This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Asym-MSF: Asymmetric Multi-scale Structural Fusion for Robust BCVA Prediction with Self-supervised Fidelity Constraints

Yifei Wang[1], Qian Zhou[1], Xu Han[1], Hua Zou[1, ✉], Haifeng Jiang[2], Yong Wang[2], and Danmin Cao[3]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] Aier Eye Hospital of Wuhan University (Wuhan Aier Eye Hospital), Wuhan, Hubei Province, China
[3] The First Affiliated Hospital of Guangzhou Medical University, Guangzhou Medical University, Guangzhou, China
zouhua@whu.edu.cn

Abstract. Postoperative Best Corrected Visual Acuity (BCVA) prediction is essential for cataract surgical planning. Existing deep learning methods commonly rely on symmetric fusion, implicitly treating all modalities as equally reliable. However, in clinical practice, auxiliary modalities (e.g., SLO, text) are often noisy or missing, while Optical Coherence Tomography (OCT) provides high-resolution structural evidence of retinal status. Symmetric fusion may distort high-fidelity OCT representations. To address this, we propose Asym-MSF, an asymmetric framework that prioritizes OCT as the foundational anchor. We introduce an Asymmetric Latent Bottleneck (ALB), which generates spatial-aware seeds from OCT to selectively compress auxiliary modalities, effectively filtering out irrelevant noise before fusion. Furthermore, we design a Multi-scale HyperNetwork. Unlike standard MLP-based generators, our module captures multi-granular structural contexts to generate position-specific kernels. To preserve anatomical fidelity during dynamic modulation, we integrate a Self-supervised Feature Restorer (SFR), which reconstructs OCT features under hybrid amplitude, directional, spatial-gradient, and frequency-consistency constraints. Experiments on a clinical dataset of 2,635 cases demonstrate that Asym-MSF achieves the best performance among the compared methods, with a Mean Absolute Error (MAE) of 0.0378. Code is available at https://github.com/wyf-cyber/Asym-MSF.

Keywords: Cataract BCVA Prediction · Asymmetric Multimodal Learning · Multimodal fusion · Multi-scale Dynamic Convolution

1 Introduction

Cataract is a leading cause of global visual impairment [2], making accurate prediction of postoperative Best Corrected Visual Acuity (BCVA) vital for surgical

planning [10,11]. Clinically, multimodal data provide complementary information for accurate prediction and follow a strict hierarchy: Optical Coherence Tomography (OCT) provides high-fidelity structural resolution of retinal layers [5], while Scanning Laser Ophthalmoscopy (SLO) and clinical text provide auxiliary context [1]. Meanwhile, real-world data are often incomplete or heterogeneous, with auxiliary modalities frequently missing or quality-compromised [15].

Despite this hierarchy, existing multimodal methods largely employ symmetric fusion paradigms, such as feature concatenation or uniform cross-attention [13,6]. These approaches inadvertently treat noisy auxiliary data as equal to the high-precision OCT signal. Consequently, when auxiliary modalities are missing or corrupted, symmetric fusion allows artifacts to propagate into the foundational representation—a phenomenon we term "feature contamination" that degrades structural integrity. To address this, we propose Asym-MSF, an asymmetric framework that prioritizes OCT as the structural anchor. Inspired by the efficiency of dynamic kernel networks [8,14,4], we identify that standard linear generators lack the inductive bias to perceive multi-scale anatomy and ensure structural fidelity. We introduce three core innovations:

