Compress to Scale: Semantic Token Reduction for High-Resolution Retinal Image Analysis

Bryan Lopez de Munain[1 [0009−0006−9152−2462]], Steven Kerr[1 [0000−0002−3643−7859]], Enrico Pellegrini[2 [0000−0001−6913−1760]], and Miguel O. Bernabeu[1 [0000−0002−6456−3756]]

[1] Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, UK
B.Lopez-De-Munain@sms.ed.ac.uk
[2] Optos plc, Dunfermline, UK

Abstract. Ultra-widefield (UWF) retinal imaging captures up to 200° of the retina and contains diagnostically important peripheral and posterior pole features that are often lost when images are aggressively downsampled for vision transformer (ViT) processing. While increasing image resolution could preserve these fine-grained pathological signals, the quadratic dependence of self-attention on token count makes high-resolution training and inference computationally prohibitive. Motivated by the observation that token importance is highly uneven in UWF retinal images, we propose a compute-constrained scaling strategy that reallocates computation from redundant image tokens to increased image resolution through semantic token reduction, enabling higher-resolution UWF image analysis under a fixed compute budget.

We instantiate this strategy using four token reduction methods—PiToMe, DiffRate, EViT, and DynamicViT—within a DINOv2 ViT-S/14 backbone constrained to approximately 45 GFLOPs. Across two retinal imaging datasets, compute-matched token reduction generally improves performance over the uncompressed 392×392 baseline, with the magnitude and statistical significance of gains varying across datasets and disease classes. On MMRDR-UWF, per-class AUC gains range from +0.016 to +0.033, with the largest improvements in disease stages characterised by fine lesions such as microaneurysms. On TOP, per-class AUC gains reach +0.022.

These results indicate that reallocating compute from redundant tokens to increased image resolution can improve performance across multiple token reduction strategies, with larger and more statistically supported gains on MMRDR-UWF than on TOP. Our findings suggest that semantic token compression provides a practical pathway for scaling UWF image resolution without increasing computational cost and that substantial resolution-driven gains remain available for retinal disease classification, particularly for diseases characterised by fine pathological features.

Keywords: Ultra-widefield Retinal Imaging · Vision Transformers · Token Compression.

1 Introduction

Retinal imaging is increasingly used for automated disease detection because it provides a non-invasive view of retinal structures and pathology [5, 16]. Ultrawidefield (UWF) imaging extends the field of view to up to 200° of the retina, capturing peripheral regions that are not visible in standard colour fundus imaging and that may contain diagnostically relevant pathology. Modern UWF systems can acquire images at resolutions approaching 4000×4000 pixels, enabling the detection of fine-grained features such as microaneurysms, drusen deposits, and subtle vascular abnormalities.

Preserving these features requires substantially higher image resolutions than are typically used in deep learning pipelines. This challenge is particularly acute for Vision Transformers (ViTs), whose self-attention mechanism scales quadratically with token count [3]. As a result, UWF images are commonly downsampled aggressively before analysis [2, 6], raising the question of how much performance is limited by the fine-grained information discarded during downsampling.

Recent token reduction methods [7] offer a potential solution. By identifying and removing or merging redundant tokens during inference, these methods reduce the computational burden of self-attention while preserving informative image content. This creates the possibility of reallocating computation from redundant tokens to increased image resolution, allowing models to process finer image detail without increasing overall computational cost.

Recent work has also explored efficiency from complementary perspectives in retinal imaging. Engelmann and Bernabeu developed RETFound-Green [5], demonstrating that a retinal foundation model can be trained and used downstream with substantially reduced data and compute, while retaining strong performance across diverse retinal tasks. Playout et al. [11] proposed a region-based approach for diabetic retinopathy classification that uses SLIC superpixels to construct region-level tokens, semantic prototype matching, and graph-based token clustering, with the goal of enabling higher-resolution Transformer inputs while maintaining a compact token representation. In contrast, our study systematically evaluates existing token-reduction methods under a matched compute budget to quantify how computation saved through token reduction can be exchanged for increased input resolution.

