Quality-Controlled Keyframe Sampling using Endoscopic Foundation Models

Haiko Middeljans[1*], Tim J.M. Jaspers[1], Martijn R. Jong[2], Rixta A.H. van Eijck van Heslinga[2], Florance C. Slooter[2], Albert J. de Groof[2], Jacques J. Bergman[2], Peter H.N. De With[1], and Fons van der Sommen[1]

[1] Department of Electrical Engineering, Architectures for Reliable Image Analysis, Eindhoven University of Technology, Eindhoven, The Netherlands
h.middeljans@tue.nl

[2] Department of Gastroenterology and Hepatology, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands

Abstract. Data collection and annotation have been a longstanding bottleneck for computer vision applications, particularly in the medical domain, where data is limited by high annotation costs and low disease prevalence. Computer-aided detection and diagnosis (CADe/x) require substantial data for clinical relevance, yet current approaches rely on manual collection of high-quality images or post hoc keyframe selection from video, both time-consuming and tied to subjective quality assessment. To accelerate CADe development, we propose a quality-controlled keyframe selection algorithm that samples clinically relevant and diverse frames from video, leveraging endoscopic foundation model features and a computer-aided quality system to reject low-quality frames. We evaluate our approach by training downstream CADe models exclusively on the extracted frames for Barrett's neoplasia detection. Results show an improvement in Area Under the Receiver Operating Characteristic curve of up to 4.7% over uniform sampling, alongside an 80% reduction in data requirements at similar performance. The code is available at github.com/HaikoMid/GastroKey.

Keywords: Video summarization · Endoscopy · Data diversity · Foundation models

1 Introduction

Esophageal cancer remains a significant global health challenge. While Barrett's esophagus (BE) is the primary precursor to esophageal adenocarcinoma (EAC), the detection of Barrett's neoplasia still presents a significant diagnostic hurdle. Early-stage neoplastic lesions are often subtle and frequently prevent detection by non-expert endoscopists [5,23]. Given the high mortality rate of EAC, early detection and intervention are critical to improve patient prognosis [4].

* corresponding author

To support endoscopists, computer-aided detection (CADe) systems have emerged, largely integrated with deep learning techniques [2,12]. Despite their potential, data scarcity and inconsistent image quality remain critical barriers to clinical translation. Data scarcity is driven by the low prevalence of BE (1% to 2%) and the even lower risk of progression to high-grade dysplasia (HGD) or EAC [26]. Moreover, most CADe models are trained on curated, expert images, creating a distribution mismatch when deployed in real-time clinical settings, where non-expert practitioners encounter lower-quality data [3,11], often further degraded by motion blur or overexposure [8].

To mitigate these issues, researchers have turned to computer-aided quality (CAQ) systems [7,3,11,10] for quality control, data augmentation [16] for robustness, or video data to increase the dataset size in the medical domain [25,18]. However, video data introduces its own limitations, as frames are often temporally correlated and their individual quality is typically lower than that of static images. Recent advancements utilize deep learning-based methods to extract relevant keyframes from videos [14]. While the results show improvement in the performance of downstream vision-language models, it remains unclear how effectively these summarization techniques translate to the nuances of medical imaging, where both diversity and quality are essential.

This work introduces a quality-controlled keyframe sampling algorithm built on an endoscopic foundation model. While this method can be integrated into various data-curation pipelines, such as automatic frame selection for annotation, we focus here on its utility in model development. We validate our approach by training CADe models for Barrett's neoplasia detection on the sampled keyframes, demonstrating superior performance over models trained via uniform sampling. These downstream gains serve as a validation of selection quality rather than the algorithm's sole objective. Through a series of experiments, we systematically evaluate various foundation models and sampling strategies, improving Area Under the Receiver Operating Characteristic curve (AUC) performance and robustness by up to 4.7% and 3.6%, respectively, over uniform sampling, while reducing data storage by 80% at similar performance to uniform sampling at higher frame rates.

