EndoVLM: An Endoscopy Vision-Language Pre-training Model via Anatomy-Guided Sparsity and Progressive Alignment

Zhenyu Yi[†1,2,4], Jianwei Xu[†1,4(✉)], Yue Hu[3], Zhongwei Qiu[1,4], Sijing Li[5], Liang Huang[3], Bin Lv[3(✉)], Ling Zhang[1,4], and Yingda Xia[1,4]

[1] DAMO Academy, Alibaba Group, Hangzhou, China
qingcheng.xjw@alibaba-inc.com
[2] Department of Gastroenterology, The First Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, China
lvbin@medmail.com.cn
[3] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[4] Hupan Lab, Hangzhou, 310023, China
[5] Zhejiang University, China

Abstract. The development of foundation models (FMs) is crucial for advancing endoscopic image analysis. However, existing endoscopy FMs mainly rely on self-supervised learning from uni-modal images or videos, overlooking the rich semantic knowledge contained in clinical reports. Furthermore, effectively leveraging these records is hindered by a fundamental modality gap: structured anatomical descriptions are not naturally mapped to specific frames within the high-redundancy, uncurated visual streams. In this paper, we present EndoVLM, a novel vision-language FM pre-trained on over 348K endoscopic examinations, each pairing a clinical report with its corresponding image collection. An Anatomy-Guided Sparse Pooling mechanism utilizes textual descriptions as queries to drive sparse attention, efficiently aggregating semantically salient frames into anatomy-specific visual representations across redundant image-sets. Next, a Progressive Semantic-Aware Alignment strategy models clinical taxonomy (anatomy and pathological status) via structured soft targets, bridging the gap from global patient-level matching to fine-grained localized alignment. Finally, a Semantic-Concentrated Masked Autoencoder is applied exclusively to these semantic-rich frames, integrating low-level visual precision with robust high-level semantic representation. Extensive experiments across various downstream tasks demonstrate that EndoVLM outperforms existing foundation models and remains competitive with task-specific methods. Remarkably, EndoVLM also exhibits robust zero-shot generalization capabilities, highlighting its potential for broader clinical application. Codes are available at https://github.com/Scatteredrain/EndoVLM.

Keywords: Endoscopy · Foundation model · Vision-language pre-training.

[†] Equal contribution; The work was done during Z. Yi's internship at DAMO Academy.

1 Introduction

Recent advances in foundation models (FMs) have catalyzed a paradigm shift in the development of medical image analysis [2]. Through self-supervised learning (SSL) on massive unlabeled datasets [9,17], FMs learn highly transferable representations for downstream tasks [11,10,18], driving breakthroughs across various medical domains, including CT [25], X-ray [28], and histopathology [31]. Furthermore, since clinical workflows naturally pair images with detailed diagnostic reports, this multimodal synergy has inspired vision-language FMs—such as BiomedCLIP [35], LLaVA-Med [14], and fVLM [21]. By aligning visual features with rich textual insights, these models build more interpretable and clinically grounded AI systems.

However, existing endoscopy-specific FMs [32,33,26,13] remain predominantly visual self-supervised models, overlooking the rich clinical insights embedded in textual reports, which not only describe anatomical landmarks and pathological findings but also provide procedural context. Meanwhile, general vision-language pre-training models (VLMs) such as CLIP [19] typically rely on a one-to-one alignment between individual samples (e.g., a single image, short video clip, or 3D medical data such as CT) and their textual descriptions. In contrast, routine gastrointestinal (GI) endoscopy produces a single comprehensive report for dozens of unannotated images or lengthy videos—which we formulate as an *image-set*. Without explicit temporal or spatial annotations, it is exceptionally difficult to align a specific frame with a particular sentence or abnormal finding within the comprehensive report [15]. For example, a gastroscopy report covers **eight** regions (*esophagus, cardia, fundus, body, antrum, angularis, pylorus, duodenum*), while a colonoscopy assesses **nine** segments (*ileum, ileocecal region, ascending colon, hepatic flexure, transverse colon, splenic flexure, descending colon, sigmoid colon, rectum*). During examination, clinicians capture multiple images of each region from various angles and document morphology, mucosal features, and any pathological findings. The lack of per-frame labels makes it nearly impossible to identify the anatomical location or mucosal features in individual images, impeding precise image- or pixel-level cross-modal alignment. Addressing this bottleneck requires novel strategies that bridge structured clinical narratives with large-scale, unstructured visual corpora.