In this work, we investigate whether semantic token reduction can be used to recover the benefits of higher-resolution UWF imaging under a fixed computational budget. We evaluate four representative token reduction methods—PiToMe [15], DiffRate [1], EViT [9], and DynamicViT [12]—across two large UWF retinal imaging datasets. By matching computational cost while varying image resolution, we quantify the extent to which semantic token reduction can unlock resolution-driven performance gains in retinal disease classification.

We make the following **contributions**:

1. We establish a compute-constrained empirical framework for studying the trade-off between image resolution and semantic token reduction in ultra-widefield retinal imaging.

2. We demonstrate that substantial classification gains remain available at resolutions far above those commonly used in UWF retinal image analysis, with resolution requirements varying considerably across disease classes.
3. We show that multiple token reduction methods, including PiToMe, DiffRate, EViT, and DynamicViT, can recover a substantial fraction of the performance gains obtained from unconstrained high-resolution inference while operating under the compute budget of a standard 392px ViT.

2 Methodology

2.1 Datasets

Training and evaluation was performed over the MMRDR-UWF [14] and TOP [8] datasets. MMRDR-UWF is a diabetic retinopathy dataset with UWF images across 5 disease classes: 32.4% with No DR, 18.6% with Mild NPDR, 20.8% with Moderate NPDR, 13.2% with Severe NPDR, and 14.9% with Proliferative DR. TOP is a multi-class UWF image dataset spanning 9 disease classes: 37.5% Normal, 25.5% with DR, 20.1% with Glaucoma, 7.5% with Retinal Detachments, 6.0% with Retinal Vein Occlusions, 3.2% with Age-related Macular Degeneration, 2.0% with Retinitis Pigmentosa, 1.7% with Macular Holes, and 0.2% with Artery Occlusions. Retinitis Pigmentosa (RP), Macular Hole (MH), and Artery Occlusion (AO) were excluded because the test split contains insufficient positive samples for reliable ROC-AUC estimation.

2.2 Token value probing

To investigate the spatial distribution of discriminative information within ultra-widefield retinal images, we performed gradient-weighted class activation mapping (GradCAM [13]) using a DINOv2 ViT-S/14 backbone (vit_small_patch14_reg4_dinov2) at an input resolution of 392×392 pixels. A linear classification head was trained on frozen DINOv2 features for each dataset using cross-entropy loss on the predefined validation splits.

GradCAM scores were computed from the output activations of the final transformer block by backpropagating the predicted class logit and obtaining a scalar importance value for each image patch. To characterise the overall distribution of token importance, patch-level GradCAM scores were aggregated across all images in each dataset and their empirical distributions were computed.

2.3 Semantic token compression

While there is some variation in design choices between methods, every token selection method discussed in this work follows the same basic structure. Beginning with an image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$, where H and W denote the image height and width and C the number of channels, we patchify the image into patches $\boldsymbol{x_p} \in \mathbb{R}^{N\times(P^2\cdot C)}$, where P is the patch size and $N = (H/P)(W/P)$ is the number

of image patches. Each patch $\boldsymbol{x_p}$ corresponds to a token $\boldsymbol{z_p^l}$, where $p \in \{1, \ldots, N\}$ indexes image patches and l indexes transformer blocks.

At some layer l, we collect tokens $\{\boldsymbol{z_p^l}\}_{p=1}^N$ (or their corresponding key vectors) to build a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertices $\mathcal{V}$ contain one token each such that $\mathcal{V}_p = \boldsymbol{z_p^l}$, and the edges $\mathcal{E}$ are constructed according to a token similarity measure. The choice of similarity measure varies by method, but common choices include cosine similarity and attention scores at a given layer.

The task of semantic token compression can then be reframed as one of graph coarsening over a similarity measure. All methods attempt to either prune tokens that are highly similar to many others or merge highly similar tokens in order to minimise the computational burden introduced by processing many redundant copies of highly similar tokens.