2 Methodology

2.1 Motivation

Generic video summarization techniques may not translate well to endoscopic imaging, where both semantic diversity and diagnostic quality must be preserved. Early approaches relied on low-level statistical characteristics [22] or depth information [21], which fail to capture the semantic complexity of endoscopic procedures [22]. We instead build on endoscopic foundation models whose latent representations have been shown to improve downstream CADe performance [1,9], making them a suitable basis for keyframe selection.

A second problem is specific to combining quality and diversity: low-quality frames are often semantically distinct from high-quality ones due to corrup-

tions such as motion blur or overexposure. Standard diversity-maximization algorithms are therefore prone to mistaking this visual difference for informativeness. This motivates filtering frame quality before diversity-based selection, rather than relying on diversity alone to discard poor-quality frames.

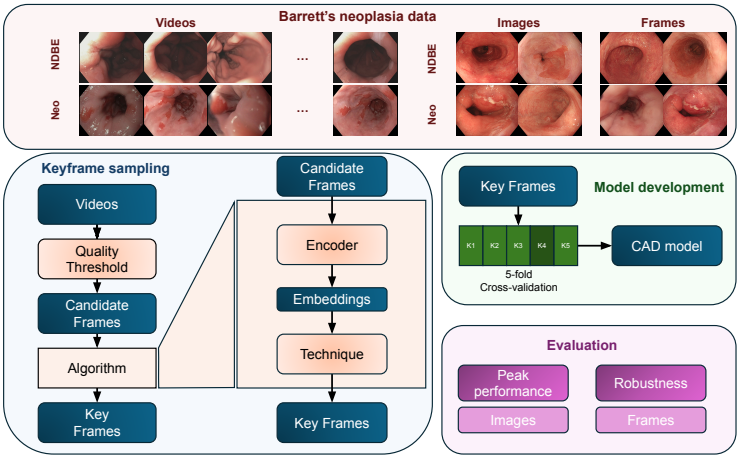


Fig. 1: Overview of the data and framework used for quality-controlled keyframe sampling, model development, and evaluation.

2.2 Proposed pipeline

To develop an algorithm that summarizes videos into relevant frames, we propose a two-stage pipeline. An overview of the pipeline is shown on the bottom left side of Fig. 1. In the first stage each frame, f_i, is processed using a CAQ model [7], which assigns a quality score $Q(f_i) \in [0, 5]$. Frames falling below a predefined quality threshold, ($T_{quality}$), are excluded from the candidate pool of frames to ensure the algorithm only considers clinically relevant frames. In the second stage, all candidate frames are passed through a pretrained foundation model encoder to extract latent representations that capture the semantic and diagnostic content of each frame, rather than relying on raw pixel-level similarity. These embeddings provide a compact, domain-informed representation in which visually similar frames lie close together in the latent space, while frames depicting distinct characteristics are further apart. Using these embeddings, we compare several sampling techniques to identify a subset of keyframes.

Quality algorithm To assess the impact of image quality on downstream performance, we implement a CAQ model [7] that effectively distinguishes between poor and good mucosal visualization. We explore the quality–diversity trade-off

by adjusting the quality threshold, since stricter thresholds boost frame quality at the expense of temporal diversity. Evaluating these variations allows us to determine the optimal balance for frame sampling in clinical settings.

Encoder selection Given the domain-specific properties of endoscopic images, various foundation model encoders are evaluated. These include domain-specific encoders developed using DINO frameworks with the GastroNet-5M [9] (GN-5M) dataset. Specifically, we consider a ResNet-50 (RN-50) and several Base Vision Transformers (ViT-B) developed using DINOv1 (Dv1), DINOv2 (Dv2) and DINOv3 (Dv3). We also include the DINOv3 baseline provided by Meta, as it has demonstrated strong transferability in the medical domain [15].

Keyframe selection technique We evaluate four latent keyframe sampling techniques.

Greedy cosine similarity. Cosine similarity has been used to remove redundant images in large-scale datasets [20]. Specifically, a greedy cosine algorithm is implemented that establishes the initial keyframe by selecting the first frame that exceeds a predefined quality threshold. Subsequent keyframes are iteratively identified using a diversity maximization strategy that calculates the similarity between all candidate frames with respect to the currently selected frames. Frame selection proceeds sequentially by adding frames that have the smallest absolute maximum similarity to the existing pool until the target frame count is satisfied.