In this work, we present EndoVLM, a novel vision-language foundation model pre-trained on a massive cohort of 348K endoscopic cases. To bridge the gap between unordered image-sets and structured reports, we introduce three core components. First, an Anatomy-Guided Sparse Pooling (AGSP) mechanism employs query-driven sparse attention to distill semantically salient frames from redundant visual streams. Subsequently, we propose a Progressive Semantic-Aware Alignment (PSAA) strategy. By modeling clinical classifications—encompassing both anatomical structures and pathological states—through soft targets, PSAA facilitates a robust transition from coarse patient-level matching to fine-grained localized alignment. Finally, a Semantic-Concentrated Masked Autoencoder (SC-MAE) is applied strictly to these filtered frames, unifying low-level geometric precision with high-level clinical semantics. Extensive experiments across diverse

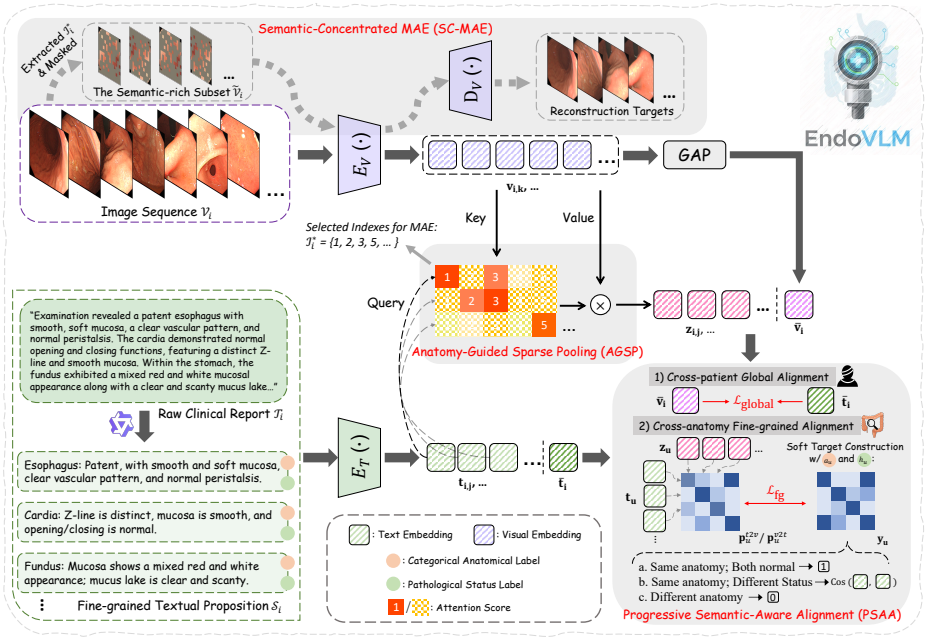

Fig. 1: **Overview of EndoVLM.** We align unordered images with reports via: AGSP for report-guided salient frame distillation and visual feature pooling, PSAA for hierarchical alignment, and SC-MAE for geometric regularization.

downstream tasks—including risk stratification, polyp segmentation, anatomy recognition, and video disease diagnosis—demonstrate that EndoVLM outperforms state-of-the-art foundation models and remains competitive with task-specific segmentation baselines using a minimal segmentation head. Remarkably, EndoVLM also exhibits robust zero-shot generalization capabilities, underscoring its immense potential as a highly adaptable backbone for scalable clinical deployment.

2 Method

We propose EndoVLM, a vision-language foundation model tailored for GI endoscopy. As shown in Fig. 1, to address the asymmetry between high-redundancy, unordered image-sets and raw clinical reports, our approach begins with a data construction pipeline, and followed by three core components: AGSP to distill semantically salient frames via anatomical queries; PSAA for hierarchical semantic alignment; and SC-MAE for robust visual representation learning.

