Focus-to-Perceive Representation Learning: A Cognition-Inspired Hierarchical Framework for Endoscopic Video Analysis

Supplementary Material

In the supplementary material, we first summarize the endoscopic video datasets used in our study including pre-training corpora and downstream benchmarks in Section A. Next, we provide implementation details for reproducibility, covering pre-training settings, evaluation protocols, and competitors, in Section B. In Section C, we present additional experimental results including surgical phase recognition, extended ablations, and further visualizations (segmentation/detection comparisons and mask visualizations). Finally, we analyze the failure case and present future research in Section D.

A. Dataset Details

The datasets used in the experiment include 7 pre-training datasets and 4 downstream task datasets. Although EndoFM-LV [26] and EndoMamba [20] extend the pre-training dataset, the extended datasets are not fully disclosed. Therefore, we still use publicly available datasets including Colonoscopic [13], SUN-SEG [7], LDPolypVideo [12], Hyper-Kvasir [2], Kvasir-Capsule [19], CholecTriplet [14], Renji-Hospital [25], PolypDiag [21], CVC-12k [1], KUMC [9], and Cholec80 [23] as evaluation benchmarks for all experiments.

A.1. Pre-training Datasets

We leverage 7 medical video datasets spanning diagnostic endoscopy, capsule endoscopy, and laparoscopic surgery as unlabeled sources for self-supervised pre-training. In total, these datasets provide 32,896 videos with 5,024,101 frames. Unless otherwise stated, we sample 30 FPS short clips with an average duration of 5 seconds from each video. Representative frames from all datasets are shown in Fig. 1, illustrating the diversity in imaging modality, anatomical site, and visual appearance.

Colonoscopic [13] is a small-scale medical video dataset collected from routine screening colonoscopies, focusing on gastrointestinal lesions (mainly colorectal polyps) observed under both white-light and narrow-band imaging. It contains 76 short colonoscopy videos centered on individual lesions and captures realistic in-procedure appearance variations such as camera motion, specular highlights, and illumination changes.

SUN-SEG [7] is a large-scale colonoscopy video benchmark for polyp-centric analysis, constructed from the SUN database collected at tertiary hospitals. It consists of 1,106 short video clips with 158,690 frames in total, covering diverse polyp sizes, morphologies, and anatomical locations under realistic screening conditions such as rapid camera motion, specular highlights, fluids, and low-contrast mucosa.

LDPolypVideo [12] is a diverse colonoscopy video dataset designed to capture real-world variability in colorectal polyps across different patients and examination settings. It contains 160 colonoscopy videos with 40,266 frames, where polyps exhibit wide variations in size, shape, texture, and viewing angle, together with challenging artifacts such as motion blur and occlusions.

Hyper-Kvasir [2] is a large-scale gastrointestinal endoscopy dataset collected from routine gastro- and colonoscopy examinations, covering both upper and lower GI tract with a broad spectrum of anatomical landmarks and pathological findings. We use only its video subset, which comprises 374 endoscopic recordings (about 10 hours, ∼0.9M frames) of real clinical procedures.

Kvasir-Capsule [19] is a large-scale video capsule endoscopy dataset consisting of 117 complete small-bowel examinations acquired with wireless capsule cameras, totaling approximately 4.7M frames. The recordings capture long-range traversal of the gastrointestinal mucosa with diverse findings and image quality variations typical of capsule endoscopy.

CholecTriplet [14] is a laparoscopic surgery video dataset built on CholecT50, containing 50 full-length cholecystectomy procedures recorded from the operative camera (about

Figure 1. Example frames of the 7 pre-training datasets used in this work.

10^5 frames at 1 FPS). It provides long, workflow-rich sequences with frequent instrument–tissue interactions and view changes that are complementary to diagnostic endoscopy.

