89	90.64	58.92	88.61
2-shot	88.59	51.19	79.02	91.38	67.63	90.69
4-shot	90.06	53.27	79.11	94.73	69.82	90.96
8-shot	91.28	56.92	81.74	95.52	73.70	91.71
16-shot	**93.47**	**77.12**	**94.43**	**97.98**	**78.54**	**96.76**

Table 5: Ablation study on few-shot learning performance with different sample sizes (Average AUC(%), best F1(%) and Accuracy Score(AS%)).

4.2 Comparison with Grading Methods

We present a multimodal framework for detecting and grading mucus anomalies in nasal endoscopy, focusing on two discharge types: clear nasal discharge (CND) and purulent nasal secretion (PUS). As shown in Table 2, our method consistently outperforms several baselines. For CND, ACC improves from 65.96% to 71.97%, and F1 from 56.93% to 69.78%; for PUS, ACC and F1 reach 84.13% and 85.93%, demonstrating stronger representation and classification robustness.

Compared with existing approaches, our framework shows notable advantages. Against GDR-Net (Che et al., 2023), it remains more robust under boundary ambiguity. Unlike CLIP (Radford et al., 2021), we introduce perception modules tailored to endoscopic imaging. Relative to OrdinalCLIP (Niu et al., 2016) and CLIP-DR (Yu et al., 2024), our method achieves higher specificity and produces grading results that are more consistent with clinical practice.

Although our main experiments are based on two nasal endoscopy datasets (CND and PUS), this choice is justified and representative. Nasal endoscopy is more challenging than other modalities (e.g., *gastroscopy, colonoscopy, cystoscopy*), which mostly visualize homogeneous wall structures. The nasal cavity involves complex anatomy (*septum, turbinates, sinuses*) and diverse mucus presentations, leading to substantial visual heterogeneity. Moreover, clear and purulent discharges are clinically informative early indicators of rhinitis, yet their low contrast and weak boundaries make them difficult to model. Thus, CND and PUS serve as high-difficulty benchmarks for robustness and transferability. To further validate the generalization ability of our method beyond nasal endoscopy, we conduct transfer experiments on colonoscopy datasets, including both polyp and mucus subsets. As shown in Table 3, the results on Hyper-Kvasir, ColonDB, and ClinicDB confirm that our framework maintains strong performance when adapting from nasal to colonoscopy domains.

Overall, our detection-then-grading strategy aligns with clinical workflows and enhances both performance and interpretability for mucus anomaly assessment.

4.3 Ablation Studies

To assess the contribution of each module, we conducted ablation studies under both zero-shot and multi-shot settings, see Table 4 and Table 5.

Zero-shot setting. Introducing attribute prompts improves F1 by 4.53% (CND) and 2.35% (PUS), and AUC by 13.74% and 10.84%. Adding the adapter yields further F1 gains of 9.71% and 9.70%, and AUC gains of 5.11% and 4.74%. Finally, applying $\mathcal{L}_{\text{rank}}$ boosts F1 by 6.25% and 7.17%, and AUC by 3.73% and 6.66%, confirming its effectiveness for severity modeling.

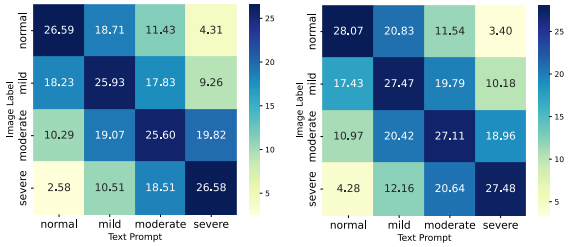
Multi-shot setting. We evaluate the model under 0-shot, 1-shot, 2-shot, 4-shot, 8-shot, and 16-shot conditions. Both CND and PUS tasks show steady improvements across F1, AUC, and AS metrics. For example, on PUS, F1 increases from 45.54 (0-shot) to 78.54 (16-shot), AUC from 75.45 to 96.76, and AS from 68.31 to 97.98. For CND, F1 rises from 43.56 to 77.12, AUC from 74.79 to 94.43, and AS from 60.76 to 93.47. Gains begin to saturate beyond 8-shot, showing that the model achieves high accuracy with limited supervision.

As shown in Table 6, the dual-branch configuration consistently outperforms the individual zero-only and few-only branches in AS and F1, highlighting their complementary strengths. Visual results confirm that the dual-branch design delivers clearer segmentation under glare or blurred boundaries, highlighting its robustness.