2.1 Preliminaries and Data Construction

Given a dataset $\mathcal{D} = \{(\mathcal{V}_i, \mathcal{T}_i)\}_{i=1}^{N}$ of N examinations, each contains an unordered image-set $\mathcal{V}_i = \{I_{i,k}\}_{k=1}^{K_i}$ and a raw report $\mathcal{T}_i$. Standard global alignment paradigms (e.g., CLIP [19]) are suboptimal in this context due to the "many-to-many" semantic mismatch between visually redundant frames and information-dense textual findings. To bridge this semantic gap, we use Qwen3 [34] with a fixed clinical schema to parse $\mathcal{T}_i$ into M_i semantic triplets $\mathcal{S}_i = \{(s_{i,j}, a_{i,j}, h_{i,j})\}_{j=1}^{M_i}$. Here, $s_{i,j}$ represents a fine-grained textual proposition of a distinct morphological finding; $a_{i,j}$ is its anatomical label (mapped to the aforementioned 17 GI regions); and $h_{i,j} \in \{0, 1\}$ indicates the pathological status (normal or abnormal).

For feature extraction, a vision encoder E_V maps frames $I_{i,k}$ to local normalized embeddings $\mathbf{F}_i^v = \{\mathbf{v}_{i,k} \in \mathbb{R}^d\}_{k=1}^{K_i}$ and a global average-pooled vector $\bar{\mathbf{v}}_i$. Concurrently, a text encoder E_T encodes the full report $\mathcal{T}_i$ and propositions $s_{i,j}$ into a global normalized embedding $\bar{\mathbf{t}}_i$ and a set of fine-grained normalized embeddings $\mathbf{H}_i^t = \{\mathbf{t}_{i,j} \in \mathbb{R}^d\}_{j=1}^{M_i}$, respectively. The final representation for the i-th examination is formulated as the tuple: $(\mathbf{F}_i^v, \bar{\mathbf{v}}_i, \bar{\mathbf{t}}_i, \{\mathbf{t}_{i,j}\}, \{a_{i,j}\}, \{h_{i,j}\})$.

2.2 Anatomy-Guided Sparse Pooling (AGSP)

Endoscopic image-sets exhibit high redundancy (e.g., repetitive normal mucosa) that dilutes sparse abnormal signals. We propose AGSP, which uses fine-grained text embeddings as queries to selectively aggregate relevant visual evidence.

For each sentence embedding $\mathbf{t}_{i,j}$, we compute dot-product relevance $r_{j,k} = \mathbf{t}_{i,j}^\top \mathbf{v}_{i,k}$ with all frames in $\mathcal{V}_i$. Instead of standard cross attention, we use sparse attention by selecting top-K similar frames $\mathcal{I}_{i,j}$. The anatomy-specific visual representation $\mathbf{z}_{i,j}$ is derived via re-normalized attention over $\mathcal{I}_{i,j}$:

$$\mathbf{z}_{i,j} = \sum_{k \in \mathcal{I}_{i,j}} \frac{e^{r_{j,k}/\tau_a}}{\sum_{m \in \mathcal{I}_{i,j}} e^{r_{j,m}/\tau_a}} \mathbf{v}_{i,k}, \tag{1}$$

where $\tau_a = 0.07$ is the temperature and $K = 3$. A small K preserves a compact set of salient frames for each anatomical query while suppressing redundant normal views. This aligns the visual modality with the anatomy described in $\mathbf{t}_{i,j}$ while filtering noise.

2.3 Progressive Semantic-Aware Alignment (PSAA)

To capture both holistic context and fine-grained mucosal and pathological nuances, we employ a progressive alignment strategy.