2.4 Token reduction methods

We evaluate four token reduction methods that share the goal of reducing tokens in later transformer layers but differ substantially in how they estimate token importance and redundancy. **PiToMe** builds a similarity graph from cosine similarity between key vectors, flagging tokens similar to many others as redundant; a protected subset of informative tokens is kept while the rest are *merged* after selected attention layers. **DiffRate** learns an input-adaptive compression schedule, progressively *pruning* tokens estimated to contribute little to the model's predictions. **EViT** ranks tokens by the attention they receive from the CLS token, retaining highly attended tokens and *fusing* the rest into representative tokens. **DynamicViT** uses lightweight prediction modules to *prune* uninformative tokens via attention masking, with pruning decisions learned jointly with the classification objective. These methods thus span both merging- and pruning-based strategies and rely on distinct importance signals—graph energy, learned schedules, CLS attention, and predicted keep-probabilities. What they share is the exploitation of token redundancy, making them well suited to trading redundant tokens for higher image resolution under a fixed compute budget.

2.5 Resolution sweep implementation details

To determine whether semantic token reduction can raise image resolution at a fixed computational budget, we compare models under matched FLOP constraints rather than matched input resolutions. Features are extracted with the official DINOv2 [10] ViT-S/14 model with four register tokens (`vit_small_patch14_reg4_dinov2`), frozen throughout. The unconstrained sweep varies resolution from 224px to 1064px in patch-size (14px) increments, giving 61 padding-free evaluation points. Compute-constrained experiments span 392px to 826px in 14px steps, the lower bound being the baseline operating point and the upper the point beyond which the required compression exceeds the practical limits of the methods considered. All compute-constrained runs are matched to the cost of an uncompressed 392px ViT-S/14 model, a target of 44.98 GFLOPs. Following

prior work, each transformer block's cost is approximated as

$$\text{FLOPs} = 24ND^2 + 4N^2D, \tag{1}$$

where the $24ND^2$ term captures the QKVO projections ($8ND^2$) together with the position-wise feed-forward network ($16ND^2$), and $4N^2D$ accounts for the attention map and value aggregation. N is the token count and D the embedding dimension. Total FLOPs sum this across all blocks while accounting for the token count induced by reduction. The CLS and four register tokens are never merged or pruned. At each resolution, a binary search finds the largest token retention ratio whose realised cost stays within budget, applied independently to each method. To isolate resolution from method-specific scheduling, a single uniform keep ratio is used across all layers rather than learned or layer-specific schedules. Experiments use the predefined benchmark splits of each dataset: for both TOP and MMRDR-UWF, the labelled validation split trains the classifier and the held-out test split is used for evaluation. Performance is reported as one-vs-rest AUC per disease class and macro-AUC across classes. Confidence intervals use 2,000 stratified bootstrap replicates [4] of the test set, taken as the empirical 2.5th and 97.5th percentiles. Macro-AUC is the unweighted mean of per-class AUCs, its interval obtained by averaging per-class AUCs within each replicate. Reported ΔAUC margins are paired bootstrap estimates—identical resampled indices applied to both models before differencing—and significance is assessed with paired bootstrap tests on the resulting ΔAUC distributions.

3 Results and Analysis

3.1 UWF images contain many low-value tokens

To investigate the spatial distribution of discriminative information in ultra-widefield retinal images, we analysed GradCAM activations from a multi-label classifier trained on the TOP dataset and a five-class diabetic retinopathy classifier trained on the MMRDR-UWF dataset (Fig. 1). The resulting activation distributions reveal a highly non-uniform pattern, with the majority of image regions contributing relatively little to the model prediction and a smaller proportion of regions exhibiting substantially higher activation. This indicates that the performance improvements observed with increasing resolution are unlikely to arise from uniformly distributed additional detail, but rather from improved representation of specific spatial regions containing clinically relevant features.

This heterogeneity in information density motivates adaptive token reduction, where computation can be prioritised towards informative regions while redundancy in less discriminative areas is reduced.

3.2 Disease Classes Exhibit Distinct Resolution Requirements

Table 1 summarises the unconstrained resolution sweep described in Section 2.5, in which frozen DINOv2 features were evaluated across input resolutions

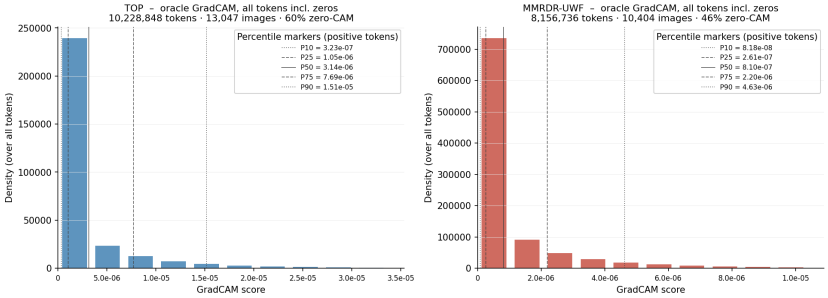

Fig. 1. Histograms of GradCAM scores of every patch in both the TOP and MMRDR-UWF datasets.

