CP-IMoE: Collaborative Prompt-Guided Interactive Mixture-of-Experts for Incomplete Multimodal Learning

Supplementary Material

1. Details of Datasets

We evaluate our CP-IMoE on three multimodal medical imaging datasets:

Harvard30k Glaucoma [9]. This dataset provides paired *fundus* and *OCT* images for glaucoma screening. We follow the binary classification protocol (*normal* vs. *glaucoma*). Fundus images offer global morphological cues, whereas OCT provides structural information, forming a complementary multimodal pair.

Derm7pt [6]. Derm7pt contains paired *clinical* and *dermoscopic* images of skin lesions. Following the Seven-Point Checklist protocol, we group diagnoses into five classes: *NEV (nevus)*, *BCC (basal cell carcinoma)*, *MEL (melanoma)*, *MISC (miscellaneous)*, and *SK (seborrheic keratosis)*. Clinical images provide macroscopic lesion context, while dermoscopy reveals fine-grained structures crucial for malignancy assessment.

MMC-AMD [15]. MMC-AMD contains paired *fundus* and *OCT* images for age-related macular degeneration (AMD) classification. We adopt a four-class setup comprising *wetAMD*, *dryAMD*, *PCV*, and *normal*. Fundus imaging captures global retinal appearance, whereas OCT provides detailed biomarkers of AMD pathology.

The class distribution in the training and test sets is provided in Tab. S1.

Dataset	Modality	Class	Train	Test
Harvard30k Glaucoma	Fundus + OCT	Normal	859	224
		Glaucoma	628	148
Derm7pt	Clinic + Derm	NEV	256	219
		BCC	19	16
		MEL	90	101
		MISC	32	40
		SK	16	19
MMC-AMD	Fundus + OCT	wetAMD	266	65
		dryAMD	50	8
		PCV	143	41
		Normal	156	39

Table S1. Class distribution of the three multimodal datasets.

2. More Results

2.1. Model Complexity

As summarized in Tab. S2, our CP-IMoE significantly reduces the training cost. Compared with full model training which means both the backbone and all expert networks are jointly optimized, CP-IMoE reduces trainable parameters from 93.3M to 21.0M and training FLOPs from 49.2G to 16.7G. This efficiency gain stems from freezing both the pretrained encoders and the expert networks, requiring only three MLP layers to be optimized during deployment. Such a lightweight design demonstrates the scalability and practicality of CP-IMoE, especially for large-scale multimodal systems where computational resources are constrained.

Methods	Trainable Params (M)	Training FLOPs (G)
Full Model Training	93.3	49.2
CP-IMoE (Ours)	21.0 ($\downarrow$ 77.5%)	16.7 ($\downarrow$ 66.1%)

Table S2. Model complexity analysis.

2.2. Impact of Backbone

As shown in Tab. S3, we investigate the impact of different backbone choices on the performance of CP-IMoE under varying missing-modality ratios on the MMC-AMD dataset. When using the ResNet50 [5] backbone with fine-tuning, CP-IMoE achieves the highest and most stable performance. In contrast, when employing the retinal foundation model RETFound [18], both frozen and fine-tuned settings exhibit a notable performance degradation. Although RETFound provides strong general retinal features, its pre-training objective (self-supervised on large-scale unpaired fundus images) lacks domain adaptation for multimodal fusion tasks, leading to suboptimal cross-modal analysis. These results suggest that foundation models may require dedicated multimodal adaptation or prompt-tuning strategies to fully realize their potential in incomplete-modality scenarios.

2.3. Model Performance of Each Class

In addition, to intuitively illustrate the comparative performance of different models and to provide a detailed analysis of each individual class in different datasets, we plot the class-wise F1 scores of compared missing modality models across all three datasets, as shown in Fig. S1. CP-IMoE consistently achieves the highest average result across the

Backbones	Missing Rate: 20%			Missing Rate: 40%			Missing Rate: 60%			Missing Rate: 80%		
	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa	Acc
ResNet50 [5] Finetuning	**86.15**	**78.62**	**85.62**	**83.96**	**78.79**	**85.62**	**83.90**	**74.63**	**83.01**	**80.79**	**72.83**	**81.70**
RETFound [18] Frozen	68.29	58.11	72.55	66.54	51.04	68.63	63.49	45.47	65.36	57.07	39.19	61.44
RETFound [18] Finetuning	47.42	48.55	67.32	46.32	47.58	67.32	45.49	48.65	67.97	48.42	50.43	68.63

Table S3. The impact of backbone selection.