Stage 1: Cross-Patient Global Alignment. We first enforce consistency between the patient-level image-set representation $\bar{\mathbf{v}}_i$ and the global report embedding $\bar{\mathbf{t}}_i$ using the symmetric InfoNCE [16] loss across a batch of size B:

$$\mathcal{L}_{\text{global}} = -\frac{1}{2B} \sum_{i=1}^{B} \left(\log \frac{e^{\langle \bar{\mathbf{v}}_i, \bar{\mathbf{t}}_i \rangle / \tau}}{\sum_{j=1}^{B} e^{\langle \bar{\mathbf{v}}_i, \bar{\mathbf{t}}_j \rangle / \tau}} + \log \frac{e^{\langle \bar{\mathbf{t}}_i, \bar{\mathbf{v}}_i \rangle / \tau}}{\sum_{j=1}^{B} e^{\langle \bar{\mathbf{t}}_i, \bar{\mathbf{v}}_j \rangle / \tau}} \right), \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is cosine similarity and τ is a learnable temperature.

Stage 2: Cross-Anatomy Fine-Grained Alignment. Global alignment alone is insufficient for distinguishing subtle mucosa attributes. Inspired by recent dense semantic alignment formulations [21,30], we perform fine-grained contrastive learning on a set of semantic units $\mathcal{U} = \{(\mathbf{z}_u, \mathbf{t}_u)\}_{u=1}^{M}$, aggregating all M valid local visual-text pairs $(\mathbf{z}_{i,j}, \mathbf{t}_{i,j})$ in the mini-batch.

We first compute the softmax-normalized image-to-text and text-to-image similarity distributions, denoted as $\mathbf{p}_u^{\mathrm{v2t}}$ and $\mathbf{p}_u^{\mathrm{t2v}}$, respectively. For any pair of units (u, k) within $\mathcal{U}$, the k-th element of these distributions is defined as:

$$p_{u,k}^{\mathrm{v2t}} = \frac{\exp(\langle \mathbf{z}_u, \mathbf{t}_k \rangle / \tau)}{\sum_{m=1}^{M} \exp(\langle \mathbf{z}_u, \mathbf{t}_m \rangle / \tau)}, \quad p_{u,k}^{\mathrm{t2v}} = \frac{\exp(\langle \mathbf{t}_u, \mathbf{z}_k \rangle / \tau)}{\sum_{m=1}^{M} \exp(\langle \mathbf{t}_u, \mathbf{z}_m \rangle / \tau)}. \tag{3}$$

To seamlessly progress from global cross-patient matching to localized alignment of anatomy and pathological status, we construct a taxonomy-aware soft target distribution $\mathbf{y}_u = [y_{u,1}, \ldots, y_{u,M}]^\top \in \mathbb{R}^M$. Let a_u and h_u be the categorical anatomical and pathological status labels of the u-th unit, respectively. The normalized target $y_{u,k}$ is computed as:

$$y_{u,k} = \frac{\hat{y}_{u,k}}{\sum_{m=1}^{M} \hat{y}_{u,m}}, \text{ with } \hat{y}_{u,k} = \begin{cases} 1, & \text{if } a_u = a_k \text{ and } h_u = h_k = 0; \\ \langle \mathbf{t}_u, \mathbf{t}_k \rangle, & \text{if } a_u = a_k \text{ and } (h_u = 1 \text{ or } h_k = 1); \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

By avoiding repulsion among identical healthy anatomies and adaptively scaling abnormal alignments via textual similarity, this formulation drives the model to autonomously disentangle anatomical and pathological semantics. The textual similarity term provides a semantic relational prior within each anatomical group, allowing related abnormal findings to share softer supervision. The fine-grained alignment loss is defined as follows, where CE denotes cross-entropy.

$$\mathcal{L}_{\mathrm{fg}} = \frac{1}{2M} \sum_{u=1}^{M} \left(\mathrm{CE}(\mathbf{y}_u, \mathbf{p}_u^{v2t}) + \mathrm{CE}(\mathbf{y}_u, \mathbf{p}_u^{t2v}) \right). \tag{5}$$

