

Proto-CAP: Prototype Memory Fusion and Curriculum-Adaptive Loss for Pediatric Myopia Visual Question Answering

Xu Yan^{1,2†}, Yanlin Wu^{1,2†}, Zixun Wang³, Jingtao Yu³, Ruihua Wei³, and Tao Li^{1,2(✉)}

¹ Nankai University, Tianjin, China

² Haihe Lab of ITAI, Tianjin, China

³ Tianjin Medical University Eye Hospital, Tianjin, China
litao@nankai.edu.cn

Abstract. Medical Visual Question Answering (Med-VQA) is pivotal for assisting clinical decision-making and automated diagnosis. While deep learning has achieved remarkable success in general VQA tasks, Med-VQA research targeting specific diseases—particularly pediatric myopia involving multimodal data—remains largely underexplored, primarily due to the scarcity of publicly available, expert-annotated datasets. In this paper, we present the Pediatric Myopia VQA dataset (PM-VQA). To the best of our knowledge, this is the first publicly available Med-VQA dataset dedicated to pediatric myopia. It comprises **2,741** multimodal images, totaling **14,781** expert-verified question-answer pairs. Furthermore, we propose an enhanced framework named Proto-CAP, designed to address the semantic gap and overfitting issues inherent in medical small-scale data. Within Proto-CAP, we introduce a Question-Guided Prototype Fusion mechanism and a Curriculum-Adaptive Poly-2 Loss to enhance prior knowledge injection and dynamically balance training difficulty. Experiments on PM-VQA demonstrate that our proposed method achieves state-of-the-art performance, reaching an accuracy of **79.19%** and an F1-score of **78.74%**, which outperforms existing methods by **2.10%** in accuracy and **2.14%** in F1-score. Code and dataset will be available at <https://github.com/PixelOptic-Dev/Proto-CAP>.

Keywords: Pediatric Myopia · Medical Visual Question Answering · Multimodal Fusion.

1 Introduction

Pediatric myopia has emerged as an increasingly severe public health challenge globally, projected to affect approximately 50% of the world’s population by 2050[4]. Its accurate diagnosis and intervention rely heavily on the comprehensive multimodal analysis of Color Fundus Photography(CFP) and Optical Coherence Tomography (OCT). Medical Visual Question Answering, capable of interpreting medical imagery and providing interactive diagnostic evidence,

presents a promising solution to mitigate the shortage of ophthalmic medical resources[11,2]. However, the direct deployment of current Med-VQA in this field is constrained by a multitude of complex factors.

One major hurdle is the scarcity of domain-specific data. Existing public datasets (e.g., VQA-RAD[7] and SLAKE[12]) primarily target general radiology scenarios and lack fine-grained annotations specific to ophthalmic lesions. Such general-purpose data fails to support models in learning the complex correlations between subtle fundus changes and refractive status in pediatric myopia[21], thereby severely impeding the development of professional-grade diagnostic tools. Another is the challenge of optimizing deep models on small-scale medical data. In the absence of large-scale pre-training data, directly training deep models is highly prone to overfitting[18]. Although strategies such as sample re-weighting[17] and static polynomial loss[8] have alleviated this issue to some extent, they are either computationally expensive or lack adaptability to training dynamics, failing to strike an optimal balance between gradient boosting and hard example mining.

To address these challenges, we propose Proto-CAP, an enhanced framework designed specifically for high-performance Med-VQA. We first construct PM-VQA—the first large-scale, expert-verified dataset dedicated to pediatric myopia. To effectively utilize this data, we design a prototype fusion module that explicitly models medical concepts via a learnable memory bank, injecting question-guided prior knowledge into visual representations. Simultaneously, we introduce a Curriculum-Adaptive Poly-2 Loss, which dynamically adjusts curriculum difficulty based on the model’s convergence status.

The main contributions are summarized as follows:

1. We create a pediatric myopia medical visual question answering dataset called PM-VQA, which is the first expert-verified dataset specifically for pediatric myopia.
2. We propose the Proto-CAP framework, featuring two core innovations: a Prototype Memory Fusion module that bridges semantic gaps via learnable medical concepts, while a Curriculum-Adaptive Poly-2 Loss dynamically balances gradient boosting and hard example mining based on real-time proficiency.
3. Extensive experiments demonstrate that Proto-CAP significantly outperforms state-of-the-art Med-VQA methods across all metrics.