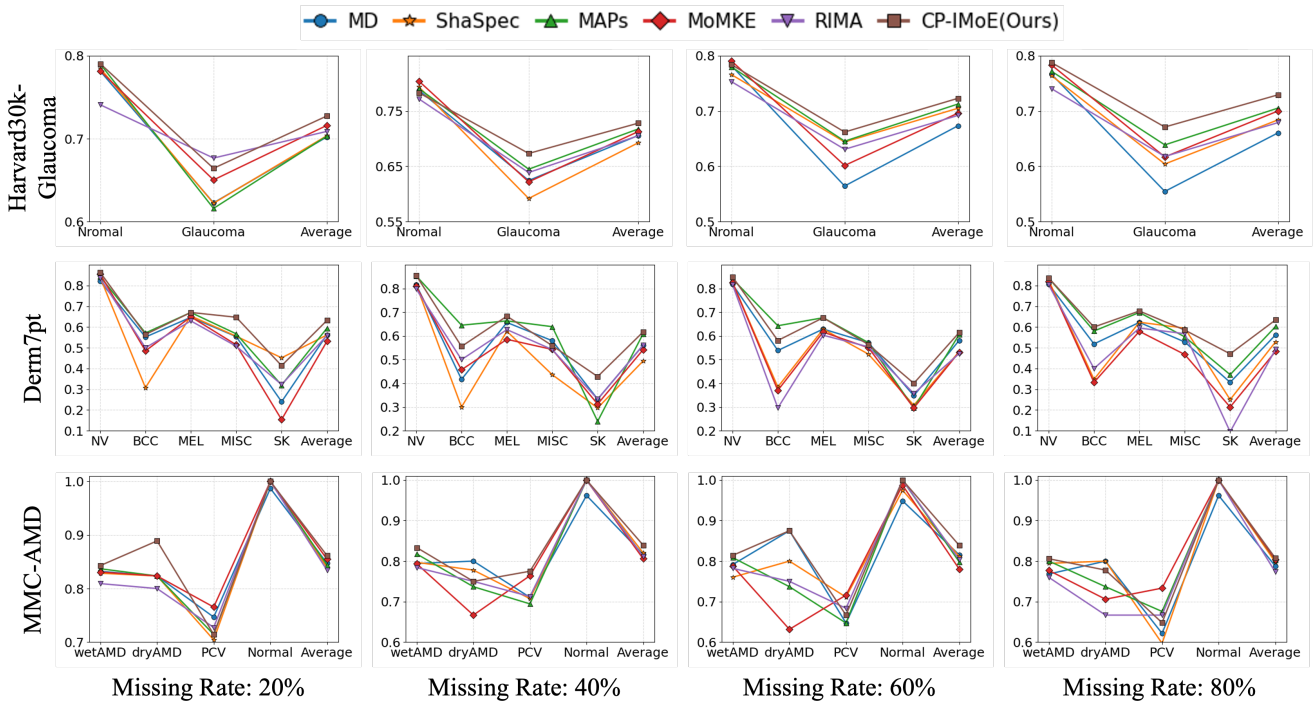

Figure S1. Class-wise F1-score comparison of missing modality models on the Harvard30k Glaucoma [9], Derm7pt [6] and MMC-AMD [15] datasets.

four missing-modality settings on the three datasets. Specifically, on Harvard30k Glaucoma [9], CP-IMoE consistently outperforms baseline MoMKE [16] across all missing ratios. It improves the average F1 by 1.3% at 20% missing rate, with the margin further increasing to 2.4% at 40% missing rate and 2.1% at 60% missing rate. Under the extreme 80% missing setting, the improvement reaches 3.1%. On Derm7pt [6], a similar trend is observed. CP-IMoE consistently outperforms baseline MoMKE [16], improving the average F1 by 4.0% at 20% missing rate, with notable gains on challenging classes such as BCC and MISC. The gap widens as missingness increases, reaching gains of 7.7% at 40% missing rate and 7.2% at 60% missing rate. Under the extreme 80% missing setting, CP-IMoE achieves its largest improvement of 13.2%, demonstrating strong resilience on fine-grained lesion categories. On MMC-AMD [15], CP-IMoE again shows superior performance across all missing ratios. Although the gain at 20% missing rate is modest (from 85.49% to 86.15%), the advantage becomes clear at moderate levels, with increases from 80.60% to 83.96% at 40% missing rate and from 78.06% to 83.90% at 60% missing rate.