Renji-Hospital [25] is a large-scale clinical endoscopy video dataset collected from routine upper and lower gastrointestinal examinations at the Baoshan Branch of Renji Hospital in Shanghai, China. It comprises 16,494 colonoscopy clips (2,491,952 frames) and 7,653 gastroscopy clips (1,170,753 frames), covering common mucosal abnormalities such as polyps and erosions under real-world screening and diagnostic conditions.

A.2. Downstream Datasets

We evaluate our representations on 4 labeled benchmarks covering disease diagnosis, polyp segmentation, polyp detection, and surgical phase recognition. Representative frames from these downstream datasets are shown in Fig. 2.

PolypDiag [21] is a large-scale endoscopy video dataset constructed from Hyper-Kvasir and LDPolypVideo, targeting lesion-level disease diagnosis. It contains 253 short endoscopic clips with 485,561 frames in total, where each clip is assigned a binary video-level label indicating the presence or absence of neoplastic lesions (polyps or early cancers). We use these video-level lesion labels to fine-tune and evaluate our model on the disease classification task.

CVC-12k [1] is a colonoscopy image dataset built from 18 video sequences, comprising 11,954 frames of which most contain at least one colorectal polyp. Each frame is annotated with a polyp region, originally provided as approximate masks around the visible lesions. Following Endo-FM, we reorganize the annotated frames into 29 short video clips (612 labeled frames) and convert the polyp regions into pixel-wise masks. This yields a supervised benchmark for evaluating polyp segmentation performance.

KUMC [9] is a colonoscopy polyp detection and classification dataset collected at the University of Kansas Medical Center. It consists of 80 video sequences curated from routine colonoscopy examinations, where frames are annotated with bounding boxes and categorical labels for individual polyps (adenomatous vs. hyperplastic). We use the same data partitioning method as in Endo-FM, which comprises 53 sequences containing approximately 19,832 labeled frames. We leverage the bounding-box and category annotations to evaluate the performance of polyp detection.

Cholec80 [23] is a laparoscopic cholecystectomy dataset comprising 80 full-length surgical videos (about 40 minutes each) recorded at 25 FPS from 13 surgeons. Each frame is annotated with one of seven surgical phases, and tool-presence labels are provided at 1 FPS. We follow the standard 40/40 train–test split and use only the phase annotations to benchmark cross-domain transfer on surgical workflow recognition.

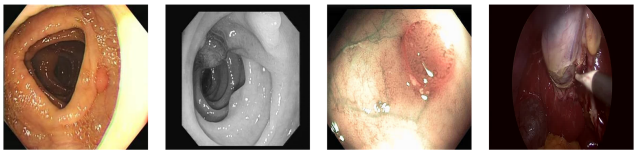

Figure 2. Example frames of the 4 downstream datasets used in this work.

B. Implementation Details

B.1. Pre-training Settings

Our *FPRL* is built based on EndoMamba-S [20], where the patch size and embedding dimension are set to 16 and 384, respectively. For every input video, we generate 3 views with spatial size 224×224 and 2 frames per view from a temporal window. The model is trained using the AdamW optimizer [11] with a base learning rate of 1.5e-4, a cosine learning rate schedule for 400 epochs, and a batch size of 64, with the first 40 epochs dedicated to linear warmup. For feature alignment, the pretrained VideoMamba-S [10] serves as the teacher model. In accordance with the EndoMamba configuration, the weight for the feature alignment loss is set to $\lambda_2 = 0.8$. The other hyperparameters of the loss functions are established as follows: $\lambda_1 = 1.0$, $\lambda_3 = 1.0$, and $\lambda_{\text{pf}} = 20$, based on preliminary experiments. The general hyperparameter settings for our *FPRL* framework during the training process are summarized in Table 1. All experiments are conducted using PyTorch [17] on a Linux machine equipped with four NVIDIA A800 GPUs.