2 Dataset

2.1 Comparison with Existing Datasets

We performed a detailed comparison between PM-VQA and existing mainstream medical VQA datasets. As shown in Table 1, current datasets exhibit certain limitations. While SLAKE[12] and VQA-RAD[7] contain multimodal data and have undergone expert validation, they focus primarily on radiology rather than ophthalmology. Conversely, ophthalmology-related datasets such as

OphthalWeChat[22], DME-VQA[19], and INSEGCAT-VQA[20] mostly lack multimodal support. Furthermore, they are largely constructed from public datasets without rigorous QA validation, making it difficult to meet the requirements for multimodal data fusion analysis in clinical practice. To this end, we create PM-VQA, a rigorously expert-verified, multimodal dataset derived from clinical practice. With **2,741** images and **14,781** QA pairs, it fills the void for high-quality clinical ophthalmic VQA data.

Table 1. Comparison of existing medical VQA datasets in terms of scale, domain, modality, data source and QA verification.

Reference	Image Number	QA Pairs	Domain	Multi-modal	Data Source	QA Verification
SLAKE[12]	642	14028	Radiology	✓	Public datasets	✓
VQA-RAD[7]	315	3515	Radiology	✓	Public datasets	✓
OphthalWeChat[22]	3469	30120	Eye	✓	Public datasets	✗
Holland et al.[5]	41926	479710	Eye	✗	Clinical	✓
DME-VQA[19]	679	13470	Eye	✗	Public datasets	✗
INSEGCAT-VQA[20]	4647	39008	Eye	✗	Public datasets	✗
PM-VQA(ours)	2741	14781	Eye	✓	Clinical	✓

2.2 Data Collection and Construction Process

As illustrated in Fig. 1, the PM-VQA dataset construction involves three phases. First, we collected multimodal data—comprising Color Fundus Photography (CFP), Optical Coherence Tomography (OCT), and clinical records—from a collaborating hospital. Following rigorous quality screening and data anonymization, two ophthalmologists formulated clinical QA pairs, which were subsequently validated by a senior expert to ensure medical accuracy. Finally, the processed data was formatted into JSON and partitioned into **10,271** training, **2,287** validation, and **2,223** testing QA pairs. This study and the use of all data were approved by the ethics committee of the Tianjin Medical University Eye Hospital [No. 2024KY-67].

3 Method

We propose a Proto-CAP enhanced framework for Medical Visual Question Answering. As illustrated in Fig. 2, this framework is designed to mimic the clinical cognitive pattern where physicians simultaneously consider multiple correlated questions during diagnosis. The model comprises three core components: a multimodal backbone based on the Medical Multi-Question Learning (MMQL)[1], a Prototype Memory Fusion module, and the Curriculum-Adaptive Poly-2 Loss function.

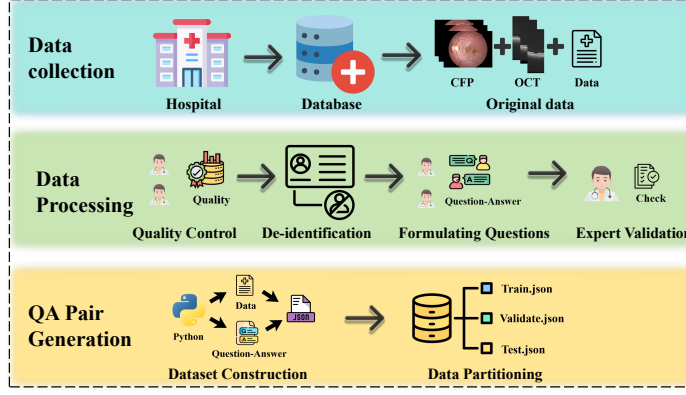


Fig. 1. Construction pipeline of the PM-VQA dataset. The process consists of three stages: Data Collection, Data Processing, and QA Pair Generation.