Overall, across all datasets and missing ratios, CP-IMoE not only achieves higher average F1 scores but also maintains much more stable performance as the missing rate increases. This highlights its strong robustness and its adaptive utilization of both modality-specific and interactive information.

2.4. Visualization Analysis of Experts Weight

Comparison of expert weight activation between the classic router and CP-IMoE on the Harvard30k Glaucoma dataset. From Fig. S2, we can observe that when the fundus modality (the secondary modality) under low missing ratios (20%–40%), the classic router can maintain relatively stable expert activation. However, as the missing

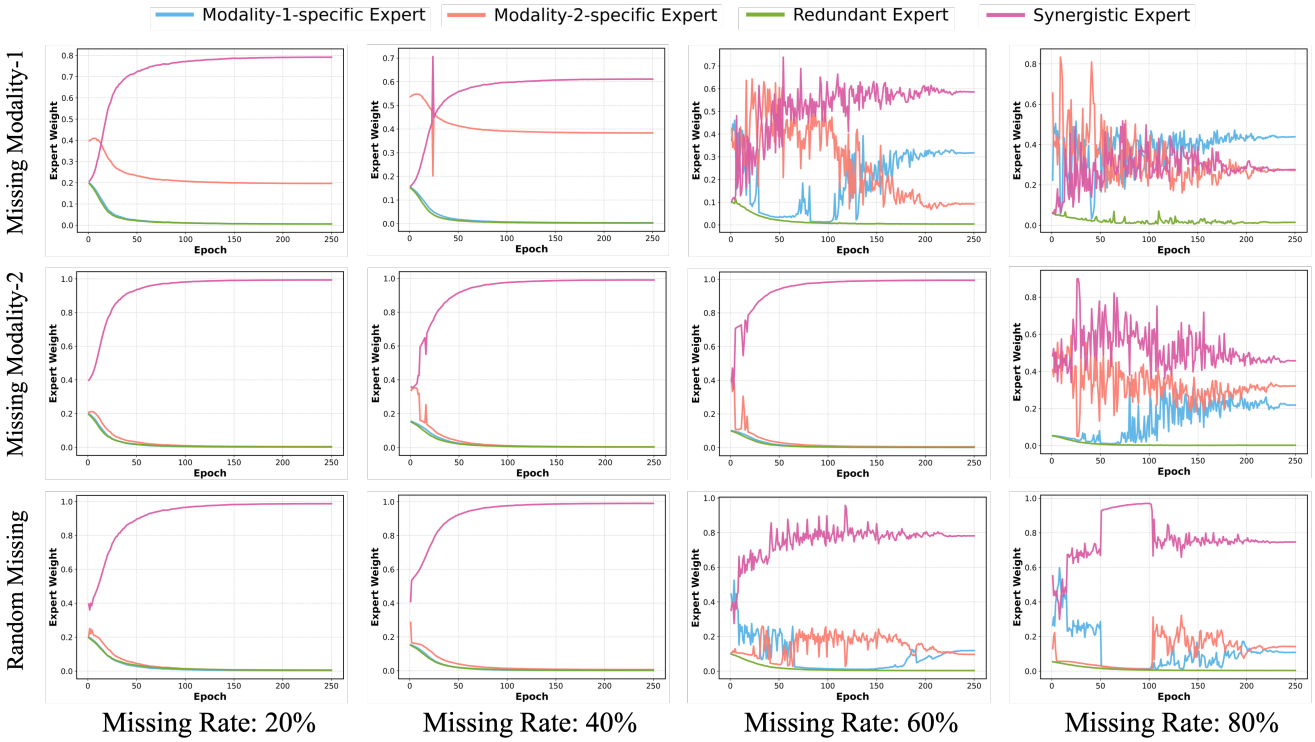

Figure S2. Expert weight visualization of classic router network under different missing-modality ratios on Harvard30k Glaucoma [9].

