Saliency Guided Longitudinal Medical Visual Question Answering

Jialin Wu
Dept. of Computer Science and Engineering
University of California, San Diego
San Diego, CA 92037
jlwu@ucsd.edu

Xiaofeng Liu
Dept. of Radiology and Biomedical Imaging
Yale University
New Haven, CT 06510
xiaofeng.liu@yale.edu

Abstract

Longitudinal medical visual question answering (Diff-VQA) requires comparing paired studies from different time points and answering questions about clinically meaningful changes. In this setting, the difference signal and the consistency of visual focus across time are more informative than absolute single-image findings. We propose a saliency-guided encoder–decoder for chest X-ray Diff-VQA that turns post-hoc saliency into actionable supervision. The model first performs a lightweight near-identity affine pre-alignment to reduce nuisance motion between visits. It then executes a within-epoch two-step loop: step 1 extracts a medically relevant keyword from the answer and generates keyword-conditioned Grad-CAM on both images to obtain disease-focused saliency; step 2 applies the shared saliency mask to both time points and generates the final answer. This closes the language–vision loop so that the terms that matter also guide where the model looks, enforcing spatially consistent attention on corresponding anatomy. On Medical-Diff-VQA, the approach attains good performance on BLEU, ROUGE-L, CIDEr, and METEOR while providing intrinsic interpretability. Notably, the backbone and decoder are general-domain pretrained without radiology-specific pretraining, highlighting practicality and transferability. These results support saliency-conditioned generation with mild pre-alignment as a principled framework for longitudinal reasoning in medical VQA.

1 Introduction

Medical Visual Question Answering (VQA) aims to answer open-ended clinical questions based on medical images, serving as a critical bridge from visual perception to clinical decision support [1]. Numerous medical VQA approaches in recent years have relied on pretrained visual or multimodal models [2, 3, 4]. However, most of these works focus on a single time-point following the natural image VQA tasks. Radiologists routinely compare current and prior studies to localize change, judge progression, and reconcile apparent discrepancies.

Difference/longitudinal visual question answering (Diff-VQA) operationalizes this workflow by conditioning answers on paired images acquired at two time points, where the difference is often the signal of interest rather than absolute appearance [5]. Recent benchmarks and methods for longitudinal chest X-rays have made this task concrete by supplying paired images, questions, and change-focused references [6, 7, 8]. Building on these resources, several approaches adapt vision–language models or design task-specific architectures to better capture temporal discrepancies, including prior work that emphasizes longitudinal pretraining [9], residual alignment in the feature or pixel space [10], or region-level retrieval and mixing [11]. However, their attention in different

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

time-points is not explicitly encouraged to be consistent, which is essential for the compare and contrast to explore the difference.

Saliency maps are a type of saliency visualization used to interpret deep learning models. In medical imaging tasks, they are widely employed to present verifiable evidence to clinicians and enhance model interpretability and trustworthiness [12]. In medical VQA tasks, researchers frequently employ attention/saliency visualizations to verify whether models focus on relevant image evidence when generating responses, reflecting the critical need for explainable and traceable reasoning processes in high-stakes medical contexts [1]. However, existing medical-VQA models often treat saliency as a post-hoc explanation [13, 14, 15] rather than incorporating it as intrinsic supervision during training. In longitudinal settings this is a missed opportunity because consistent focus on corresponding anatomy across the two time points is essential to answering difference-type questions faithfully.

We introduce a saliency-guided longitudinal VQA framework that makes saliency actionable during learning. The method has two design principles: (i) make the two images geometrically comparable and (ii) ensure that what the model says it cares about also determines where it looks at both time points, inspired by natural image co-attention [16, 17]. Specifically, we have two modules. • **Micro pre-alignment.** A lightweight CNN-based module applies a near-identity affine warp to the current study to mitigate small pose and scale variations without overfitting or erasing true changes [18]. • **Keyword-conditioned shared saliency.** Within each training epoch we run a two-step loop akin to the Expectation Maximization Algorithm [19]. First, a language model extracts one clinically salient keyword from the ground-truth answer. We compute keyword-conditioned Grad-CAM on both time points and take their union to form a shared saliency mask. Second, we re-encode masked images and generate the answer with a multimodal decoder, which ties linguistic supervision to spatial evidence and encourages consistent attention on the same anatomical regions across time.

The main contributions can be summarized as:

•We formalize a simple and effective way to enforce *spatially consistent attention* across paired images by using keyword-conditioned, shared saliency as a training signal for Diff-VQA.

•We couple this with a minimal pre-alignment module that improves longitudinal comparability while preserving true differences.

•We demonstrate competitive results on Medical-Diff-VQA using only general-domain pretrained backbones and decoders, yielding strong practicality and inherent interpretability without radiology-specific pretraining.

2 Methods

We use the longitudinal chest radiograph Diff-VQA dataset, Medical-Diff-VQA [5, 6], which constructs samples from paired studies of the same subject at two time points together with a difference-focused question–answer pair. The dataset is derived from MIMIC-CXR [20] and MIMIC-CXR-JPG [21] and was obtained from PhysioNet [22]. All usage follows the PhysioNet credentialed-access license and de-identification guidelines. The corpus contains 164,223 samples, split into 131,556 for training, 16,278 for validation, and 16,389 for testing.

An overview appears in Figure 1. The pipeline has two components: a micro image registration module and a keyword-conditioned saliency extraction module, followed by image–text encoders and a multimodal decoder.