Multi-level cosine similarity. To capture both fine-grained and semantic image characteristics, we extend the greedy cosine approach by concatenating hierarchical latent representations extracted from various network depths. The aforementioned greedy selection algorithm is then applied directly to these aggregated multi-scale feature vectors.

K-means clustering. Since clustering in high-dimensional spaces is not trivial [13,19], the features are first reduced using principal component analysis (PCA) to a two-dimensional space. In this reduced space, k-means clustering is performed, where the number of clusters is set to the requested amount of keyframes. The embeddings with the lowest Euclidean distance to the cluster centers are selected as key frames.

Max-volume sampling. For max-volume [17,24,6] sampling, we follow an implementation similar to the approach of Li *et al.* [14], but employ PCA in place of SVD.

2.3 Evaluation protocol

We evaluate our summarization method by extracting a fixed number of keyframes to train Barrett's neoplasia CADe models. For comparison, we train identical

models using standard uniformly sampled frames. We then benchmark both approaches by analyzing their peak performance and robustness using images and frames as illustrated in Fig. 1.

Sample rate To assess the influence of the sample rate on the performance of the CADe models, an ablation experiment is performed where the sample rate is changed from 0.25 to 0.1 and 0.5 frames per second (FPS), respectively. These experiments give insight into how the algorithms excel in video summarization with various data constraints.

3 Experimental Setup

This section introduces the dataset and implementation details used in this work for the downstream experiments.

3.1 Dataset

This study utilizes Barrett's esophagus videos and images categorized as non-dysplastic (NDBE) or neoplastic (Neo). Visual examples of the data are shown at the top of Fig. 1. The video dataset comprises 720 clips (431 NDBE from 58 patients; 289 neoplastic from 80 patients) with an average duration of 19.5 seconds (max 117.7s). A clinical research fellow trimmed the Neo clips to ensure continuous lesion visibility. All data were captured using Olympus HQ190 or EZ1500 gastroscopes paired with CV-190 or X1 processors (Olympus Corp., Tokyo, Japan). Extracted video frames were partitioned into five patient-level folds to prevent data leakage during CADe model training and validation, as conceptually illustrated in the top right of Fig. 1. To benchmark downstream performance, 2 holdout test sets consisting of images or frames are used. The frame test set serves as a robustness benchmark, since frames sampled from video clips introduce corruptions such as motion blur and compression artifacts. Detailed dataset quantities, patient counts, and video durations are summarized in Table 1.

3.2 Implementation Details

The experiments are carried out using an H100 GPU (NVIDIA Corp., CA, USA) equipped with 96 GB of VRAM. For sampling, the resolution is fixed to the native foundation model resolution. For the downstream task, the models are trained for a maximum of 10 epochs with an initial learning rate of 1×10^{-5}. This learning rate is reduced by a factor of 10 if no validation improvements are observed for 5 epochs. All downstream experiments apply standard data augmentation, including resized cropping, flipping, and rotation.