from 224 px to 1064 px without token reduction on the TOP multi-label disease classification and MMRDR-UWF five-class diabetic retinopathy grading tasks. Increasing image resolution yielded substantial but highly class-dependent improvements in classification performance. The optimal resolution varied considerably between disease classes, ranging from 420px for MMRDR-UWF Grade 4 to 1064px for glaucoma and Grade 0 classification. Several classes continued to benefit from resolution increases far beyond commonly used input sizes, with gains of up to +0.058 AUC for AMD, +0.047 AUC for Grade 1 diabetic retinopathy, and +0.032 AUC for Grade 3 diabetic retinopathy. Notably, the classes exhibiting the largest improvements typically achieved their peak performance at resolutions above 700px, indicating that diagnostically relevant information remains inaccessible at standard operating resolutions. This variation implies that a single deployment resolution may not be uniformly optimal across tasks; in practice, resolution selection may therefore need to reflect the target disease or application. Dynamically routing images to different resolutions is a natural extension, but is outside the scope of the present controlled study.

3.3 PiToMe Recovers A Substantial Fraction of the Performance Gains Available from High-Resolution Imaging

Table 2 evaluates PiToMe under the fixed compute budget of approximately 45 GFLOPs introduced in Section 2.5, comparing compute-constrained token reduction against the 392 px baseline on the TOP multi-label disease classification and MMRDR-UWF diabetic retinopathy grading tasks. Despite processing only a fraction of the token set, PiToMe recovered a substantial share of the improvement available from full-token high-resolution inference: 22.8% of the macro-AUC gain on TOP and 54.4% on MMRDR-UWF. The recovered fraction was strongly class-dependent and tracked how far each class's optimal resolution lay above the baseline. Classes whose peak performance was reached only moderately above 392px were captured almost entirely—RVO, for instance, matched and marginally exceeded its no-merge gain (110.5%) at 462px—whereas classes demanding very high resolution were captured only in part, such as AMD at

Table 1. Best-performing full-resolution sweep without token reduction compared with the 392px baseline. ΔAUC is reported as mean ± 1.96SE from paired bootstrap estimates.

Class	Baseline	Best AUC	Res. (px)	ΔAUC	Prev.	p
TOP						
AMD	0.807	0.865	868	$+0.058 \pm 0.036$	3.5%	0.003
RVO	0.827	0.846	476	$+0.019 \pm 0.020$	5.9%	0.069
Gla	0.862	0.882	1064	$+0.021 \pm 0.015$	19.2%	0.008
DR	0.860	0.882	966	$+0.022 \pm 0.013$	25.1%	0.002
RD	0.932	0.946	588	$+0.014 \pm 0.013$	7.6%	0.025
DM	0.854	0.867	854	$+0.014 \pm 0.012$	29.3%	0.026
Macro	0.857	0.881	–	$+0.025 \pm 0.009$	–	< 0.001
MMRDR-UWF						
Grade 0	0.878	0.898	1064	$+0.020 \pm 0.012$	32.4%	0.001
Grade 1	0.700	0.748	980	$+0.047 \pm 0.026$	18.9%	< 0.001
Grade 2	0.768	0.787	700	$+0.019 \pm 0.021$	20.0%	0.058
Grade 3	0.865	0.897	714	$+0.032 \pm 0.016$	14.4%	< 0.001
Grade 4	0.947	0.953	420	$+0.006 \pm 0.007$	14.4%	0.102
Macro	0.831	0.856	–	$+0.025 \pm 0.009$	–	< 0.001

37.9% using 434px inputs against an 868px unconstrained optimum. Grade 4 diabetic retinopathy, whose no-merge gain was itself not significant, showed no compute-constrained improvement. These results indicate that semantic token reduction captures a meaningful and frequently majority portion of the benefits associated with high-resolution retinal imaging without incurring prohibitive computational cost.