These results validate the effectiveness of each component and the robustness of our framework

Setting	CND			PUS		
	AS	F1	AUC	AS	F1	AUC
Zero-only	60.76	43.56	74.79	68.31	45.54	75.45
Few-only	89.24	42.17	82.01	93.75	47.69	81.97
Dual-branch	**90.06**	**53.27**	**79.11**	**94.73**	**69.82**	**90.96**

Table 6: Ablation study under the 4-shot setting on CND and PUS datasets, evaluating zero-only, few-only, and dual-branch settings to validate the architecture.



(a) Baseline **(b) Ours**

Figure 5: Comparison of two grading strategies on PUS dataset: (a) direct grading baseline, (b) detection-guided grading with $\mathcal{L}_{\text{rank}}$ constraint.

across various supervision regimes.

4.4 Visualization Analysis

As shown in Figure 3, visual results on the CND and PUS datasets highlight the effectiveness of our detection-guided grading framework. Compared to the zero-shot baseline in (a), the Two-Branch Ensemble in (b) yields more accurate and concentrated anomaly maps, even under challenging conditions like blurred edges and specular reflections. As severity increases, the heatmaps become sharper and better aligned with semantic cues, reflecting strong grading awareness and interpretability.

Figure 4 further demonstrates that the model effectively aligns attribute prompts with localized mucus regions. From mild to severe cases, activations intensify, and category-wise similarity becomes more distinct, enabling reliable severity quantification.

In Figure 5, the similarity matrix of baseline grading (a) shows off-diagonal confusion between adjacent categories, whereas our rank-aware model (b) achieves diagonal dominance, indicating clearer inter-class boundaries. This confirms that integrating anomaly localization with $\mathcal{L}_{\text{rank}}$ improves both semantic alignment and grading consistency.

Overall, our framework supports precise localization and interpretable severity estimation, offering clinically valuable insights for automated mucus anomaly assessment.

5 Conclusion

This study proposes the MMAG framework for objective inflammation severity assessment through nasal mucus characteristics. The framework consists of three core modules: (1) attribute prompt construction for clinical feature representation, (2) anomaly-aware region alignment for precise lesion localization, and (3) rank-aware assessment modeling severity ordinal relationships. Experiments demonstrate MMAG's accurate grading capability for both CND and PUS discharges. Future work will focus on borderline mucus feature extraction and fine-grained recognition algorithms for secretions in hidden nasal areas to enhance clinical discrimination.

Limitations

Although our framework demonstrates promising performance in grading mucus anomalies under varied endoscopic conditions, it still faces challenges in extremely low-illumination regions. Specifically, in dark nasal cavities where the endoscope light is insufficient or obstructed, transparent or low-contrast mucus may be overlooked or misclassified as absent. This can lead to underestimation of inflammation severity, which may affect clinical interpretation in real-world scenarios. Future work may explore adaptive enhancement or uncertainty modeling techniques to improve robustness in underexposed areas.

Ethical Considerations. This study was approved by the institutional ethics committee of a certified medical center (Approval No. LLYPJ2025091-01).Written consent for clinical photographs was obtained from all participants or their guardians.

Acknowledgments

This work was supported by the National Natural Science Foundation of Hunan Province (Grant no. 2025JJ70028, 2025JJ81178, 2024JJ9550), the Scientific Research Project of Education Department of Hunan Province (Grant no. 24A0401), and the Hunan Provincial Graduate Research and Innovation Project (2025).

References

Kuldeep Acharia, Pooja Thakur, Puspen Dasgupta, Sabyasachi Gon, Dharitri Mukherjee, Anwesha Dandapath, Apu Dey, and Abhishek Patra. 2025. The

role of diagnostic nasal endoscopy and computed tomography scan in chronic rhinosinusitis in adults: A study of clinical correlation. *Indian Journal of Otolaryngology and Head & Neck Surgery*, pages 1–6.

İlker Burak Arslan and İbrahim Çukurova. 2025. Rhinorrhea: Pathogenesis, diagnosis, and treatment. In *Pediatric Airway Diseases*, pages 327–336. Springer.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, and 1 others. 2021. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. 2015. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111.

Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, and 1 others. 2020. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283.

Theodoros Bozinis, Nikolaos Passalis, and Anastasios Tefas. 2021. Improving visual question answering using active perception on static images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 879–884. IEEE.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer.