3.1 Medical Multi-Question Learning Backbone

To capture correlations among questions, we adopt the Medical Multi-Question Learning (MMQL)[1], consolidating all questions Q_k related to an image v_k into a unified sequence T for joint reasoning. The concatenation is formulated as follows:

$$T = q_1 \circ q_2 \circ \dots \circ q_i \circ a_i \circ \dots \circ q_{|Q_k|} \quad (1)$$

where $\circ$ denotes the concatenation operation. During training, answers are randomly replaced with [MASK] tokens based on a probability p . Remaining ground-truth answers serve as context prompts, guiding the model to infer diagnoses from known information.

The sequence T is encoded by a pre-trained BERT-Base[6] to extract textual embeddings $Q_{emb} \in \mathbb{R}^{L \times D}$. Concurrently, the image I is processed by a Swin Transformer[14] to obtain visual features $V \in \mathbb{R}^{49 \times D}$. Before multimodal fusion, these features are enhanced by the Prototype Memory Fusion module to improve the representation of clinical concepts.

3.2 Prototype Memory Fusion

While MMQL provides inter-question context, the model may still lack domain-specific medical knowledge across samples. To address this limitation, we design a **Prototype Fusion** module positioned prior to the Transformer fusion stage, serving to enhance the visual features.

We maintain a learnable memory matrix $M \in \mathbb{R}^{K \times D}$, where K denotes the number of medical concept prototypes. To retrieve the medical knowledge most relevant to the current multi-question sequence T , we first apply global average pooling to the textual embeddings Q_{emb} to obtain a global vector q_{global} .

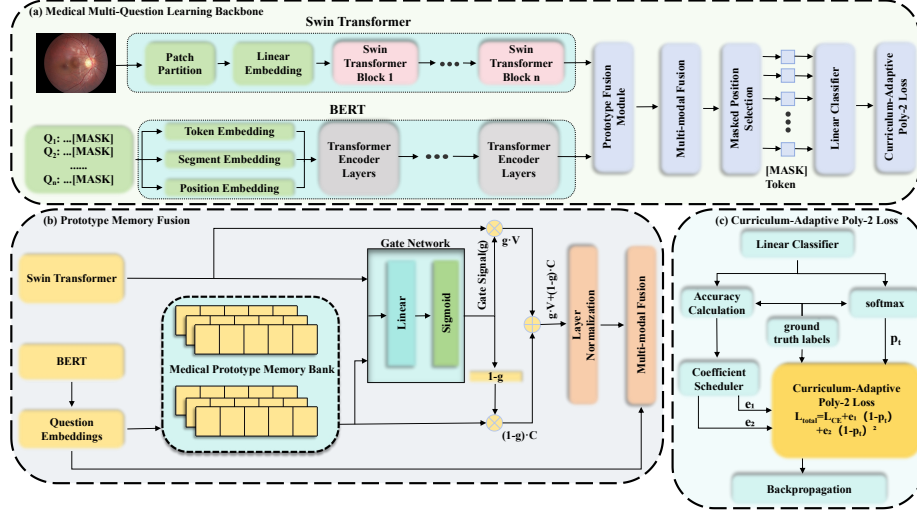


Fig. 2. (a) The MMQL backbone encodes images and masked question sequences via Swin Transformer and BERT for joint prediction. (b) Prototype Memory Fusion injects retrieved concept features C into visual features V via a gating mechanism defined as $g \cdot V + (1 - g) \cdot C$. (c) Curriculum-Adaptive Poly-2 Loss dynamically adjusts coefficients e_1 and e_2 based on training accuracy to balance sample difficulty.

This vector serves as the query to compute attention scores against the memory matrix:

$$Attn_{proto} = \text{Softmax}((q_{global} W_q) \cdot M^T) \quad (2)$$

Based on these attention weights, we aggregate the prototypes to yield the concept features C that best align with the current context.