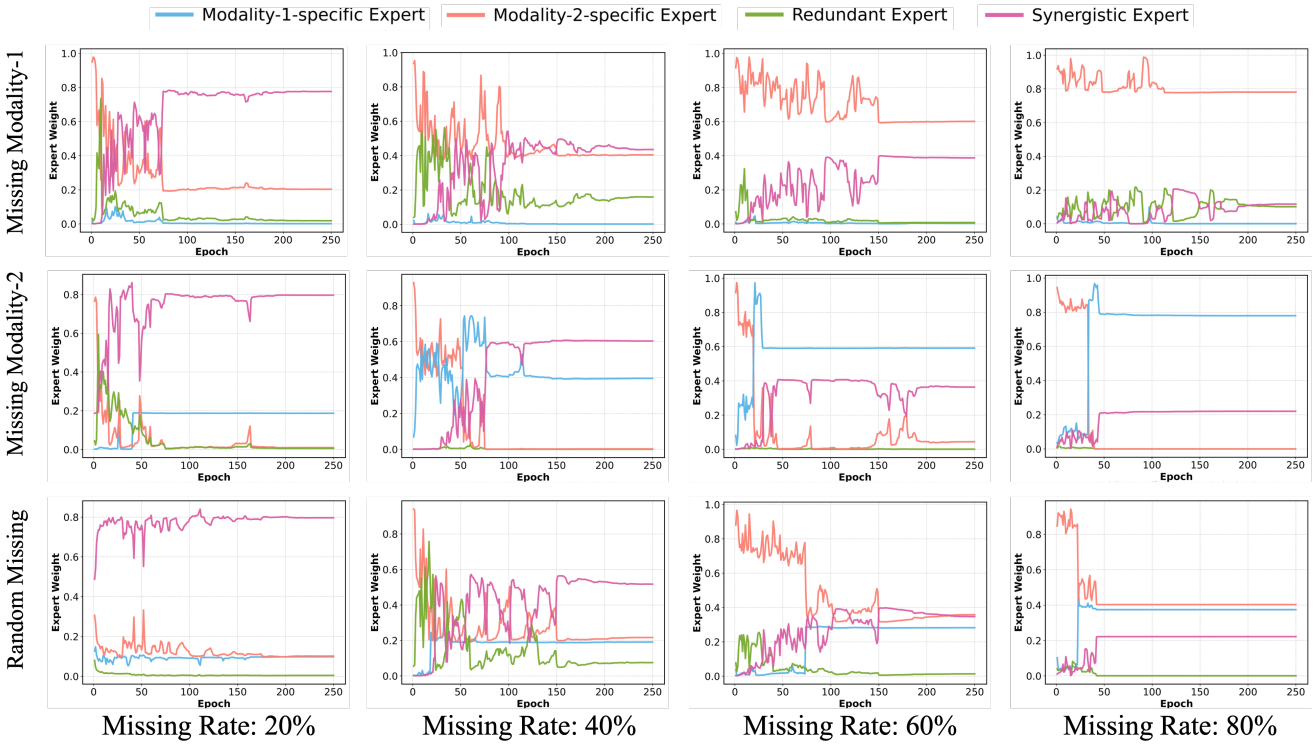

Figure S3. Expert weight visualization of our CP-IMoE under different missing-modality ratios on Derm7pt [6].