B.2. Evaluation Settings

For downstream fine-tuning, we utilize the following setup on a single NVIDIA A800 GPU. *1) PolypDiag*: We sample 8 frames at a resolution of 224×224 from each video as input, utilizing a pre-trained model to initialize the backbone and appending randomly initialized linear layers, and train for 20 epochs. The SGD optimizer is employed, with the learning rate set to 1e-3, momentum to 0.9, and batch size to 4. *2) CVC-12k*: A TransUNet [3] equipped with *FPRL* as the backbone is implemented. The AdamW optimizer is used to optimize the overall parameters by setting the learning rate as 1e-4, weight decay as 5e-2 and the batch size as 1. We resize the spatial resolution to 224×224 and fine-tune for 150 epochs. *3) KUMC*: We implement an STFT [27] with our pre-trained model as backbone for generating a feature pyramid. We resize the spatial size to 640×640 and train for 24k iterations. The SGD optimizer is used to optimize the overall parameters by setting the learning rate as 2.5e-3, weight decay as 1e-4 and momentum as 0.9. *4) Cholec80*: We utilize SV-RCNet [8] for endoscopic surgical phase recognition, in which *FPRL* serves as a temporal

Table 1. Pre-training settings.

Hyperparameter	Value
Sampling strategies	
temporal window length	50
number of view frames	2
crop size	224×224
mask ratio	0.9
Optimizing settings	
optimizer	AdamW
learning rate schedule	Cosine
weight decay	0.05
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
patch size	16
base learning rate	1.5×10^{-4}
warm-up epochs	40
pretraining epochs	400
batch size of each GPU	64
feature dimension	384
Loss functions	
weight of reconstruction loss	$\lambda_1 = 1$
weight of feature align loss	$\lambda_2 = 0.8$
weight of contrastive loss	$\lambda_3 = 1$
weight of squeezing loss	$\lambda_{\text{pf}} = 20$

module for extracted features. The input frames are resized to 224×224 and the model is trained for 25 epochs. Both *FPRL* and randomly initialized modules are updated during downstream fine-tuning. For the evaluation metrics, following the previous works, we use F1 score for PolypDiag, Dice for CVC-12k, F1 score for KUMC, and accuracy for Cholec80.

B.3. Competitors

We compare the proposed *FPRL* with several recent SOTA approaches for endoscopy video analysis. These approaches include:

- FAME [4] utilizes foreground–background merging to alleviate background bias and enhance motion-aware video representations.
- ProViCo [16] proposes a probabilistic video contrastive learning scheme that models clip-wise uncertainty in the latent space.
- VCL [18] focuses on jointly learning static and dynamic concepts to improve video representation modeling.
- ST-Adapter [15] inserts lightweight spatio-temporal adapters into frozen image backbones for parameter-efficient image-to-video transfer.
- VideoMAE [22] is a masked autoencoder baseline that learns video representations by reconstructing highly masked video tubes.
- Endo-FM [25] develops a transformer-based foundation model pre-trained on large-scale endoscopic videos.
- DropMAE [28] augments masked autoencoding with spatial-attention dropout to better capture temporal correspondences for downstream tracking and segmentation tasks.
- VideoMAE V2 [24] scales VideoMAE with dual masking and large-scale pre-training to build a general video foundation model.
- M^{2}CRL [6] integrates multi-view masked modeling with contrastive learning, tailored for endoscopic video pre-training.
- VideoMamba [10] introduces a state-space-based video backbone that models long-range spatio-temporal dynamics with linear complexity.
- EndoFM-LV [26] extends Endo-FM to a minute-level pre-training framework on long endoscopy video sequences.
- EndoMamba [20] is an efficient endoscopic foundation model built on bidirectional and causal Mamba blocks under a hierarchical pre-training scheme.

C. Additional Experimental Results

C.1. Surgical Phase Recognition

To assess the generalizability of *FPRL* to long-horizon reasoning, we further evaluate it on surgical phase recognition using the Cholec80 dataset. We follow the standardized protocol adopted in prior works, training on the official training videos and reporting frame-wise phase classification accuracy on the test split. From Table 2, we can observe that the proposed *FPRL* achieves 85.3 ± 8.0% accuracy, outperforming recent self-supervised and foundation-model baselines such as M^{2}CRL, VideoMamba, EndoFM-LV, and EndoMamba, which indicates that the representations learned by *FPRL* transfer well to long-horizon workflow modeling.