The retrieved concept features are subsequently broadcasted and injected into the original visual features V . We employ a Gating Mechanism to dynamically regulate the injection intensity, ensuring the preservation of intrinsic image features while incorporating external knowledge:

$$g = \sigma(W_g \cdot [V; C]) \quad (3)$$

$$V_{enhanced} = \text{LayerNorm}(g \cdot V + (1 - g) \cdot C) \quad (4)$$

Finally, the enhanced visual features $V_{enhanced}$ replace the original V and are fed into the main Transformer alongside the textual embeddings Q_{emb} for final interaction and prediction.

3.3 Curriculum-Adaptive Poly-2 Loss

During the training process of Medical VQA, the learning state of the model evolves dynamically. Traditional loss functions typically rely on fixed param-

ters, lacking sensitivity to the model’s actual current capabilities. To address this limitation, we refine the polynomial loss framework and propose a **Curriculum-Adaptive Poly-2 Loss**. This loss function eliminates the dependency on static parameters by dynamically modulating polynomial coefficients based on current prediction accuracy, thereby facilitating self-paced learning. The total loss function is defined as a polynomial form based on Taylor expansion:

$$L_{total} = L_{CE} + e_1 \cdot (1 - P_t) + e_2 \cdot (1 - P_t)^2 \quad (5)$$

L_{CE} is the Cross-Entropy loss and P_t is the target prediction probability. The coefficients e_1 and e_2 dynamically adapt to the model’s current proficiency: e_1 follows a cosine decay to accelerate initial convergence, while e_2 increases linearly to emphasize hard samples (low P_t) during later training stages.

Through this mechanism, when the model exhibits low accuracy, the loss function prioritizes the correction dominated by the e_1 term to facilitate rapid convergence. Conversely, as the model achieves high accuracy, the influence of the e_2 term is amplified, compelling the model to focus on hard samples with lower prediction probabilities, thereby further enhancing generalization capability.

4 Experiment

4.1 Implementation Details

In the experiments, the input images are resized to 224×224 to match the input requirements of the Swin Transformer. We adopt Swin Transformer[14] and BERT[6] pre-trained on ImageNet-1K and Wikipedia, respectively, as the visual and textual backbones. We use the AdamW[15] optimizer with differential learning rates to fine-tune different components: the learning rate is set to 5×10^{-6} for the BERT encoder, 1×10^{-5} for the Swin Transformer and the fusion Transformer, and initialized at 1×10^{-4} for the classifier and other randomly initialized parameters. The proposed Curriculum-Adaptive Poly-2 Loss is employed as the objective function. Training is conducted for 80 epochs with a batch size of 16. Our framework was implemented using PyTorch, and all experiments were conducted on a single Nvidia RTX A6000 GPU.

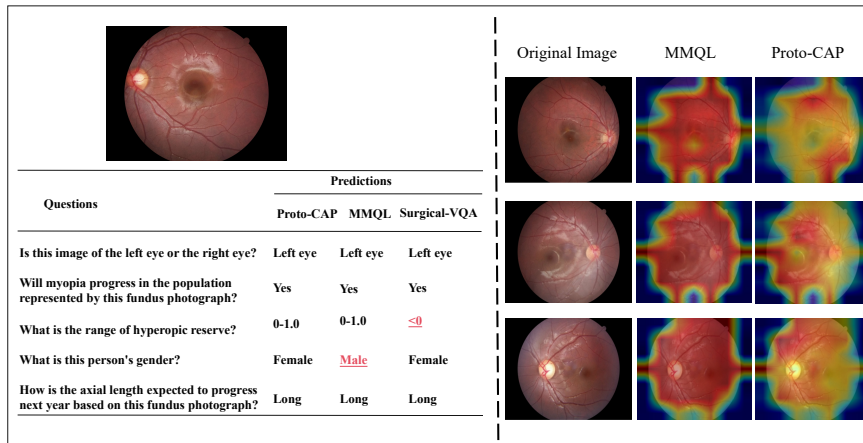
4.2 Comparisons With Other Methods

To evaluate the effectiveness of our proposed Proto-CAP, we conducted comprehensive experiments and compared it with several methods in the medical visual question answering domain. The baseline models include CMSA[3], M-Mixup[13], M2I2[10], MUMC[9], MEVF-BAN[23], Surgical-VQA[16], and MMQL[1].