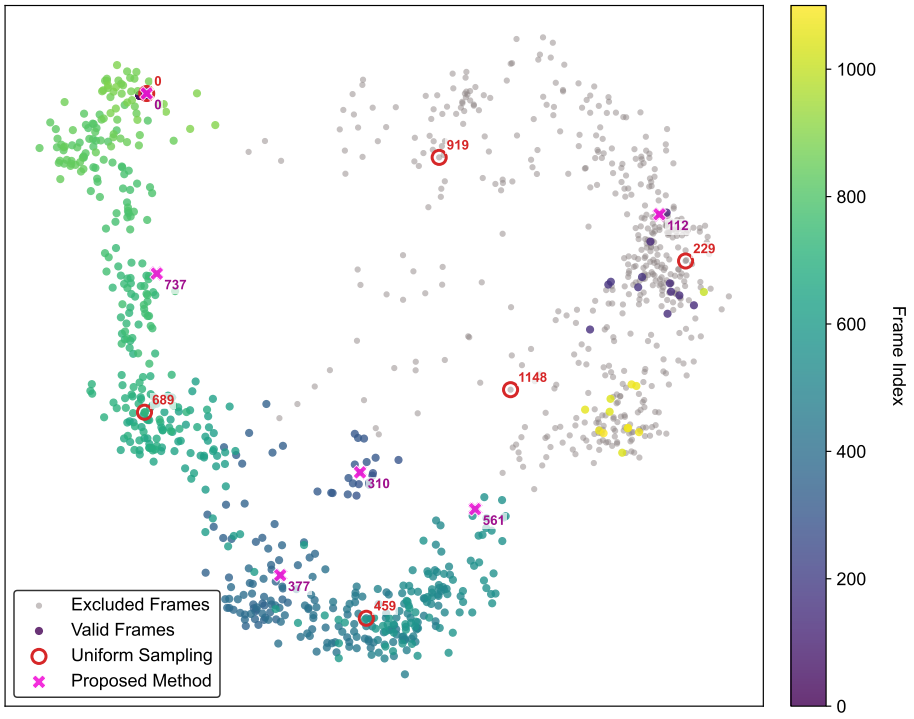
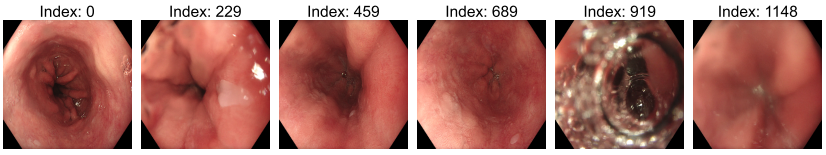
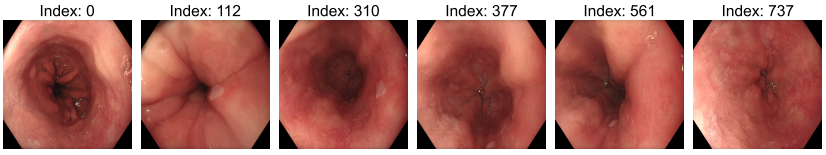


Fig. 2: A 2D latent visualization of the video frames. The excluded frames due to low quality are shown in gray. The selected indices of the frames are shown for both uniform sampling and the proposed method.



(a) Uniform sampling



(b) Proposed method

Fig. 3: Comparison of selected keyframes: (a) Uniform sampling and (b) the proposed method.

Table 1: Comprehensive dataset and video duration statistics for both non-dysplastic Barrett's esophagus (NDBE) and dysplastic Barrett's esophagus (Neo) patients.

Test Sets (Count / pt)			**Video Dataset**	
Data Type	**Neo**	**NDBE**	**Range (s)**	**Video Count**
Images	316 (42 pt)	432 (56 pt)	0 – 10	118
Frames	1018 (39 pt)	1624 (56 pt)	10 – 20	396
			20 – 30	111
			30 – 40	38
			40+	57

Table 2: Downstream evaluation of CADe models developed using sampled frames. Models are compared using the AUC (↑) across images and frames.

Encoder	**Quality Threshold**	**Technique**	**AUC** (↑)	
			Images	**Frames**
—	—	uniform	0.828 ± 0.014	0.789 ± 0.012
RN-50 (Dv1 GN-5M)	2	cosine similarity	0.831 ± 0.019	0.786 ± 0.022
ViT-B (Dv2 GN-5M)			0.825 ± 0.017	0.787 ± 0.023
ViT-B (Dv3)			0.837 ± 0.012	0.788 ± 0.016
ViT-B (Dv3 GN-5M)			0.841 ± 0.029	0.797 ± 0.027
ViT-B (Dv3 GN-5M)	3	cosine similarity	0.838 ± 0.018	0.794 ± 0.011
	2		0.841 ± 0.029	0.797 ± 0.027
	1		0.838 ± 0.014	0.782 ± 0.014
ViT-B (Dv3 GN-5M)	2	cosine similarity	0.841 ± 0.029	0.797 ± 0.027
		multi-level cosine	**0.843 ± 0.014**	**0.798 ± 0.021**
		kmeans centers	0.833 ± 0.017	0.789 ± 0.015
		max-volume	0.826 ± 0.021	0.785 ± 0.024