2.4 Semantic-Concentrated Masked Autoencoder (SC-MAE)

To efficiently capture low-level textures, SC-MAE applies the MAE paradigm [9] exclusively to the semantic-rich subset $\tilde{\mathcal{V}}_i = \{I_{i,k} \in \mathcal{V}_i \mid k \in \mathcal{I}_i^*\}$, where $\mathcal{I}_i^* = \bigcup_j \mathcal{I}_{i,j}$ denotes the unified index set aggregated from all anatomy-specific frames retrieved by AGSP. It masks 75% of patches in $\tilde{\mathcal{V}}_i$, and reconstruct pixels via a decoder D_V, yielding an MSE loss $\mathcal{L}_{\mathrm{MAE}}$. The overall training objective is as follows, where λ_1, λ_2 are balancing weights.

$$\mathcal{L} = \mathcal{L}_{\mathrm{global}} + \lambda_1 \mathcal{L}_{\mathrm{fg}} + \lambda_2 \mathcal{L}_{\mathrm{MAE}}. \tag{6}$$

Table 1: Performance comparison. We report F1 (%) for PolypDiag, macro AUC/F1 (%) for LIMUC, Dice (%) for polyp segmentation (CVC-12k, Kvasir (Kva), ClinicDB (Clinic), Endoscene (Endo), ColonDB, ETIS). †: pre-trained with our dataset. '–': not reported and no public code or weights. The best and second-best results are in bold and underlined.

Model	Polyp Diag	LIMUC	CVC -12k	Seen		Unseen		
				Kva	Clinic	Endo	Colon	ETIS
General FMs								
MAE [9]	91.1	93.4/72.7	83.6	88.4	88.5	84.0	75.5	74.7
CLIP [19]	90.0	92.0/69.2	80.0	88.3	90.0	85.4	75.8	71.6
DINOv2 [17]	93.3	93.8/73.6	83.8	90.7	91.4	88.6	79.8	77.8
DINOv3 [23]	93.9	93.9/73.7	84.5	91.0	91.1	89.1	<u>81.6</u>	77.7
BiomedCLIP [35]	90.2	92.6/72.0	81.5	88.0	87.8	82.8	72.7	66.1
MAE† [9]	93.7	93.1/73.6	84.2	88.4	88.5	83.9	75.5	74.7
CLIP† [19]	90.1	85.2/60.2	66.2	60.1	70.6	28.8	32.8	25.6
DINOv3† [23]	94.8	94.0/<u>74.0</u>	<u>85.8</u>	91.4	92.2	89.5	<u>81.6</u>	80.1
Endo SSL								
EndoFM [32]	90.7	93.0/72.9	73.9	87.7	87.1	83.2	71.2	63.5
EndoFM-LV [33]	<u>96.3</u>	83.3/57.4	83.2	62.2	67.7	42.5	35.0	29.8
EndoMamba [26]	95.0	92.1/70.1	85.4	87.5	87.8	86.2	69.4	60.2
EndoDINO [4]	–	93.7/70.6	–	–	–	–	–	–
GastroNet-5M [13]	–	**95.5**/–	–	–	–	–	–	–
Task specific model (Segmentation)								
Polyp-PVT [5]	–	–/–	–	<u>91.7</u>	<u>93.7</u>	90.0	80.8	78.7
VM-UNet [20]	–	–/–	–	91.3	92.6	88.6	79.8	76.1
PolyMamba [7]	–	–/–	–	**91.9**	**94.0**	<u>90.4</u>	81.5	**82.9**
EndoVLM	**97.3**	<u>94.5</u>/**74.4**	**86.4**	**91.9**	93.1	**90.8**	**82.8**	<u>82.0</u>

3 Experiments and Results

Pre-training dataset. We retrospectively collected over 400K endoscopic examinations from two medical centers[1], each consisting of a clinical report and its corresponding image package. By leveraging Qwen3 [34] to analyze report content, we filtered out ineligible cases—including post-operative exams, incomplete procedures (defined as gastroscopy reports missing any of the 8 standard anatomical regions, or colonoscopy reports missing any of the 9 intestinal landmarks), and other non-conforming studies. The final curated dataset contains 348K examinations, comprising more than 18.6M endoscopic images.