We evaluate the performance using four standard metrics: Accuracy, Precision, Recall, and F1-score. The quantitative comparison results on the PM-VQA dataset are presented in Table 2. Our Proto-CAP achieves the state-of-the-art performance with an Accuracy of 79.19% and an F1-score of 78.74%. Specifically,

Table 2. Comparison of performance metrics on the PM-VQA .

Method	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
CMSA[3]	50.07	39.49	50.07	41.15
M-Mixup[13]	53.58	30.65	53.58	38.68
M2I2[10]	57.76	48.79	57.76	51.76
MUMC[9]	63.52	62.06	63.52	61.86
MEVF-BAN[23]	65.76	65.82	65.76	65.48
Surgical-VQA[16]	69.09	71.77	69.09	67.48
MMQL[1]	77.09	78.35	77.09	76.60
Proto-CAP(ours)	79.19	79.41	79.19	78.74

**Fig. 3.** The left panel shows the prediction results of different models on pediatric myopia-related questions. The right panel displays the visualization of attention heatmaps, comparing the interpretability of MMQL and Proto-CAP.

it surpasses the second-best method, MMQL, by a significant margin of 2.10% in Accuracy and 2.14% in F1-score.

Furthermore, we provide a qualitative comparison to intuitively demonstrate the superiority of our model. As illustrated in Fig. 3, the left panel shows specific question-answering cases on a pediatric myopia fundus image. Proto-CAP successfully answers all questions correctly, whereas MMQL struggles with demographic features (misclassifying the patient's gender as "Male") and Surgical-VQA fails on fine-grained clinical assessments (predicting the hyperopic reserve as "<0" instead of "0-1.0"). The right panel of Fig. 3 visualizes the attention heatmaps. Compared to the original MMQL, Proto-CAP generates more concentrated and precise focal areas around the key anatomical structures .

The remarkable reasoning capability of Proto-CAP stems fundamentally from its mechanistic breakthroughs. While the traditional MMQL excels at capturing

inter-question dependencies, its lack of medical priors hinders the alignment of localized visual anomalies with deep clinical semantics. Addressing this bottleneck, our Prototype Memory Fusion module successfully bridges the vision-language semantic gap by explicitly injecting parameterized ophthalmic concepts into the visual space. This innovation forces the network to move beyond fitting superficial image features and pivot towards authentic pathology-aware reasoning. Consequently, the model’s attention dynamically converges on the most diagnostically valuable clinical regions, ultimately generating predictions that are not only accurate but also fully consistent with ophthalmic clinical logic.

4.3 Ablation Studies

To evaluate each component’s contribution, we conducted ablation studies on the PM-VQA dataset using the MMQL backbone as the baseline. The quantitative results are summarized in Table 3. Integrating the Prototype Memory Fusion (+Proto) increases Accuracy by 1.18% and F1-score by 0.98%, demonstrating that dynamically injecting medical concepts enhances feature representation. Applying the Curriculum-Adaptive Poly-2 Loss (+CAP) alone yields a 0.65% Accuracy and 0.74% F1-score improvement, validating its ability to dynamically balance optimization between hard and easy samples. Finally, combining both modules (+Proto-CAP) achieves the best overall performance (79.19% Accuracy, 78.74% F1-score), outperforming the baseline by 2.10% and 2.14%, respectively. This highlights their strong complementarity: the Proto module provides enriched semantic features, while the Curriculum-Adaptive Poly-2 Loss ensures they are optimally learned.

Table 3. Ablation study on the proposed Proto and CAP modules.

Method	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
baseline	77.09	78.35	77.09	76.60
+Proto	78.27	77.63	78.27	77.58
+CAP	77.74	77.50	77.74	77.34
+Proto-CAP	79.19	79.41	79.19	78.74

5 Conclusion

In this paper, we propose Proto-CAP, a novel Med-VQA framework for pediatric myopia assessment. Built on a Medical Multi-Question Learning backbone, it simultaneously processes multiple related clinical inquiries. Proto-CAP introduces two key innovations: a Prototype Memory Fusion module that integrates learnable medical concepts into visual features, and a Curriculum-Adaptive Poly-2 loss that dynamically adapts to sample difficulty during training. Extensive experiments on the PM-VQA dataset demonstrate that Proto-CAP significantly

outperforms state-of-the-art models, establishing a robust baseline for pediatric ophthalmology and broader computer-aided diagnosis.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 62272248), the Natural Science Foundation of Tianjin (No. 23JCZDJC01010), the Tianjin Key Medical Discipline Construction Project (No. TJYXZDXK-3-004A-2) and Haihe Lab of ITAI Science and Technology Project (No. 25HHXCSS00007).