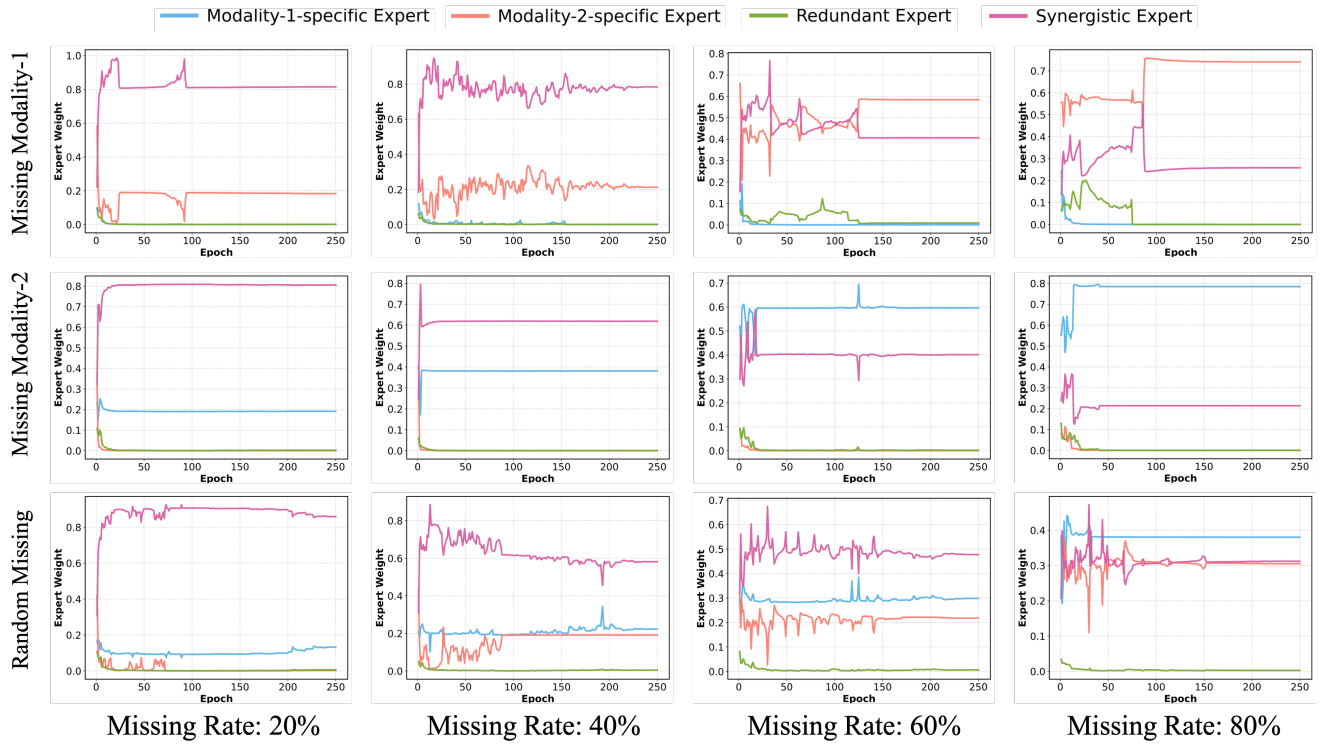

Figure S4. Expert weight visualization of our CP-IMoE under different missing-modality ratios on MMC-AMD [15].

ratio increases, the routing process becomes chaotic, and the expert weights fluctuate dramatically, indicating that the router struggles to select appropriate experts for downstream inference. In contrast, when the OCT modality (the primary modality) is missing, the router collapses much earlier. Even at low missing ratios (20%–60%), the classic router consistently over-relies on the synergistic expert. When the missing ratio reaches 80%, the router again becomes unstable, assigning expert weights in a highly irregular manner. Under random-missing conditions, the routing behavior lies between the two extremes of missing the primary or secondary modality. The router exhibits collapse at low missing ratios (20%–40%), and as the missing ratio exceeds 60%, it re-enters a disordered state, though less severe than when the primary modality is missing but stronger than the secondary-modality case.

In contrast, our proposed CP-IMoE consistently achieves adaptive and stable expert routing across all missing-modality settings (Fig. 6), effectively maintaining balanced expert utilization and preventing representation collapse.

Expert weight activation on Derm7pt and MMC-AMD dataset. To further validate the generalization ability of CP-IMoE across different datasets, we visualize the expert activation on the Derm7pt and MMC-AMD datasets, as shown in Fig. S3 and Fig. S4. Visualizations across multiple datasets consistently reveal a clear trend: when the missing-modality ratio is low (e.g., $< 40\%$), the model tends to rely more on the synergistic expert, as cross-modal synergy provides richer discriminative cues. However, as the missing ratio increases (e.g., $> 60\%$), the synergistic information diminishes, and the model gradually shifts its reliance toward the available modality-specific experts. In addition, the redundant expert is rarely selected under any setting, indicating that the model naturally avoids representations that do not contribute meaningful discriminative information. This consistent pattern highlights the robustness and cross-dataset generalizability of the proposed expert routing mechanism.

3. More Analysis

Why collaborative prompt (CLP) is effective for representation collapse? Two questions: (1) Why do MoE-based multimodal models suffer from representation collapse, where the router degenerates into selecting a single expert for most samples? (2) Why is this issue exacerbated under incomplete-modality conditions? The key reason lies in the nature of the router network input, which is directly derived from multimodal feature representations. These representations often share highly similar statistical structures across different samples; thus, when a modality is missing, the router network input becomes solely determined by the remaining modality. This further reduces

the variability in the router's input. As a result, the router observes low inter-sample diversity, leading to saturated softmax activations and heavily skewed routing coefficients (e.g., approaching 0 or 1).