Table 2. Surgical phase recognition.

Method	Venue	Year	Accuracy
FAME [4]	CVPR	2022	81.9 ± 9.2
ProViCo [16]	CVPR	2022	82.3 ± 8.5
ST-Adapter [15]	NeurIPS	2022	81.0 ± 8.7
EndoSSL [5]	MICCAI	2023	83.0 ± 8.0
Endo-FM [25]	MICCAI	2023	82.8 ± 9.1
M^{2}CRL [6]	NeurIPS	2024	83.5 ± 8.8
VideoMamba [10]	ECCV	2024	84.1 ± 8.9
EndoFM-LV [26]	JBHI	2025	85.1 ± 7.9
EndoMamba [20]	MICCAI	2025	84.4 ± 8.4
FPRL	Ours	-	**85.3 ± 8.0**

Table 3. Ablation on model architecture variants.

Model Variant	Performance (%)		
	Cla.	Seg.	Det.
EndoMamba-T	92.2 ± 0.4	83.8 ± 0.3	81.7 ± 0.9
EndoMamba-S	**95.2 ± 0.3**	86.1 ± 0.1	**89.8 ± 0.1**
EndoMamba-M	92.3 ± 0.1	**87.4 ± 0.5**	89.5 ± 0.2

Table 4. Ablation on Decoder depth and CVMFC blocks.

Stacked Layers		Performance (%)		
Decoder	CVMFC	Cla.	Seg.	Det.
4	1	95.2 ± 0.3	**86.1 ± 0.1**	**89.8 ± 0.1**
	2	94.4 ± 0.5	85.3 ± 0.8	85.2 ± 1.0
	3	92.9 ± 0.4	86.1 ± 0.7	87.4 ± 0.9
8	1	93.8 ± 0.4	85.4 ± 0.5	84.6 ± 0.4
	2	**96.0 ± 0.7**	85.5 ± 0.8	87.7 ± 0.8
	3	89.9 ± 0.6	84.9 ± 0.2	86.7 ± 0.9

Table 5. Ablation on loss formulations.

Loss Formulations		Performance (%)		
Similarity	Regression	Cla.	Seg.	Det.
Cosine	ℓ_1	92.3 ± 0.5	85.0 ± 0.8	87.3 ± 1.0
	ℓ_2	**95.2 ± 0.3**	**86.1 ± 0.1**	**89.8 ± 0.1**
Cross-Entropy	ℓ_1	91.1 ± 0.5	80.3 ± 0.7	85.9 ± 0.9
	ℓ_2	92.9 ± 0.7	82.6 ± 0.8	87.5 ± 0.8

Table 6. Ablation on the number of sampled frames.

Number of Frames	Performance (%)		
	Cla.	Seg.	Det.
1	94.4 ± 1.6	86.0 ± 0.9	88.6 ± 1.1
2	**95.2 ± 0.3**	**86.1 ± 0.1**	**89.8 ± 0.1**
3	93.0 ± 0.5	85.2 ± 0.4	85.4 ± 0.1

C.2. Additional Ablations

Ablation on Different Architectures. As shown in Table 3, scaling the backbone from EndoMamba-T to EndoMamba-S consistently improves performance across all three downstream tasks, particularly in detection, which shows an improvement of +8.1%. However, further scaling to EndoMamba-M primarily benefits segmentation while slightly degrading classification and detection performance. This suggests that the small variant provides a more favorable balance between capacity and performance.

Ablation on Auxiliary Module. Table 4 analyzes the impact of decoder depth and the number of cross-view masked feature completion (CVMFC) blocks. Notably, a shallow 4-layer decoder equipped with a single CVMFC block achieves optimal overall performance. In contrast, deeper decoders and multiple CVMFC blocks yield only marginal gains or have adverse effects. This observation suggests that excessive temporal decoding may lead to over-smoothing of features, ultimately hindering dense prediction capabilities.