3.4 Resolution-Driven Performance Gains Generalise Across Token Reduction Methods

To test whether the observed gains stem from increased image resolution rather than a specific token reduction algorithm, we repeated the fixed-compute (45 GFLOPs) experiments of Section 2.5 on the TOP and MMRDR-UWF datasets, using four representative methods: PiToMe, DiffRate, EViT, and DynamicViT. Each is independently calibrated to the same 44.98 GFLOP budget via binary search over its token-retention parameter, and reported at its best-performing resolution.

Results appear in Table 3. All four methods yield positive macro-AUC differences relative to the 392 px baseline, but the evidence is stronger on MMRDR-UWF than on TOP. On TOP, gains range from +0.0057 (PiToMe) to +0.0098 (DynamicViT); DiffRate, EViT, and DynamicViT reach statistical significance, whereas PiToMe shows a positive but non-significant trend. On MMRDR-UWF, all four methods produce statistically significant gains.

Table 2. PiToMe under a fixed compute budget of approximately 45 GFLOPs compared with the 392px baseline. Results are reported at the best-performing resolution for each class. %Max denotes the fraction of the no-merge resolution improvement recovered. ΔAUC is reported as mean ± 1.96SE from paired bootstrap estimates.

Class	Baseline	PiToMe	Res. (px)	ΔAUC	%Max	Prev.	p
TOP							
AMD	0.807	0.830	434	$+0.022 \pm 0.035$	37.9%	3.5%	0.198
RVO	0.827	0.847	462	$+0.021 \pm 0.027$	110.5%	5.9%	0.115
Gla	0.862	0.871	560	$+0.009 \pm 0.015$	42.9%	19.2%	0.224
DR	0.860	0.876	462	$+0.016 \pm 0.011$	72.7%	25.1%	0.009
RD	0.932	0.936	420	$+0.004 \pm 0.012$	28.6%	7.6%	0.527
DM	0.854	0.859	420	$+0.006 \pm 0.009$	42.9%	29.3%	0.230
MMRDR-UWF							
Grade 0	0.878	0.894	462	$+0.016 \pm 0.010$	80.0%	32.4%	0.001
Grade 1	0.700	0.733	504	$+0.033 \pm 0.022$	70.2%	18.9%	0.006
Grade 2	0.768	0.784	448	$+0.016 \pm 0.020$	84.2%	20.0%	0.117
Grade 3	0.865	0.887	448	$+0.022 \pm 0.015$	68.8%	14.4%	0.003
Grade 4	0.947	0.947	462	$+0.000 \pm 0.011$	0.0%	14.4%	0.913

4 Conclusion

We investigated whether semantic token reduction can be used to unlock the benefits of higher-resolution ultra-widefield retinal imaging without incurring the prohibitive computational costs associated with full-token vision transformers. Across two UWF retinal datasets, we found that disease classes exhibit markedly different resolution requirements and that several classes continue to realise substantial performance gains at resolutions well beyond those typically used in retinal image analysis. These findings suggest that diagnostically relevant information is often lost when images are aggressively downsampled to satisfy computational constraints.

We further showed that semantic token reduction enables models to recover a substantial fraction of the performance improvements obtainable from unconstrained high-resolution inference while operating under fixed computational budgets. Importantly, this behaviour was observed across multiple token reduction methods, indicating that the effect is not specific to a single algorithm but instead reflects a broader opportunity to trade redundant tokens for increased image resolution.

Taken together, our results demonstrate that image resolution remains an underexploited axis for improving retinal disease classification and that semantic token reduction offers a practical pathway for accessing this information efficiently. Future work should investigate whether these findings extend to larger retinal datasets, additional ophthalmic imaging modalities, and foundation models trained at substantially higher native resolutions.

Table 3. Comparison of token-reduction methods under a fixed compute budget of approximately 45 GFLOPs. Results are reported at the best-performing resolution for each method. %Max denotes the fraction of the no-merge resolution improvement recovered. ΔAUC is reported as mean ± 1.96SE from paired bootstrap estimates.