4 Results and Discussion

The experimental results are presented in Table 2. The findings indicate that the pre-trained DINOv3 model, when fine-tuned on GN-5M, achieves the highest downstream performance. Moreover, the model yields optimal results on images when applying a quality threshold of 2, which filters out poor-quality candidate frames. Although the performance gains on images are minimal with an increasing quality threshold, the improvements on frames are more pronounced, demonstrating increased robustness. The multi-level cosine sampling technique appears to be the most effective method to sample diverse frames from latent representations. Given that it achieves both superior peak performance and robustness, we select this algorithm as the proposed method for the remainder of this work.

Table 3: Data efficiency. Comparison of AUC (↑) performance across sampling rates on images and frames.

FPS	**Technique**	**AUC** (↑)	
		Images	**Frames**
0.10	uniform	0.794 ± 0.011	0.757 ± 0.015
	multi-level cosine	0.831 ± 0.020	0.784 ± 0.020
0.25	uniform	0.828 ± 0.014	0.789 ± 0.012
	multi-level cosine	**0.843 ± 0.014**	0.798 ± 0.021
0.50	uniform	0.832 ± 0.014	0.783 ± 0.022
	multi-level cosine	0.841 ± 0.014	**0.802 ± 0.016**

When compared to the baseline, this proposed method showed increases in peak performance and robustness AUC of 0.015 and 0.009, respectively.

A qualitative analysis of the proposed keyframe selection algorithm is presented in Fig. 2. Video frames are projected into a 2D representation using PCA based on the foundation model extracted embeddings. The color gradient denotes the temporal progression of the video, revealing the underlying semantic trajectory. Frames failing to meet the quality threshold are marked in gray and excluded from the selection process. Selected keyframes are highlighted with distinct markers, comparing the temporal coverage of standard uniform sampling against the proposed method. The visual appearance of the corresponding keyframes is detailed in Fig. 3. While the proposed method identifies diverse, clinically relevant keyframes that span the temporal progression of the procedure, uniform sampling fails to account for image degradation, often selecting frames compromised by overexposure, blur, or obscuring artifacts.

To evaluate the summarization effectiveness of our method, Table 3 summarizes performance across varying sampling rates (0.1, 0.25, and 0.5 FPS). As expected, uniform sampling demonstrates a strong dependency on the sampling frequency, with increased performance at higher rates and significant degradation at lower rates due to reduced video coverage. Conversely, our proposed algorithm maintains high downstream performance, even at the lowest sampling rate of 0.1 FPS, outperforming the uniform baseline with an AUC of 0.037 on images. Moreover, we show that CADe models trained on our method's keyframes at 0.1 FPS match the performance of a model trained on uniformly sampled frames at 0.5 FPS, an 80% reduction in data requirements. From the perspective of video summarization, this stability is highly desirable and confirms the summarization performance, showing that aggressive temporal compression can be achieved without losing valuable information for downstream tasks.

5 Conclusions and Future Work

This work proposes a keyframe sampling algorithm for endoscopic video summarization. Quantitative experiments demonstrate the effectiveness of the sampling algorithm, showing superior downstream performance over uniform sampling. These findings suggest that leveraging latent features from endoscopic foundation models, paired with quality-informed sampling algorithms, provides a more representative and diagnostically rich keyframe selection than traditional uniform sampling. By producing a versatile and information-rich set of keyframes, this sampling framework can serve as a preprocessing pipeline across multiple distinct endoscopic vision tasks. For future work, the algorithm could be adapted to automatically select a proper amount of keyframes from a video based on the quality and diversity per video without input from the user. Furthermore, evaluation of the method on extended endoscopic procedures and other environments, such as surgical scenes, will be essential to validate its broader clinical benefit.