Ablation on Loss Formulations. As reported in Table 5, the combination of cosine similarity with an ℓ_2 regression term significantly outperforms all other loss formulations across classification, segmentation, and detection tasks. Substituting either cosine similarity with cross-entropy or ℓ_2 with ℓ_1 results in a marked decline in performance, thereby confirming that regressing continuous teacher features using a cosine + ℓ_2 objective is more effective than alignment based on discrete classification.

Ablation on the Number of Sampled Frames. Table 6 investigates the effects of the number of sampled frames. Utilizing two sparsely sampled frames yields optimal overall performance and surpasses single-frame training. However, incorporating three frames consistently detracts from performance across all tasks. This decline is likely due to the introduction of redundant non-semantic motion present in endoscopic videos when additional views are included.

C.3. Additional Visualization Results

Qualitative Results for Segmentation. Fig. 3 shows a visual comparison of segmentation results between our method and other self-supervised pre-training approaches on the CVC-12k dataset. Polyps frequently present with indistinct boundaries and considerable variations in shape, which pose significant challenges for accurate segmentation. Nevertheless, our approach consistently yields superior results when compared to other leading self-supervised methods. In particular, as shown in the first and second rows for larger polyps, while other methods often misclassify or overlook certain lesion regions, our approach effectively distinguishes polyps from normal tissues. Moreover, although all methods exhibit comparable performance in segmenting isolated small polyps, only our method successfully delineates all polyps when multiple small instances are present together within a single frame.

Qualitative Results for Detection. Fig. 4 presents the detection results obtained from our method and other competing self-supervised pre-training models applied to the KUMC dataset. Our approach exhibits superior performance in both boundary recognition and localization of small polyps. Although all methods can generally identify the actual lesion area in high-contrast scenarios (as shown in the first and second rows), other methods tend to incorporate extraneous normal tissue, whereas our approach achieves a higher level of localization precision. Moreover, even under challenging conditions, such as low contrast, blurred polyp boundaries, or complex backgrounds, our ap-