Downstream task setup. To evaluate the versatility and robustness of EndoVLM, we conducted experiments across various tasks: (i) *PolypDiag for video polyp diagnosis* [27]. Following EndoFM-LV [33], we sampled 16 frames per video. Frame-level features extracted by our image-based backbone were aggregated via mean pooling to form a video-level representation, which was then clas-

[1] The study was conducted in accordance with the Declaration of Helsinki and approved by the institutional review board of the lead medical center (No. 2024-KLS-489-02).

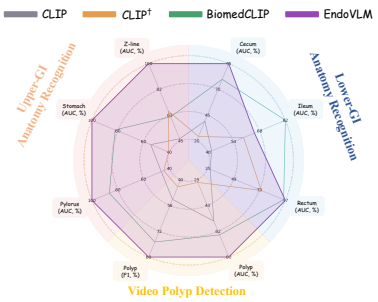

Fig. 2: Zero-shot performance comparison.

Table 2: Ablation study of different loss components. $\mathcal{L}_{\text{glo}}$: Global Alignment; $\mathcal{L}_{\text{fg}}$: Fine-Grained Alignment; $\mathcal{L}_{\text{MAE}}$: Masked Autoencoder.

Components			LIMUC	CVC-12k
$\mathcal{L}_{\text{glo}}$	$\mathcal{L}_{\text{fg}}$	$\mathcal{L}_{\text{MAE}}$	**AUC**	**Dice**
✓			85.2	66.2
✓	✓		94.1	85.6
		✓	93.1	84.2
✓	✓	✓	**94.5**	**86.4**

sified using a linear head. (ii) *CVC-12k for video polyp segmentation* [1] and (iii) *Kvasir-SEG* [12], *ClinicDB* [1], *ColonDB* [24], *ETIS* [22], *EndoScene* [29] *for polyp segmentation generalization assessment*. A linear segmentation head is implemented by reshaping the backbone's patch tokens into spatial feature maps $(B, D, H/P, W/P)$, followed by **a single convolutional layer and bilinear interpolation for final dense prediction**. We adopt the same data splits and evaluation protocol as EndoFM-LV [33] on CVC-12k, and strictly follow the setup of Polyp-PVT [5] for generalization assessment, where "seen" and "unseen" denote whether a dataset is used during downstream fine-tuning. (iv) *LIMUC for ulcerative colitis severity grading* [18]. We append a linear classification head to the EndoVLM, and follow the same experimental settings as GastroNet5M [13]. (v) *Zero-Shot Transfer*. We assess the model's generalizability via zero-shot anatomical recognition on Hyper-Kvasir [3] (covering upper & lower GI tracts) and video disease diagnosis on PolypDiag [27], directly leveraging the pre-aligned visual-semantic space without task-specific fine-tuning.

Implementation details. The vision encoder and language encoder are instantiated as ViT-B/16 [6] and PubMedBERT [8], respectively. All input images are resized to 224×224 for pre-training. The entire framework is trained for 100 epochs on NVIDIA A800 GPUs with a total batch size of 96. We optimize the network using the AdamW optimizer with a base learning rate of 1.5e-4 and a weight decay of 0.05. The loss weights λ_1 and λ_2 are empirically set to 1.

Comparison experiment. We compare our method against recent SOTA methods, including general visual and vision-language FMs (DINOv2/v3 [17,23], MAE [9], CLIP [19], BiomedCLIP [35]), endoscopy-specific visual SSL models (EndoFM [32], EndoFM-LV [33], EndoMamba [26], EndoDINO [4], GastroNet-5M [13]), and task-specific segmentation models [5,7,20]. We also pre-train the general FMs from scratch on our dataset. Our approach surpasses all FMs, trailing only GastroNet-5M on LIMUC. Notably, (1) on video tasks (PolypDiag/CVC-12k), EndoVLM exceeds video-specific FMs [32,33,26] using a simple image-level mean-pooling aggregation, bypassing complex temporal modeling. (2) For dense