Prior works typically tackle MoE representation collapse from three angles, (i) *Architectural or objective regularization*, e.g., top-k routing to encourage usage diversity [13]; (ii) *Training strategies*, such as temperature annealing [2], stochastic perturbations [7]; (iii) *Expert specialization*, including pretraining or auxiliary tasks that enforce distinct expert roles to reduce overlap [10, 12].

Inspired that, our approach integrates insights from training strategies and expert specialization. We first pretrain each expert to establish well-structured and complementary expert functions. Then, instead of altering the routing mechanism, we inject structured information into the router network input, acting as both a controllable perturbation and a semantic bias. The static prompt reshapes global decision regions by conditioning routing on the presence/absence of modalities, while the dynamic prompt refines local neighborhoods by encoding discriminative cues from the available streams.

Modality Specificity. As shown in Fig. 6, the modalities present clear specificity, meaning that they contribute unequally to diagnosis. In ophthalmic disease analysis, OCT plays a more essential role, so missing OCT leads to a large drop in performance. In contrast, missing fundus images has a smaller impact. Among the three settings (OCT-missing, random-missing, and fundus-missing), the model performs best when fundus is missing. A similar trend is observed in skin disease diagnosis, where dermoscopy provides more discriminative information than clinical images.

4. Discussion and Future Works

Experiments with replacing the backbone using a retinal foundation model reveal an important and somewhat disappointing observation: unlike the rapid advancements seen in natural-image foundation models such as CLIP [11], LLMs [8], and VLMs [17], medical foundation models-especially multimodal ones-are still underdeveloped. A primary challenge is generalization. Current medical foundation models often fail to transfer effectively to downstream tasks and typically remain restricted to narrow vertical domains. For example, ophthalmic foundation models are applicable only to eye diseases, while dermatology foundation models work exclusively for skin-related tasks. This limited cross-domain usability poses a significant barrier to the broader adoption of foundation models in medical AI.

In contrast, natural-image foundation models benefit from massive, highly diverse datasets, enabling strong generalization across various vertical applications, including emotion recognition [4], anomaly detection [14], and embodied intelligence [3], often without task-specific fine-tuning. Motivated by this gap, an important future direction is to develop general-purpose medical foundation models that can generalize across multiple clinical domains and produce robust representations without requiring extensive fine-tuning, ultimately advancing the field of multimodal medical intelligence.

Furthermore, as shown in Fig. 6 and discussed in Section 3 of Supplementary Material, medical modalities exhibit strong modality specificity. This implies that multimodal inputs are not always necessary: in certain scenarios, additional modalities may introduce noise and even degrade performance [1]. Therefore, another promising research direction is to explore modality-specific modeling that selectively activates or suppresses modalities as needed. Such advancements will facilitate the development of interpretable, deployable, and lightweight multimodal diagnostic models.

References

[1] Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In *Proceedings of the 41st International Conference on Machine Learning*, pages 5608–5628, 2024. 5

[2] Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, et al. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems*, 35:34600–34613, 2022. 5

[3] Jie Feng, Jinwei Zeng, Qingyue Long, Hongyi Chen, Jie Zhao, Yanxin Xi, Zhilun Zhou, Yuan Yuan, Shengyuan Wang, Qingbin Zeng, et al. A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science. *arXiv preprint arXiv:2504.09848*, 2025. 5

[4] Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1736, 2024. 5

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[6] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2): 538–546, 2018. 1, 2, 3

[7] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. 5

[8] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In

European Conference on Computer Vision, pages 323–340. Springer, 2024. 5

[9] Yan Luo, Yu Tian, Min Shi, Tobias Elze, and Mengyu Wang. Eye fairness: A large-scale 3d imaging dataset for equitable eye diseases screening and fair identity scaling, 2024. 1, 2, 3

[10] Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 2025. 5

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5

[12] Amélie Royer, Ilia Karmanov, Andrii Skliar, Babak Ehteshami Bejnordi, and Tijmen Blankevoort. Revisiting single-gated mixtures of experts. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022. 5

[13] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. 5

[14] Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. Large language models for forecasting and anomaly detection: A systematic literature review. *arXiv preprint arXiv:2402.10350*, 2024. 5

[15] Weisen Wang, Xirong Li, Zhiyan Xu, Weihong Yu, Jianchun Zhao, Dayong Ding, and Youxin Chen. Learning two-stream cnn for multi-modal age-related macular degeneration categorization. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4111–4122, 2022. 1, 2, 4

[16] Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 438–446, 2024. 2

[17] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024. 5

[18] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023. 1, 2