Method	Macro AUC	Res. (px)	ΔAUC	%Max	p
TOP					
PiToMe	0.8626	406	$+0.0057 \pm 0.0077$	22.8%	0.145
DiffRate	0.8651	406	$+0.0083 \pm 0.0078$	33.2%	0.039
EViT	0.8664	476	$+0.0096 \pm 0.0088$	38.4%	0.033
DynamicViT	0.8667	476	$+0.0098 \pm 0.0088$	39.2%	0.034
MMRDR-UWF					
PiToMe	0.8451	462	$+0.0136 \pm 0.0079$	54.4%	< 0.001
DiffRate	0.8415	462	$+0.0100 \pm 0.0082$	40.0%	0.012
EViT	0.8447	462	$+0.0132 \pm 0.0080$	52.8%	< 0.001
DynamicViT	0.8419	462	$+0.0104 \pm 0.0080$	41.6%	0.009

Acknowledgments. B.L.d.M. is supported by a PhD studentship funded by Optos plc.

Disclosure of Interests. B.L.d.M. holds a student internship position with Optos plc and is supported by a PhD studentship funded by Optos plc. E.P. is an employee of Optos plc. The remaining authors declare no competing interests.

References

1. Chen, M., Shao, W., Xu, P., Lin, M., Zhang, K., Chao, F., Ji, R., Qiao, Y., Luo, P.: Diffrate: Differentiable compression rate for efficient vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17164–17174 (October 2023)
2. Cui, T., Lin, D., Yu, S., Zhao, X., Lin, Z., Zhao, L., Xu, F., Yun, D., Pang, J., Li, R., Xie, L., Zhu, P., Huang, Y., Huang, H., Hu, C., Huang, W., Liang, X., Lin, H.: Deep learning performance of ultra-widefield fundus imaging for screening retinal lesions in rural locales. JAMA Ophthalmology **141**(11), 1045–1051 (2023). https://doi.org/10.1001/jamaophthalmol.2023.4650
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
4. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall/CRC (1993)
5. Engelmann, J., Bernabeu, M.O.: Training a high-performance retinal foundation model with half-the-data and 400 times less compute. Nature Communications **16**, 6862 (2025). https://doi.org/10.1038/s41467-025-62123-z

6. Engelmann, J., McTrusty, A.D., MacCormick, I.J.C., Pead, E., Storkey, A., Bernabeu, M.O.: Detecting multiple retinal diseases in ultra-widefield fundus imaging and data-driven identification of informative regions with deep learning. Nature Machine Intelligence **4**, 1143–1154 (2022). https://doi.org/10.1038/s42256-022-00566-5
7. Haurum, J.B., Escalera, S., Taylor, G.W., Moeslund, T.B.: Which tokens to use? investigating token reduction in vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 773–783 (2023). https://doi.org/10.1109/ICCVW60793.2023.00085
8. Kameoka, M.: masumoto RP data RP optos (2018). https://doi.org/10.6084/m9.figshare.7403831.v1, https://doi.org/10.6084/m9.figshare.7403831.v1, dataset, version 1
9. Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not all patches are what you need: Expediting vision transformers via token reorganizations. In: International Conference on Learning Representations (2022)
10. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
11. Playout, C., Legault, Z., Duval, R., Boucher, M.C., Cheriet, F.: A region-based approach to diabetic retinopathy classification with superpixel tokenization. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15005, pp. 36–45. Springer Nature Switzerland (October 2024). https://doi.org/10.1007/978-3-031-72086-4_4
12. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. In: Advances in Neural Information Processing Systems. vol. 34, pp. 13937–13949 (2021)
13. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (October 2017)
14. Tang, Z., Wang, L., Guo, Z., Liang, Q., Xue, S., Feng, C., Ran, L., Chen, L., Wang, X., Li, J.: A multimodal retinal image dataset for diabetic retinopathy detection using foundation models. Scientific data **13** (03 2026). https://doi.org/10.1038/s41597-026-07005-9
15. Tran, H.C., Nguyen, D.M., Nguyen, D.M., Nguyen, T.T., Le, N., Xie, P., Sonntag, D., Zou, J.Y., Nguyen, B.T., Niepert, M.: Accelerating transformers with spectrum-preserving token merging. Advances in Neural Information Processing Systems (2025)
16. Zhou, Y., et al.: A foundation model for generalizable disease detection from retinal images. Nature (2023)