Haoxuan Che, Yuhan Cheng, Haibo Jin, and Hao Chen. 2023. Towards generalizable diabetic retinopathy grading in unseen domains. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 430–440. Springer.

Choubo Ding, Guansong Pang, and Chunhua Shen. 2022. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7388–7398.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. 2020. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer.

Hong-Seng Gan, Muhammad Hanif Ramlee, Zimu Wang, and Akinobu Shimizu. 2025. A review on medical image segmentation: Datasets, technical models, challenges and solutions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(1):e1574.

Matteo Gelardi, Rossana Giancaspro, Elisa Boni, Mario Di Gioacchino, Giulia Cintoli, Michele Cassano, and Maria Teresa Ventura. 2025. Rhinitis in the geriatric population: Epidemiological and cytological aspects. *Geriatrics*, 10(2):50.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. 2022. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 98–107.

Liujie Hua, Qianqian Qi, and Jun Long. 2024. P2 random walk: self-supervised anomaly detection with pixel-point random walk. *Complex & Intelligent Systems*, 10(2):2541–2555.

Liujie Hua, Xiu Su, Yueyi Luo, Shan You, and Jun Long. 2025. Hieclip: Hierarchical clip with explicit alignment for zero-shot anomaly detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. 2024. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11375–11385.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications*, 15(1):654.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2016. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. 2022. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328.

Ahmad R Sedaghat, Ryan A Cotter, Isam Alobid, Saad Alsaleh, Wilma Terezinha Anselmo-Lima, Manuel Bernal-Sprekelsen, Rakesh K Chandra, Jannis Constantinidis, Wytske J Fokkens, Christine Franzese, and 1 others. 2025. Nasal endoscopy score thresholds to trigger consideration of chronic rhinosinusitis treatment escalation and implications for disease control. *Rhinology*, 63(1):54–62.

Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. 2015. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644.

Sanjar Usmonov and Khondamir Jurayev. 2025. Acute purulent sinusitis: clinical course, diagnosis and treatment methods. *International Journal of Medical Sciences And Clinical Research*, 5(01):69–71.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and 1 others. 2023. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

Xinkun Wang, Yifang Wang, Senwei Liang, Feilong Tang, Chengzhi Liu, Ming Hu, Chao Hu, Junjun He, Zongyuan Ge, and Imran Razzak. 2025. Robust multimodal learning for ophthalmic disease grading via disentangled representation. *arXiv preprint arXiv:2503.05319*.

Xian Wu, Hong Xiao, and Lin Ma. 2025. The application of computational fluid dynamics in hepatic portal vein haemodynamics research: a narrative review. *Quantitative Imaging in Medicine and Surgery*, 15(3):2605.

Yuzhe Yang, Hao Wang, and Dina Katabi. 2022. On multi-domain long-tailed recognition, imbalanced domain generalization and beyond. In *European Conference on Computer Vision*, pages 57–75. Springer.

Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. 2023. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24490–24499.

Qinkai Yu, Jianyang Xie, Anh Nguyen, He Zhao, Jiong Zhang, Huazhu Fu, Yitian Zhao, Yalin Zheng, and Yanda Meng. 2024. Clip-dr: Textual knowledge-guided diabetic retinopathy grading with ranking-aware prompting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 667–677. Springer.

Xinpan Yuan, Mingzhu Huang, Liujie Hua, Changhong Zhang, Jianuo Ju, Shaomin Xie, and Wenguang Gan. 2025a. Men-ad: Multimodal zero-shot detection of mucus anomalies in endoscopy. In *International Conference on Intelligent Computing*, pages 85–97. Springer.

Xinpan Yuan, Siming Jin, Liujie Hua, Guihu Zhao, Changhong Zhang, and Yuan Guo. 2025b. Of-ar relation aware representation learning for lesion image segmentation and grading. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xinpan Yuan, Junhua Kuang, Liujie Hua, Guihu Zhao, Changhong Zhang, and Jiabao Li. 2025c. A novel single continuous shot multiple lesions endoscopy report generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yunshan Zhong, Jiawei Hu, You Huang, Yuxin Zhang, and Rongrong Ji. 2024a. Erq: Error reduction for post-training quantization of vision transformers. In *International Conference on Machine Learning (ICML)*.

Yunshan Zhong, You Huang, Jiawei Hu, and Rongrong Ji Yuxin Zhang. 2024b. Towards accurate post-training quantization of vision transformers via error reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), CCF-A*.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. 2023. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*.