1) Asymmetric Latent Bottleneck (ALB): Instead of standard concatenation or transformer blocks [12], ALB generates "spatial-aware seeds" from OCT to actively query and compress auxiliary modalities, similar to the cross-attention mechanism in Perceiver architectures [7]. This bottleneck ensures that only information relevant to the foundational OCT structure is retained, filtering out noise from incomplete auxiliary data.
2) Multi-scale HyperNetwork for DCM: Standard kernel generators struggle with the complex, multi-granular variations of retinal layers. We propose a Multi-scale HyperNetwork within the Dynamic Convolutional Modulation (DCM) block. By aggregating features through parallel dilated convolutions with varied receptive fields [16], it generates position-specific kernels sensitive to both localized lesions and global topological structures.
3) Self-supervised Feature Restorer (SFR): To reduce the risk that dynamic modulation overfits to auxiliary context or introduces auxiliary-induced distortions, SFR enforces a self-supervised reconstruction constraint on the OCT representation. It reconstructs the original OCT signal from the generated kernels using a novel hybrid loss that constrains amplitude (MSE), structural continuity via spatial gradients, and spectral integrity through Fast Fourier Transforms (FFT) [9].

Experiments on 2,635 clinical cases show that Asym-MSF outperforms the compared multimodal baselines under the same dataset split and evaluation protocol, achieving an MAE of 0.0378.

2 Methodology

Following the established feature extraction backbone [15], we utilize modality-specific encoders to project heterogeneous inputs into a unified D-dimensional

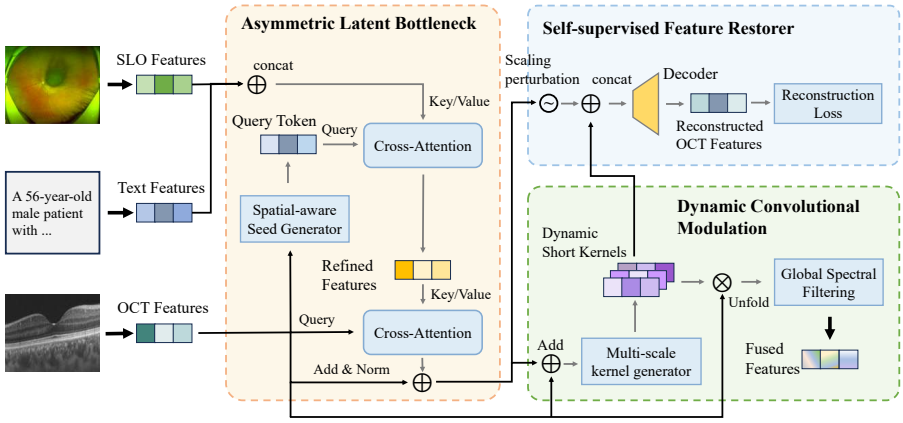

Fig. 1. Overview of our proposed Asym-MSF framework, highlighting the Asymmetric Latent Bottleneck, Dynamic Convolutional Modulation, and the hybrid constraints in the Feature Restorer.

latent space. This yields aligned feature sequences $\mathbf{F}_{oct}, \mathbf{F}_{slo}, \mathbf{F}_{text} \in \mathbb{R}^{B\times L\times D}$, where B is the batch size and L is the sequence length. As illustrated in Fig. 1, our Asym-MSF framework establishes $\mathbf{F}_{oct}$ as the structural anchor, adaptively modulated by auxiliary contexts via an Asymmetric Latent Bottleneck (ALB) and a Dynamic Convolutional Modulation (DCM), constrained by a Self-supervised Feature Restorer (SFR).

2.1 Asymmetric Latent Bottleneck (ALB)

To effectively aggregate heterogeneous signals (SLO and Text) without overwhelming the OCT structural baseline, we propose the ALB. Instead of using predefined learnable tokens, ALB dynamically generates N spatial-aware latent seeds from the OCT features, ensuring the compression is guided by the foundational anatomy.