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

References

1. Chen, Q., Bian, M., Xu, H.: Mmql: Multi-question learning for medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 480–489. Springer (2024)
2. Eslami, S., De Melo, G., Meinel, C.: Does clip benefit visual question answering in the medical domain as much as it does in the general domain? arXiv preprint arXiv:2112.13906 (2021)
3. Gong, H., Chen, G., Liu, S., Yu, Y., Li, G.: Cross-modal self-attention with multi-task pre-training for medical visual question answering. In: Proceedings of the 2021 international conference on multimedia retrieval. pp. 456–460 (2021)
4. Holden, B.A., Fricke, T.R., Wilson, D.A., Jong, M., Naidoo, K.S., Sankaridurg, P., Wong, T.Y., Naduvilath, T.J., Resnikoff, S.: Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050. *Ophthalmology* **123**(5), 1036–1042 (2016)
5. Holland, R., Taylor, T.R., Holmes, C., Riedl, S., Mai, J., Patsiamanidi, M., Mitsopoulou, D., Hager, P., Müller, P., Paetzold, J.C., et al.: Specialized curricula for training vision language models in retinal image analysis. *NPJ Digital Medicine* **8**(1), 532 (2025)
6. Kenton, J.D.M.W.C., Toutanova, L.K., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. vol. 1, pp. 4171–4186. Minneapolis, Minnesota (2019)
7. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. *Scientific data* **5**(1), 1–10 (2018)
8. Leng, Z., Tan, M., Liu, C., Cubuk, E.D., Shi, X., Cheng, S., Anguelov, D.: Polyloss: A polynomial expansion perspective of classification loss functions. arXiv preprint arXiv:2204.12511 (2022)
9. Li, P., Liu, G., He, J., Zhao, Z., Zhong, S.: Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 374–383. Springer (2023)
10. Li, P., Liu, G., Tan, L., Liao, J., Zhong, S.: Self-supervised vision-language pre-training for medial visual question answering. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)

11. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z.: Medical visual question answering: A survey. *Artificial Intelligence in Medicine* **143**, 102611 (2023)
12. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 1650–1654. IEEE (2021)
13. Liu, L., Su, X.: How well apply multimodal mixup and simple mlps backbone to medical visual question answering? In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2648–2655. IEEE (2022)
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
16. Seenivasan, L., Islam, M., Krishna, A.K., Ren, H.: Surgical-vqa: Visual question answering in surgical scenes using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 33–43. Springer (2022)
17. Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., Meng, D.: Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems* **32** (2019)
18. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical image analysis* **63**, 101693 (2020)
19. Tascon-Morales, S., Márquez-Neila, P., Sznitman, R.: Consistency-preserving visual question answering in medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 386–395. Springer (2022)
20. Tascon-Morales, S., Márquez-Neila, P., Sznitman, R.: Localized questions in medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 361–370. Springer (2023)
21. Ting, D.S., Peng, L., Varadarajan, A.V., Keane, P.A., Burlina, P.M., Chiang, M.F., Schmetterer, L., Pasquale, L.R., Bressler, N.M., Webster, D.R., et al.: Deep learning in ophthalmology: the technical and clinical considerations. *Progress in retinal and eye research* **72**, 100759 (2019)
22. Xu, P., Gong, X., Chen, X., Zhang, W., Yang, J., Yan, B., Yuan, M., Zheng, Y., He, M., Shi, D.: Benchmarking large multimodal models for ophthalmic visual question answering with ophthalwechat. arXiv preprint arXiv:2505.19624 (2025)
23. Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M.: Medical visual question answering via conditional reasoning. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2345–2354 (2020)