Acknowledgments. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-18266. This publication is part of the project "You won't find what you don't image: Exposing blind spots in endoscopic cancer screening" (project number: 19091) of the research program Talentprogramma Veni-TTW which is financed by the Dutch Research Council (NWO).

References

1. Boers, T.G., Fockens, K.N., van der Putten, J.A., Jaspers, T.J., Kusters, C.H., Jukema, J.B., Jong, M.R., Struyvenberg, M.R., de Groof, J., Bergman, J.J., et al.: Foundation models in gastrointestinal endoscopic ai: Impact of architecture, pretraining approach and data efficiency. Medical Image Analysis **98**, 103298 (2024)
2. Bretthauer, M., Ahmed, J., Antonelli, G., Beaumont, H., Beg, S., Benson, A., Bisschops, R., De Cristofaro, E., Gibbons, E., Häfner, M., et al.: Use of computer-assisted detection (cade) colonoscopy in colorectal cancer screening and surveillance: European society of gastrointestinal endoscopy (esge) position statement. Endoscopy **57**(06), 667–673 (2025)
3. De Groof, A.J.: Artificial intelligence (ai) systems for detection of barrett's neoplasia: time to bridge domain gaps and explore human–ai interaction (2024)
4. Desai, M., Lieberman, D., Srinivasan, S., Nutalapati, V., Challa, A., Kalgotra, P., Sundaram, S., Repici, A., Hassan, C., Kaminski, M.F., et al.: Post-endoscopy barrett's neoplasia after a negative index endoscopy: a systematic review and proposal for definitions and performance measures in endoscopy. Endoscopy **54**(09), 881–889 (2022)
5. Frazzoni, L., Arribas, J., Antonelli, G., Libanio, D., Ebigbo, A., Van Der Sommen, F., De Groof, A.J., Fukuda, H., Ohmori, M., Ishihara, R., et al.: Endoscopists' diagnostic accuracy in detecting upper gastrointestinal neoplasia in the framework of artificial intelligence studies. Endoscopy **54**(04), 403–411 (2022)
6. Goreinov, S.A., Oseledets, I.V., Savostyanov, D.V., Tyrtyshnikov, E.E., Zamarashkin, N.L.: How to find a good submatrix. In: Matrix Methods: Theory, Algorithms And Applications: Dedicated to the Memory of Gene Golub, pp. 247–256. World Scientific (2010)