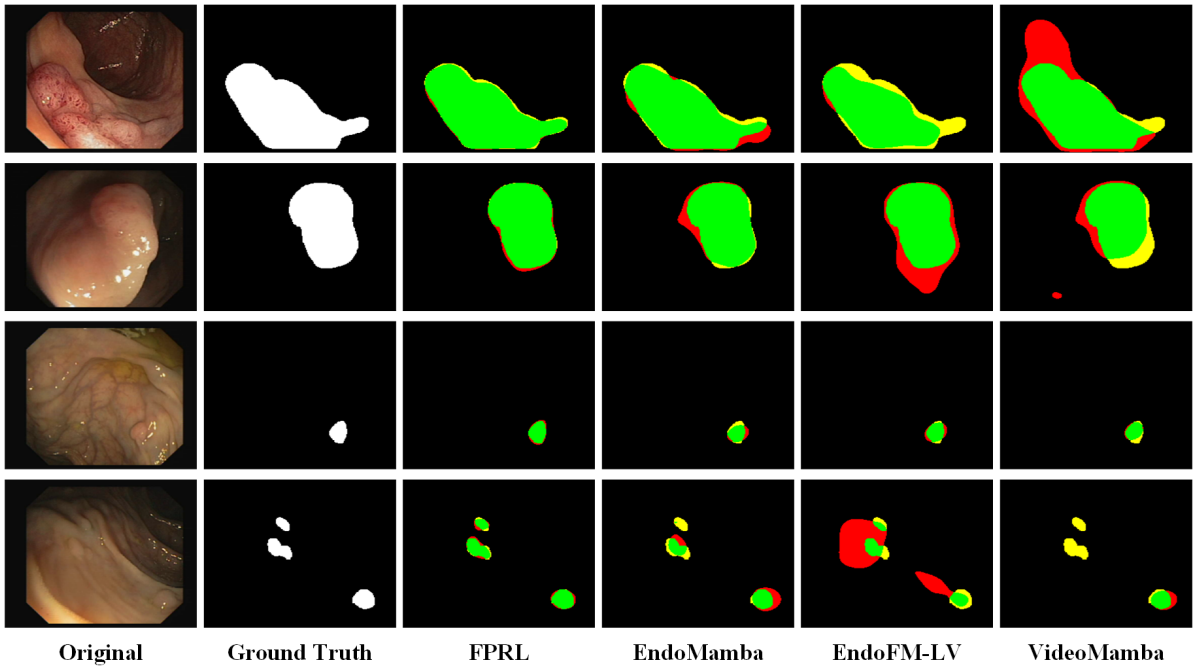

Figure 3. Qualitative results for segmentation task on the CVC-12k dataset, where green, red, and yellow regions represent the true positive, false positive, and false negative, respectively.

proach remains robust in accurately identifying polyp locations. This capability is attributed to the effective representation learning of salient lesion regions facilitated by our proposed *Static Semantic Focus* mechanism.

Mask Generation Process. Figure 5 illustrates our teacher-prior adaptive masking strategy. We generate a saliency map from the frozen teacher network, capturing lesion-aware global priors, while a lightweight network produces a complementary map emphasizing view-specific cues such as local contrast. These are fused into a unified importance map, where we employ a Top-K sampling strategy to select the most informative regions, retaining only the corresponding patches while masking out the remainder. This process suppresses irrelevant regions like background and specular reflections, directing model focus to clinically meaningful structures. The resulting masks (shown in the last column) concentrate on polyp interiors and boundaries, promoting more focused reconstruction and stable pre-training.

D. Failure Case Analysis

D.1. Failure Case

As reported in Table 6, the single-frame setting consistently underperforms the two-frame configuration, which appears at odds with our design objective of suppressing dynamic redundancy and non-semantic motion by keeping each view as short as possible. We argue that this discrepancy mainly stems from the imperfect quality of real endoscopic videos. In practice, many sequences contain frames that are heavily affected by motion blur, abrupt camera shake, illumination flicker, specular highlights, or transient occlusions from tools and fluids (see Fig. 6). Under the single-frame regime, such degradations directly contaminate the only available observation in a view, thereby corrupting the supervision signal for reconstruction and making the learned representations highly sensitive to occasional low-quality frames.

D.2. Future Research

From a broader perspective, the above failure case indicates that the gap between the single-frame and two-frame settings is largely driven by data quality issues rather than by the sparse-view design itself. This suggests three main directions for further improvement. First, the frame sampling strategy could be made more robust to low-quality observations, so that views are less likely to be dominated by severe blur, illumination fluctuation, or transient occlusions. Second, the pre-training framework could be refined to leverage sparse multi-frame information in a more systematic manner, utilizing additional frames primarily to enhance robustness while still guiding the model to focus on semantically meaningful dynamics. Finally, dataset curation and augmentation strategies that explicitly account for typical endoscopic artifacts may further narrow the performance gap observed in Table 6. We leave a systematic exploration of these directions for future research.