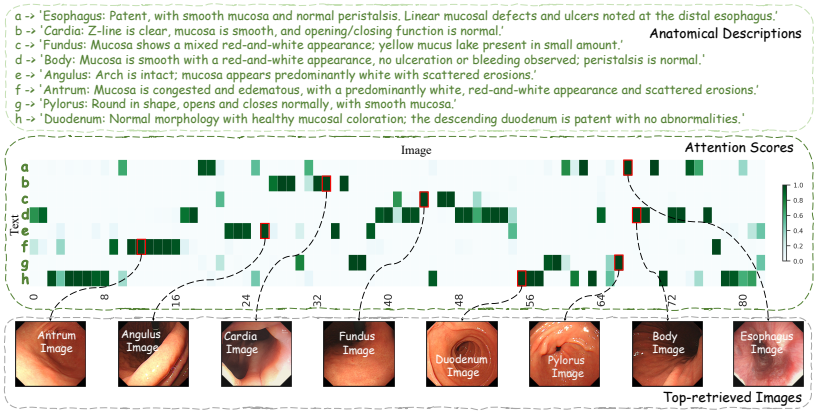

Fig. 3: Visualization of AGSP **attention scores**. The heatmap illustrates similarities between unordered image embeddings and anatomical text embeddings, where *a*-*h* denote specific **anatomical descriptions** in a gastroscopy report. Highlights identify the **top-retrieved images** for each specific anatomy.

prediction tasks, EndoVLM achieves competitive or superior segmentation performance using only a single convolutional layer with bilinear interpolation, without complex decoders used in task-specific models [5,20,7]. Superior results on unseen ColonDB and ETIS datasets further validate its robust generalization.

Zero-shot evaluation. We evaluate zero-shot performance on three clinical classification tasks using the prompt: *"This is an image of {cls}."* As shown in Fig. 2, EndoVLM significantly outperforms general (CLIP) and medical-specific (CLIP†, BiomedCLIP) FMs. It achieves near-perfect transferability ($\sim$ 100% AUC) in Upper-GI anatomy recognition and surpasses BiomedCLIP by 18% AUC in the challenging video disease diagnosis task, validating that our taxonomy-aware alignment yields transferable representations without task-specific fine-tuning. The lower ileum result reflects a category-specific limitation, as ileum descriptions are often procedural and less uniform in colonoscopy reports.

Ablation study. Ablation results in Table 2 validate the efficacy of our proposed components. Global alignment ($\mathcal{L}_{\text{glo}}$) alone–mimicking standard CLIP paradigm–yields poor performance (85.2% AUC and 66.2% Dice), as coarse patient-level matching struggles to capture localized lesion semantics. Integrating fine-grained alignment ($\mathcal{L}_{\text{fg}}$) triggers a significant leap, boosting AUC by 8.9% and Dice by 19.4%, which highlights the critical role of our taxonomy-aware soft targets in aligning detailed anatomical and pathological features. Finally, SC-MAE ($\mathcal{L}_{\text{MAE}}$) further refines representations to achieve the best performance (94.5% AUC and 86.4% Dice), demonstrating that pixel-level reconstruction provides complementary geometric regularization to semantic alignment.

Qualitative analysis and interpretability. To validate the interpretability of our AGSP module, we visualize the attention scores between the sentence embeddings $\mathbf{t}_{i,j}$ and image embeddings $\mathbf{v}_{i,k}$. As shown in Fig. 3, the attention

matrix exhibits highly localized and target-specific activations. EndoVLM accurately retrieves the most representative frames for eight distinct upper-GI regions (e.g., Esophagus, Duodenum) from a redundant set of over 80 images.

4 Conclusion

We present EndoVLM, the first GI VLM pre-trained on a large-scale dataset of unordered GI endoscopy image-sets and clinical reports. To bridge the profound semantic gap between redundant visual frames and structured narratives, we propose a unified framework that seamlessly integrates semantics-driven salient frame distillation, progressive cross-modal alignment, and geometric reconstruction. Pre-trained on a massive cohort of 348K cases, EndoVLM demonstrates exceptional adaptability and robust zero-shot generalization. It significantly outperforms existing uni-modal foundation models and complex task-specific architectures across diverse downstream tasks, providing a highly scalable foundation for next-generation AI-assisted endoscopy.