First, we apply a 1D convolution (kernel size 3) followed by GELU activation and Adaptive Average Pooling to $\mathbf{F}_{oct}$ to condense it into a fixed number of latent queries $\mathbf{Q}_{seed} \in \mathbb{R}^{B\times N\times D}$ (where $N = 64$):

$$\mathbf{Q}_{seed} = \text{AdaptiveAvgPool}(\text{GELU}(\text{Conv1D}(\mathbf{F}_{oct}))) \tag{1}$$

Next, the auxiliary modalities are concatenated as $\mathbf{F}_{aux} = [\mathbf{F}_{slo}, \mathbf{F}_{text}] \in \mathbb{R}^{B\times(L_{slo}+L_{text})\times D}$. We employ a two-stage cross-attention bottleneck. In the compression stage, the OCT-derived seeds query the auxiliary features to distill a refined, compact context $\mathbf{Z}_{aux} \in \mathbb{R}^{B\times N\times D}$:

$$\mathbf{Z}_{aux} = \text{MultiHeadAttn}(\text{query} = \mathbf{Q}_{seed}, \text{key} = \mathbf{F}_{aux}, \text{value} = \mathbf{F}_{aux}) \tag{2}$$

In the feedback stage, the original OCT sequence queries this compressed summary $\mathbf{Z}_{aux}$ to fetch position-specific auxiliary information. The output is fused via a residual connection and LayerNorm, yielding the localized auxiliary context $\mathbf{C}_{aux} \in \mathbb{R}^{B\times L\times D}$:

$$\mathbf{C}_{aux} = \text{LayerNorm}(\mathbf{F}_{oct} + \text{MultiHeadAttn}(\mathbf{F}_{oct}, \mathbf{Z}_{aux}, \mathbf{Z}_{aux})) \tag{3}$$

2.2 Dynamic Convolutional Modulation (DCM)

To integrate $\mathbf{C}_{aux}$ into the foundational OCT stream position-by-position, we design a modulation scheme containing a Multi-scale Kernel Generator and Global Spectral Filtering.

Multi-scale Dynamic Kernels. We generate token-specific kernels by fusing the localized auxiliary context with the original OCT feature, i.e., $\mathbf{X}_{comb} = \mathbf{C}_{aux} + \mathbf{F}_{oct}$. As shown in Fig. 2, we feed $\mathbf{X}_{comb}$ into a parallel 4-branch 1D convolutional network to capture complex contextual patterns. The branches consist of kernel-size-1 convolution and three kernel-size-3 1D convolutions with respective dilation rates of 1, 2, and 3, granting varied receptive fields.

The concatenated multi-scale features are projected to $\mathbb{R}^{B\times L\times (D\cdot K)}$, where $K = 3$ is the target dynamic kernel size. To prevent parameter explosion and ensure training stability, we normalize and bound the generated weights using a `tanh` function scaled by a learnable amplitude parameter s (initialized to 5.0):

$$\mathbf{W}_{dyn} = s \cdot \tanh(\text{LayerNorm}(\text{Proj}(\text{Concat}(\mathbf{X}_{b1}, \mathbf{X}_{b2}, \mathbf{X}_{b3}, \mathbf{X}_{b4})))) \tag{4}$$

The tensor is reshaped to $\mathbb{R}^{B\times L\times D\times K}$. During forward propagation, $\mathbf{F}_{oct}$ is locally unfolded into sliding windows of size K. We perform element-wise multiplication between the sliding windows and $\mathbf{W}_{dyn}$, followed by aggregation, to achieve position-aware structural modulation, denoted as $\mathbf{F}_{short}$.

Global Spectral Filtering. To complement the local receptive field ($K = 3$), we adopt FFT-based filtering [14] to model global token interactions. Let $\mathcal{T}$ denote transposition from $B \times L \times D$ to $B \times D \times L$, and let $\mathcal{R}_{L_p}$ and $\mathcal{R}_{L_p}^{-1}$ denote Real FFT and inverse Real FFT with length $L_p = 2^{\lceil \log_2 L \rceil}$. The spectral branch is defined as

$$\begin{aligned} \Delta\mathbf{F} &= \mathcal{T}^{-1}\mathcal{C}_L\mathcal{R}_{L_p}^{-1}\big(\mathcal{R}_{L_p}(\mathcal{T}(\mathbf{F}_{short})) \odot \mathbf{W}_{freq}\big)\,, \\ \mathbf{F}_{out} &= \text{LN}(\mathbf{F}_{oct} + \text{Dropout}(\Delta\mathbf{F}))\,, \end{aligned} \tag{5}$$

where $\mathcal{C}_L$ crops the output back to length L, and $\mathbf{W}_{freq} \in \mathbb{C}^{1\times D\times (L_p/2+1)}$.