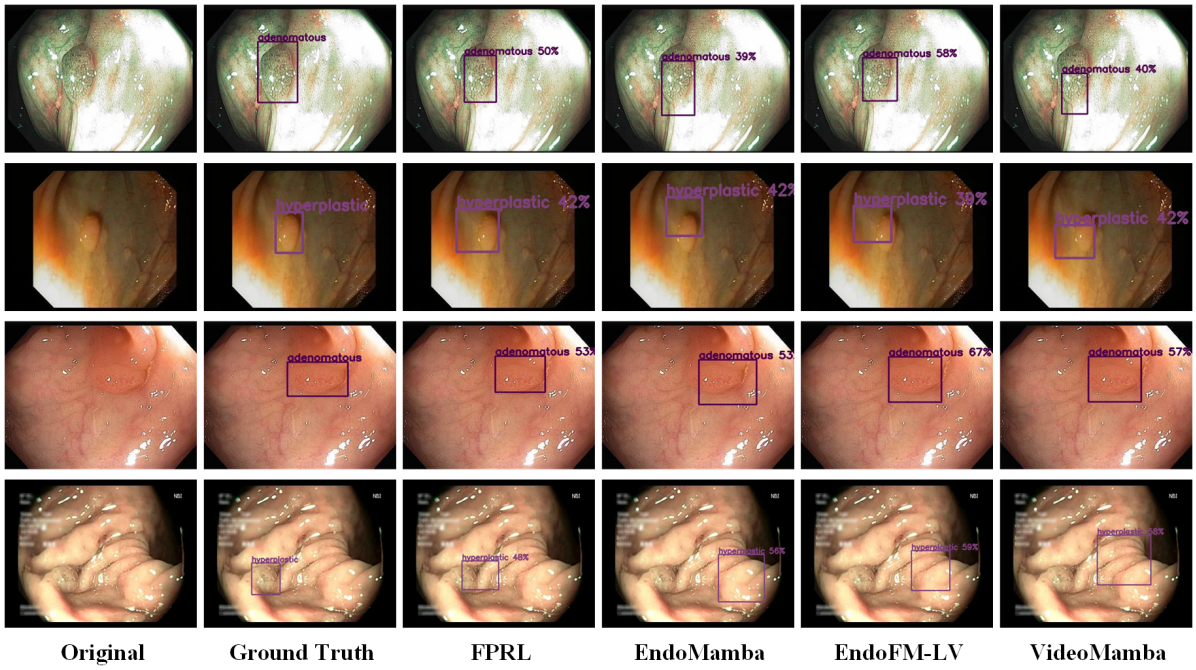

Figure 4. Qualitative results for detection task on the KUMC dataset.

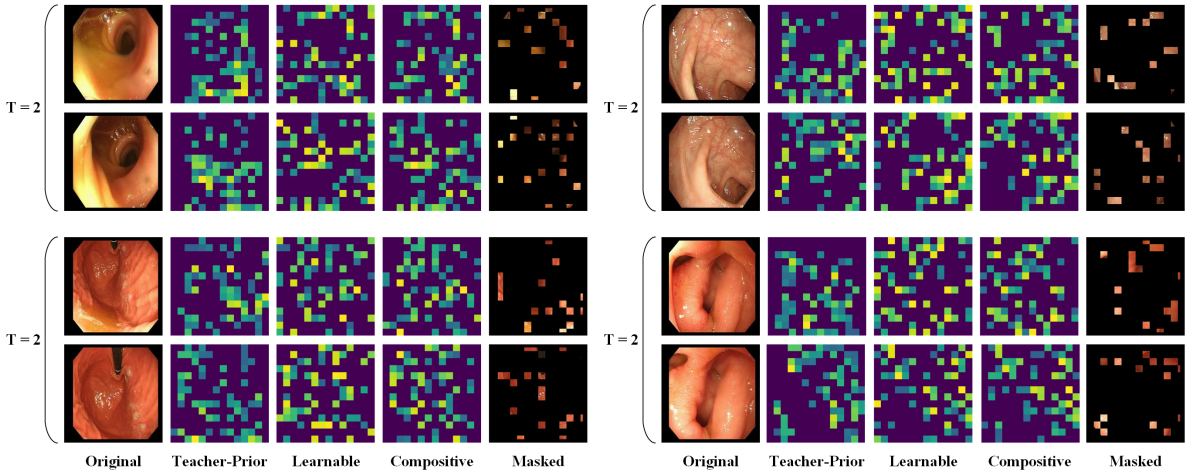

Figure 5. Illustration of our teacher-prior adaptive masking strategy. We visualize the feature heatmaps from both the teacher and the lightweight network (columns 2 & 3). The important regions (column 4) are then selected through a weighted aggregation of these heatmaps, where visible patches are sampled via a Top-K strategy.