7. Jaspers, T.J.M., Jong, M.R., Kusters, C.H.J., Boers, T.G.W., van Heslinga, R.A.H.V.E., Jukema, J.B., de Groof, A.J., Bergman, J.J., With, P.H.N.D., van der Sommen, F.: Self-supervised learning for automated image quality assessment in endoscopy. In: Colliot, O., Mitra, J. (eds.) Medical Imaging 2025: Image Processing. vol. 13406, p. 134061A. International Society for Optics and Photonics, SPIE (2025). https://doi.org/10.1117/12.3045901, https://doi.org/10.1117/12.3045901
8. Jaspers, T.J., Boers, T.G., Kusters, C.H., Jong, M.R., Jukema, J.B., De Groof, A.J., Bergman, J.J., de With, P.H., van der Sommen, F.: Robustness evaluation of deep neural networks for endoscopic image analysis: Insights and strategies. Medical Image Analysis **94**, 103157 (2024)
9. Jong, M.R., Boers, T.G., Fockens, K.N., Jukema, J.B., Kusters, C.H., Jaspers, T.J., van Heslinga, R.v.E., Slooter, F.C., Struyvenberg, M.R., Bisschops, R., et al.: Gastronet-5m: A multicenter dataset for developing foundation models in gastrointestinal endoscopy. Gastroenterology (2025)
10. Jong, M.R., van Heslinga, R.A.v.E., Kusters, C.H., Jaspers, T.J., Boers, T.G., Duits, L.C., Pouw, R.E., Weusten, B.L., Alkhalaf, A., van der Sommen, F., et al.: Evaluation of an improved computer-aided detection system for barrett's neoplasia in real-world imaging conditions. Endoscopy **57**(12), 1327–1337 (2025)
11. Jong, M.R., Jaspers, T.J., van Heslinga, R.A.v.E., Jukema, J.B., Kusters, C.H., Boers, T.G., Pouw, R.E., Duits, L.C., van der Sommen, F., de Groof, A.J., et al.: The development and ex vivo evaluation of a computer-aided quality control system for barrett's esophagus endoscopy. Endoscopy **57**(07), 709–716 (2025)
12. Kusters, C.H., Boers, T.G., Jaspers, T.J., Jong, M.R., van Heslinga, R.A.v.E., Jukema, J.B., Fockens, K.N., de Groof, A.J., Bergman, J.J., van der Sommen, F., et al.: Designing a computer-aided detection system for barrett's neoplasia: Insights in architectural choices, training strategies and inference approaches. Computer Methods and Programs in Biomedicine **269**, 108891 (2025)
13. Lederman, R.R., Silva-Sánchez, D., Chen, Z., Mordant, G., Balanov, A., Bendory, T.: The catastrophic failure of the k-means algorithm in high dimensions, and how hartigan's algorithm avoids it (2026), https://arxiv.org/abs/2602.09936
14. Li, P., Abdullaeva, I., Gambashidze, A., Kuznetsov, A., Oseledets, I.: Maxinfo: A training-free key-frame selection method using maximum volume for enhanced video understanding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7198–7207 (2026)
15. Li, Y., Wu, Y., Lai, Y., Hu, M., Yang, X.: Meddinov3: How to adapt vision foundation models for medical image segmentation? arXiv preprint arXiv:2509.02379 (2025)
16. Middeljans, H., Kusters, C.H., Jaspers, T.J., Jong, M.R., Van Eijck Van Heslinga, R.A., Slooter, F.C., De Groof, A.J., Bergman, J.J., Van Der Sommen, F., et al.: Wavedamp: Enhancing natural robustness in endoscopy through wavelet-based frequency damping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 604–614 (2025)
17. Mikhalev, A., Oseledets, I.V.: Rectangular maximum-volume submatrices and their applications. Linear Algebra and its Applications **538**, 187–211 (2018)
18. Mobadersany, P., Parmar, C., Damasceno, P.F., Fadnavis, S., Chaitanya, K., Li, S., Schwab, E., Xiao, J., Surace, L., Mansi, T., et al.: Harnessing temporal information for precise frame-level predictions in endoscopy videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 295–305. Springer (2024)

19. Molchanov, V., Linsen, L.: Overcoming the curse of dimensionality when clustering multivariate volume data. In: VISIGRAPP (3: IVAPP). pp. 29–39 (2018)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
21. Sasmal, P., Paul, A., Bhuyan, M.K., Iwahori, Y., Kasugai, K.: Extraction of keyframes from endoscopic videos by using depth information. IEEE Access **9**, 153004–153011 (2021)
22. Schoeffmann, K., Del Fabro, M., Szkaliczki, T., Böszörmenyi, L., Keckstein, J.: Keyframe extraction in endoscopic video. Multimedia Tools and Applications **74**(24), 11187–11206 (2015)
23. Schölvinck, D.W., Van Der Meulen, K., Bergman, J.J., Weusten, B.L.: Detection of lesions in dysplastic barrett's esophagus by community and expert endoscopists. Endoscopy **49**(02), 113–120 (2017)
24. Sozykin, K., Chertkov, A., Schutski, R., Phan, A.H., Cichocki, A.S., Oseledets, I.: Ttopt: A maximum volume quantized tensor train-based optimization and its application to reinforcement learning. Advances in neural information processing systems **35**, 26052–26065 (2022)
25. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International conference on medical image computing and computer-assisted intervention. pp. 101–111. Springer (2023)
26. Weusten, B.L., Bisschops, R., Dinis-Ribeiro, M., Di Pietro, M., Pech, O., Spaander, M.C., Baldaque-Silva, F., Barret, M., Coron, E., Fernández-Esparrach, G., et al.: Diagnosis and management of barrett esophagus: European society of gastrointestinal endoscopy (esge) guideline. Endoscopy **55**(12), 1124–1146 (2023)