2.3 Self-supervised Feature Restorer (SFR)

A persistent risk in adaptive modulation is that the dynamic kernels $\mathbf{W}_{dyn}$ might overfit to the auxiliary context and discard the critical OCT structural

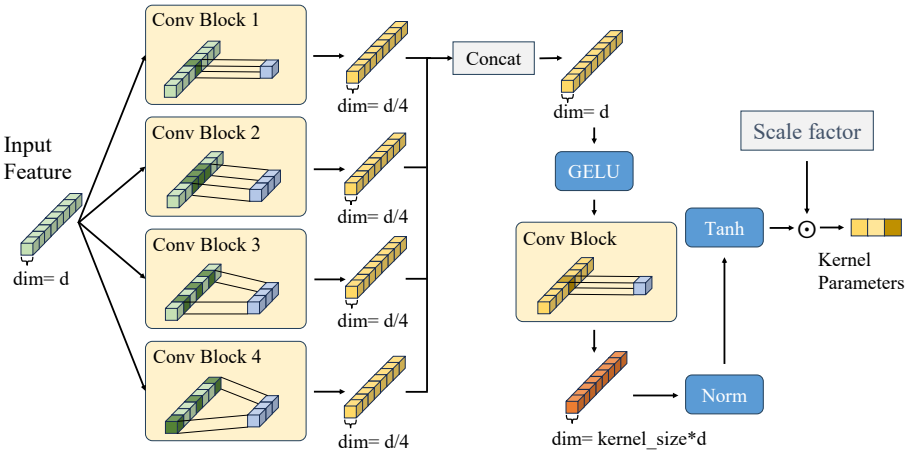

Fig. 2. Overview of our proposed multi-scale kernel generator.

fingerprints. To encourage structural fidelity, we introduce the SFR as a pretext regularization task.

Unlike standard decoders, SFR regularizes the generated kernels through an OCT reconstruction pretext task. We first apply an additive random perturbation to the auxiliary context, $\tilde{\mathbf{C}}_{aux} = \mathbf{C}_{aux} + \epsilon$, to prevent the restorer from relying solely on context shortcuts[3]. The flattened dynamic kernel is independently projected back to D dimensions via a non-linear path $\mathcal{P}_{ker}$. The concatenated pair is then processed by a token-wise MLP decoder $\mathcal{D}$:

$$\hat{\mathbf{F}}_{oct} = \mathcal{D}(\mathcal{P}_{ker}(\text{Flatten}(\mathbf{W}_{dyn})) \oplus \tilde{\mathbf{C}}_{aux}) \tag{6}$$

This pretext task forces the multi-scale generator to deeply embed the pristine OCT structural properties into the kernel weights. We optimize this via a hybrid multi-domain loss function comprehensively constraining the amplitude, geometry, and frequency components:

$$\mathcal{L}_{rec} = \mathcal{L}_{mse}(\hat{\mathbf{F}}_{oct}, \mathbf{F}_{oct}) + \lambda_1 \mathcal{L}_{cos} + \lambda_2 \mathcal{L}_{grad} + \lambda_3 \mathcal{L}_{fft} \tag{7}$$

where $\mathcal{L}_{cos} = 1 - \text{CosSim}(\hat{\mathbf{F}}_{oct}, \mathbf{F}_{oct})$ aligns feature directions. Crucially, we introduce a Spatial Gradient loss $\mathcal{L}_{grad}$ to penalize structural discrepancies between adjacent tokens ($\Delta\mathbf{F}^{(l)}_{oct} = \mathbf{F}^{(l+1)}_{oct} - \mathbf{F}^{(l)}_{oct}$), alongside an MSE loss in the Fourier domain $\mathcal{L}_{fft}$ to regularize spectral consistency. The balancing coefficients are empirically set to $\lambda_1 = 0.1$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.1$.