Acknowledgments. This study was supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2023C03050).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)
2. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
3. Borgli, H., Thambawita, V., Smedsrud, P.H., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data **7**(1), 283 (2020)
4. Dermyer, P., Kalra, A., Schwartz, M.: Endodino: A foundation model for gi endoscopy. arXiv preprint arXiv:2501.05488 (2025)
5. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Fu, R., Hu, S., Zheng, X., Tang, C., Liu, X.: Polymamba: Spatial-prior guided mamba for polyp segmentation with high-frequency enhancement. In: MICCAI. pp. 455–465. Springer (2025)

8. Gu, Y., Tinn, R., Cheng, H., et al.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare **3**(1), 1–23 (2021)
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
10. Hu, Q., Yi, Z., Zhou, Y., Huang, F., Liu, M., Li, Q., Wang, Z.: Monobox: Tightness-free box-supervised polyp segmentation using monotonicity constraint. In: AAAI. vol. 39, pp. 3572–3580 (2025)
11. Hu, Q., Yi, Z., Zhou, Y., Peng, F., Liu, M., Li, Q., Wang, Z.: Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In: MICCAI. pp. 531–541. Springer (2024)
12. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International conference on multimedia modeling. pp. 451–462. Springer (2019)
13. Jong, M.R., Boers, T.G., Fockens, K.N., et al.: Gastronet-5m: A multicenter dataset for developing foundation models in gastrointestinal endoscopy. Gastroenterology (2025)
14. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. NeurIPS **36**, 28541–28564 (2023)
15. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: CVPR. pp. 9879–9889 (2020)
16. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
17. Oquab, M., Darcet, T., Moutakanni, T., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
18. Polat, G., Kani, H.T., Ergenc, I., Ozen Alahdab, Y., Temizel, A., Atug, O.: Improving the computer-aided estimation of ulcerative colitis severity according to mayo endoscopic score by using regression-based deep learning. Inflamm. Bowel Dis. **29**(9), 1431–1439 (2023)
19. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)
20. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. ACM Transactions on Multimedia Computing, Communications and Applications (2024)
21. Shui, Z., Zhang, J., Cao, W., Wang, S., Guo, R., Lu, L., Yang, L., Ye, X., Liang, T., Zhang, Q., et al.: Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. arXiv preprint arXiv:2501.14548 (2025)
22. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International journal of computer assisted radiology and surgery **9**(2), 283–293 (2014)
23. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
24. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging **35**(2), 630–644 (2015)
25. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: CVPR. pp. 20730–20740 (2022)

26. Tian, Q., Liao, H., Huang, X., Yang, B., Lei, D., Ourselin, S., Liu, H.: Endomamba: an efficient foundation model for endoscopic videos via hierarchical pre-training. In: MICCAI. pp. 224–234. Springer (2025)
27. Tian, Y., Pang, G., Liu, F., Liu, Y., Wang, C., Chen, Y., Verjans, J., Carneiro, G.: Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In: MICCAI. pp. 88–98. Springer (2022)
28. Tiu, E., et al.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022)
29. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering **2017**(1), 4037190 (2017)
30. Wang, R., Tang, F., Yao, Q., Yan, R., Zhang, X., Huang, Z., Lai, H., He, Z., Tao, X., Jiang, Z., et al.: Simcrop: Radiograph representation learning with similarity-driven cross-granularity pre-training. In: MICCAI. pp. 563–573. Springer (2025)
31. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature **634**(8035), 970–978 (2024)
32. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: MICCAI. pp. 101–111. Springer (2023)
33. Wang, Z., Liu, C., Zhu, L., Wang, T., Zhang, S., Dou, Q.: Improving foundation model for endoscopy video analysis via representation learning on long sequences. IEEE J. Biomed. Health Inform. (2025)
34. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
35. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)