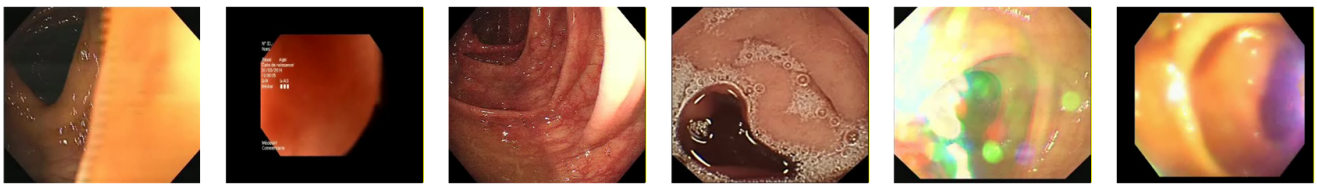

Figure 6. Examples of low-quality sampling frames.

References

[1] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015. 1, 2

[2] 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, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283, 2020. 1

[3] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. 2

[4] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9716–9726, 2022. 3

[5] Roy Hirsch, Mathilde Caron, Regev Cohen, Amir Livne, Ron Shapiro, Tomer Golany, Roman Goldenberg, Daniel Freedman, and Ehud Rivlin. Self-supervised learning for endoscopic video analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 569–578. Springer, 2023. 3

[6] Kai Hu, Ye Xiao, Yuan Zhang, and Xieping Gao. Multi-view masked contrastive representation learning for endoscopic video analysis. *Advances in Neural Information Processing Systems*, 37:47987–48014, 2024. 3

[7] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022. 1

[8] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Sv-rcnet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE Transactions on Medical Imaging*, 37 (5):1114–1126, 2017. 2

[9] Kaidong Li, Mohammad I Fathan, Krushi Patel, Tianxiao Zhang, Cuncong Zhong, Ajay Bansal, Amit Rastogi, Jean S Wang, and Guanghui Wang. Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. *PLOS One*, 16(8):e0255809, 2021. 1, 2

[10] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2024. 2, 3

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 2

[12] Yiting Ma, Xuejin Chen, Kai Cheng, Yang Li, and Bin Sun. Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 387–396. Springer, 2021. 1

[13] Pablo Mesejo, Daniel Pizarro, Armand Abergel, Olivier Rouquette, Sylvain Beorchia, Laurent Poincloux, and Adrien Bartoli. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE Transactions on Medical Imaging*, 35(9):2051–2063, 2016. 1

[14] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022. 1

[15] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 3

[16] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14711–14721, 2022. 3

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[18] Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Static and dynamic concepts for self-supervised video representation learning. In *European Conference on Computer Vision*, pages 145–164. Springer, 2022. 3

[19] Pia H Smedsrud, Vajira Thambawita, Steven A Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L Eskeland, et al. Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data*, 8(1):142, 2021. 1

[20] Qingyao Tian, Huai Liao, Xinyan Huang, Bingyu Yang, Dongdong Lei, Sebastien Ourselin, and Hongbin Liu. Endomamba: An efficient foundation model for endoscopic videos via hierarchical pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 224–234. Springer, 2025. 1, 2, 3

[21] Yu Tian, Guansong Pang, Fengbei Liu, Yuyuan Liu, Chong Wang, Yuanhong Chen, Johan Verjans, and Gustavo Carneiro. Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2022. 1, 2

[22] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35:10078–10093, 2022. 3

[23] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36 (1):86–97, 2016. 1, 2

[24] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 3

[25] Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023. 1, 2, 3

[26] Zhao Wang, Chang Liu, Lingting Zhu, Tongtong Wang, Shaoting Zhang, and Qi Dou. Improving foundation model for endoscopy video analysis via representation learning on long sequences. *IEEE Journal of Biomedical and Health Informatics*, 2025. 1, 3

[27] Lingyun Wu, Zhiqiang Hu, Yuanfeng Ji, Ping Luo, and Shaoting Zhang. Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 302–312. Springer, 2021. 2

[28] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14561–14571, 2023. 3