This pretext task encourages the generator to encode OCT structural priors into $\mathbf{W}_{dyn}$ rather than relying solely on auxiliary context. Since SFR is only an auxiliary regularizer, its decoder is discarded at inference; the deployed model retains the regularized, OCT-preserving kernel generator without extra cost.

3 Experiments and Results

3.1 Dataset and Implementation Details

Clinical Dataset. Following the experimental setup in [15], the dataset comprises 2,635 affected eyes from 1,960 cataract patients. The multimodal dataset contains: (1) **Clinical Text**: patient demographics, fundus disease history, and ophthalmologists' diagnostic descriptions; and (2) **Preoperative Images**: Optical Coherence Tomography (OCT) and Scanning Laser Ophthalmoscopy (SLO).

OCT and clinical text are routinely available, while only 988 eyes possess paired SLO images, corresponding to a missing rate of approximately 63%. This real-world modality incompleteness motivates our focus on uncertain SLO availability rather than synthetic modality dropout, which would restrict evaluation to the smaller complete-modality subset and may introduce selection bias.

The ground truth label is postoperative BCVA, measured using the Logarithm of the Minimal Angle of Resolution (LogMAR). We use the same eye-level 8:1:1 split and evaluation protocol as prior BCVA prediction studies for fair comparison. All compared methods are evaluated under the same split and protocol.

Implementation. Experiments were implemented using PyTorch on 8× NVIDIA RTX 3090 GPUs. All input images were resized to 224×224. We utilized the Adam optimizer with an initial learning rate of 0.003 ($\beta_1 = 0.9, \beta_2 = 0.99$). The batch size was set to 64. The hybrid multi-domain reconstruction loss weight is set to 0.1, the weights of the other loss components and the temperature coefficient τ are set to be the same as in [15]. The model was trained for 100 epochs using a StepLR scheduler, which decays the learning rate by a factor of 0.5 every 10 epochs. Performance is evaluated using Accuracy (Acc, percentage of samples within an error of ± 0.10 LogMAR) and Mean Absolute Error (MAE).

3.2 Quantitative Performance

For baselines originally proposed for other multimodal tasks, we adapt their fusion or missing-modality learning strategies to postoperative BCVA prediction and evaluate all methods under the same dataset split and protocol. As shown in Table 1, Asym-MSF achieves the best performance on our clinical dataset, with an MAE of 0.0378. This corresponds to a 17.3% error reduction over the strongest contemporary baseline, UML [15]. The improvement is more evident compared with earlier symmetric fusion methods. For example, CTT-Net [13] and MAF-Net [6] show limited precision, partly due to insufficient exploitation of semantic context or the instability of modality-level synthesis. Although MMIN [17] and IML-Net [18] further address missing modalities through feature-level imputation, such generative strategies may still introduce biases into high-fidelity OCT representations when auxiliary modalities are absent or unreliable.

A particularly relevant comparison is with DyKen-Hyena [14], which provides the structural foundation for dynamic kernel modulation. When directly applied to our dataset, DyKen-Hyena obtains an MAE of 0.0482. This suggests that its

MLP-based kernel generator is less effective in capturing the complex multi-scale anatomical patterns in OCT scans, especially under a high SLO missing rate of approximately 63

By contrast, Asym-MSF reduces the MAE by 21.6% compared with DyKen-Hyena. This improvement validates the effectiveness of our asymmetric design. The Multi-scale HyperNetwork generates more anatomy-aware dynamic kernels for both local lesions and global structural context, while ALB filters incomplete or noisy auxiliary information before modulation. Moreover, SFR further constrains the modulation process to preserve OCT structural fidelity, leading to more robust postoperative BCVA prediction.

Table 1. Comparison with state-of-the-art methods on the clinical cataract dataset. "•" denotes the modality is used, "○" denotes it is not used or unavailable. Note that SLO is missing in ∼63% of cases. DyKen-Hyena and Asym-MSF are both configured with OCT as the foundational modality.

Method	Input Modalities				MAE (↓)	Acc (↑)
	Text	OCT	SLO	B-Scan		
CTT-Net [13]	•	•	○	○	0.1682	88.73%
MAF-Net [6]	•	•	•	•	0.1761	85.42%
MMIN [17]	•	•	•	•	0.1193	91.69%
IML-Net [18]	•	•	•	•	0.1224	94.31%
UML [15]	•	•	•	○	0.0457	96.25%
DyKen-Hyena [14]	•	•	•	○	0.0482	96.28%
Asym-MSF (Ours)	•	•	•	○	**0.0378**	**98.60%**

3.3 Ablation Studies

Effectiveness of Key Components. We conducted an ablation analysis to evaluate the individual contributions of the ALB, DCM, and SFR modules. Since SFR regularizes the kernels generated by DCM, it cannot form a valid standalone variant without DCM. Therefore, we assess its contribution by comparing the full model with the corresponding no-SFR variant.

As shown in Table 2, replacing the symmetric baseline with the OCT-driven ALB reduces the MAE from 0.0457 to 0.0418, confirming its ability to filter auxiliary noise. The standalone Multi-scale DCM significantly enhances structural modeling, achieving an MAE of 0.0391. Interestingly, combining ALB and DCM without the restorer leads to a slight performance dip (0.0419 MAE), suggesting that increased modulation complexity requires explicit regularization. This is addressed by the SFR module, which leverages hybrid reconstruction losses to align dynamic kernels with anatomical truth, ultimately yielding the optimal MAE of 0.0378.

Notably, while Accuracy saturates at 98.60% across most configurations, the continuous MAE metric remains highly sensitive to architectural improvements. This discrepancy arises because Accuracy is a discrete measure based on a ± 0.10 LogMAR threshold, whereas MAE captures fine-grained regression precision. Thus, the consistent reduction in MAE proves that our asymmetric framework provides a more robust and precise prognosis even when the categorical accuracy reaches a plateau.

Table 2. Ablation study of Asym-MSF components. Baseline: Symmetric fusion baseline (UML [15]).

Baseline	ALB	DCM	SFR	MAE (↓)	Accuracy (↑)
✓				0.0457	96.25%
✓	✓			0.0418	**98.60%**
✓		✓		0.0391	**98.60%**
✓	✓	✓		0.0419	**98.60%**
✓		✓	✓	0.0406	**98.60%**
✓	✓	✓	✓	**0.0378**	**98.60%**

Foundational Modality Selection. Table 3 validates our clinical hierarchy by evaluating the performance of different foundational anchors. Treating clinical text as the primary modality—where images are used to modulate text features as in the original DyKen-Hyena [14]—yields an inferior MAE of 0.0390. This result confirms that clinical summaries, while informative, lack the dense anatomical evidence required for precise surgical prognosis. In contrast, establishing OCT as the structural anchor achieves the highest precision (MAE 0.0378). This aligns with clinical practice, where high-resolution retinal imaging provides direct structural evidence related to postoperative visual potential, while SLO and text serve as auxiliary modifiers that refine the localized context within the foundational stream.

Table 3. Ablation on the choice of foundational modality. The strategy with OCT as the anchor significantly outperforms text-based or symmetric fusion.

Foundation	Auxiliary	Strategy	MAE (↓)	Accuracy (↑)
-	-	Symmetric (UML)	0.0457	96.25%
Text	OCT + SLO	Asymmetric	0.0390	**98.60%**
OCT	Text + SLO	Asymmetric	**0.0378**	**98.60%**

4 Conclusion

We proposed Asym-MSF, an asymmetric multimodal framework for postoperative BCVA prediction. By using OCT as the structural anchor, ALB filters auxiliary noise, DCM performs multi-scale structure-aware modulation, and SFR preserves OCT fidelity through self-supervised reconstruction. Experiments on 2,635 clinical cases show that Asym-MSF achieves superior performance under severe modality incompleteness, highlighting the value of clinically grounded asymmetric fusion for ophthalmic prognosis.

Acknowledgments. This work was partly supported by the China University Industry-Academia-Research Innovation Fund (Grant No. 2025XJS023), the Dongguan Science and Technology of Social Development Program (Grant No. 20231800935472), and the Xiangjiang Public Welfare Foundation Technology Innovation and Application Development Project (Grant No. KY24013).

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

References

1. Bille, J.F. (ed.): High Resolution Imaging in Microscopy and Ophthalmology: New Frontiers in Biomedical Optics. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-16638-0
2. Bourne, R.R.A., Steinmetz, J.D., Saylan, M., et al.: Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the right to sight: an analysis for the global burden of disease study. The Lancet Global Health **9**(2), e144–e160 (2021). https://doi.org/10.1016/S2214-109X(20)30489-7
3. Geirhos, R., et al.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z
4. Ha, D., Dai, A.M., Le, Q.V.: Hypernetworks. In: International Conference on Learning Representations (2017)
5. Huang, D., et al.: Optical coherence tomography. Science **254**(5035), 1178–1181 (1991). https://doi.org/10.1126/science.1957169
6. Huang, Z., Lin, L., Cheng, P., Peng, L., Tang, X.: Multi-modal brain tumor segmentation via missing modality synthesis and modality-level attention fusion. arXiv preprint arXiv:2203.04586 (2022). https://doi.org/10.48550/arXiv.2203.04586
7. Jaegle, A., et al.: Perceiver IO: A general architecture for structured inputs and outputs. In: International Conference on Learning Representations (2022)
8. Jia, X., De Brabandere, B., Tuytelaars, T., Van Gool, L.: Dynamic filter networks. In: Advances in Neural Information Processing Systems 29. pp. 667–675 (2016)
9. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13919–13929 (2021). https://doi.org/10.1109/ICCV48922.2021.01366
10. Kansal, V., Schlenker, M., Ahmed, I.I.K.: Interocular axial length and corneal power differences as predictors of postoperative refractive outcomes after cataract surgery. Ophthalmology **125**(7), 972–981 (2018). https://doi.org/10.1016/j.ophtha.2018.01.021

11. Ting, D.S.W., et al.: Deep learning in ophthalmology: The technical and clinical considerations. Progress in Retinal and Eye Research **72**, 100759 (2019). https://doi.org/10.1016/j.preteyeres.2019.04.003
12. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
13. Wang, J., Wang, J., Chen, T., Zheng, W., Xu, Z., Wu, X., Xu, W., Ying, H., Chen, D.Z., Wu, J.: Ctt-net: A multi-view cross-token transformer for cataract postoperative visual acuity prediction. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 835–839. IEEE (2022). https://doi.org/10.1109/BIBM55620.2022.9995392
14. Wang, Y., Wang, W., Luo, Y.: DyKen-Hyena: Dynamic kernel generation via cross-modal attention for multimodal intent recognition. arXiv preprint (2025). https://doi.org/10.48550/arXiv.2509.09940
15. Yang, T., Zhou, Q., Zou, H., Jiang, H., Wang, Y.: UML: A unified multimodal learning framework for cataract postoperative visual acuity prediction with uncertain missing modalities. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10889124
16. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations (2016)
17. Zhao, J., Li, R., Jin, Q.: Missing modality imagination network for emotion recognition with uncertain missing modalities. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2608–2618. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.203
18. Zhou, Q., Zou, H., Jiang, H., Wang, Y.: Incomplete multimodal learning for visual acuity prediction after cataract surgery using masked self-attention. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol. 14226, pp. 735–744. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43990-2_69