2.1 Micro Image Registration Module

Given a main image $I_{\text{main}} \in \mathbb{R}^{3\times H\times W}$ and a reference image $I_{\text{ref}} \in \mathbb{R}^{3\times H\times W}$, a shallow CNN predicts 2D affine parameters $\Theta = [A\ \mathbf{t}] \in \mathbb{R}^{2\times 3}$. We warp only the main image with a differentiable grid sampler:

$$\mathbf{x} = A\,\mathbf{x}_{\text{tgt}} + \mathbf{t}.$$

To keep the transform near identity and avoid erasing true anatomical change, we regularize

$$\mathcal{L}_{\text{reg}} = w_{\text{small}}\|\Theta - I\|_F^2 + w_{\text{det}}(\det(A) - 1)^2 + w_{\text{trans}}\|\mathbf{t}\|_2^2,$$

with $w_{\text{small}} = 10^{-4}$, $w_{\text{det}} = 10^{-5}$, and $w_{\text{trans}} = 10^{-6}$. The registered image is $\widehat{I}_{\text{main}}$.

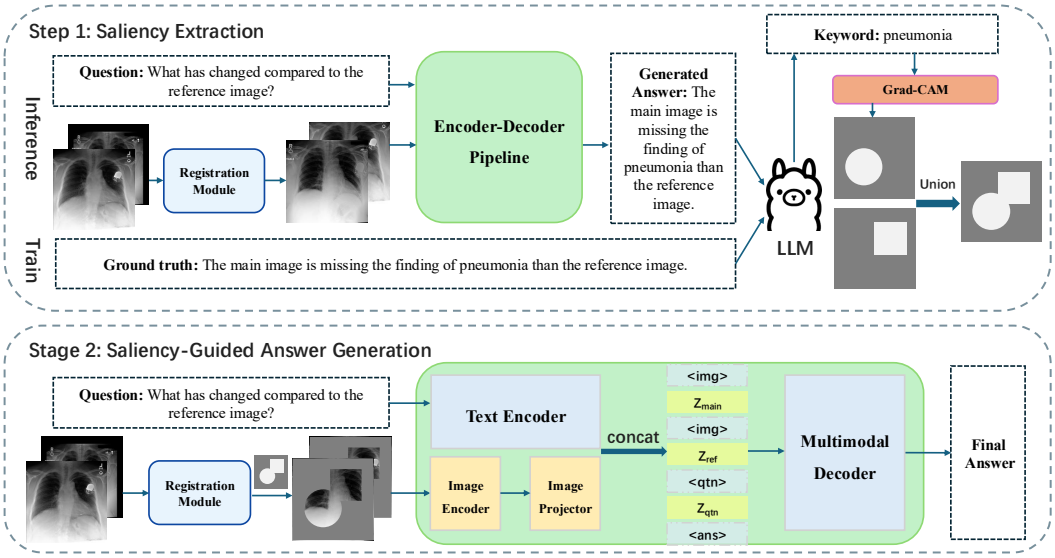

Figure 1: Overview of our proposed method. The step 1 is for saliency extraction, and step 2 is for answer generation. Saliency extraction and application are based on the after-registration images. Grad-CAM uses the keyword as the explanation target. Z_{main}, Z_{ref}, Z_{qtn} are extracted representations for the main image, reference image and question. $\langle\cdot\rangle$ denotes special tokens for separation.

2.2 Saliency Extraction Module

We compute gradient-based saliency on the registered pair using Grad-CAM [23, 24]. A single clinical keyword is extracted from the answer using Llama 3:70B [25, 26] and used as the target concept for saliency. We obtain maps S_{main} and S_{ref} on $\widehat{I}_{\text{main}}$ and I_{ref}, respectively, then form a shared mask by element-wise maximum, $S = \max(S_{\text{main}},\ S_{\text{ref}})$. After min–max normalization of S to $[0, 1]$, we apply it multiplicatively to both images:

$$I'_{\text{main}} = S \odot \widehat{I}_{\text{main}}, \qquad I'_{\text{ref}} = S \odot I_{\text{ref}}.$$

This encourages consistent focus on corresponding anatomy at both time points while retaining non-salient context with attenuated weight.

2.3 Encoders and Decoders

Image encoder and projector. We use ResNet-50 [27] pretrained on ImageNet-1k [28]. From its penultimate feature map $\mathbb{R}^{H\times W\times C}$ we form a token sequence $\mathbb{R}^{N\times C}$ with $N = HW$. A projector with one linear layer, one 8-head Transformer encoder [29], and a two-layer MLP maps image tokens to the text-representational space.

Text encoder. Questions are tokenized with embeddings shared by the decoder, then added with a learnable positional embedding and passed through six 12-head Transformer encoder layers.

Multimodal decoder. A GPT-2 [30] decoder from HuggingFace [31] consumes masked image tokens and question tokens to generate the answer. We add special tokens $\langle$pad$\rangle$, $\langle$img$\rangle$, $\langle$qtn$\rangle$, and $\langle$ans$\rangle$. Denote the main image representation as Z_{main}, the reference image representation as Z_{ref}, and the question representation as Z_{qtn} .The input sequence is

$$\textbf{concat}(\langle\text{img}\rangle, Z_{\text{main}}, \langle\text{img}\rangle, Z_{\text{ref}}, \langle\text{qtn}\rangle, Z_{\text{qtn}}, \langle\text{ans}\rangle).$$

2.4 Training and Inference

Training. During training, the ground-truth answer provides the keyword for saliency. The total loss is the sum of registration and language modeling terms, $\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{LM}}$. We run a 1-epoch

warm-up without saliency masking, then enable the two-step loop for the remaining epochs of a 16-epoch schedule.

Inference. We use a two-pass procedure. First, the decoder generates a preliminary answer without masking, from which we extract a keyword. Second, we compute keyword-conditioned saliency on both images, apply the shared mask, and regenerate the final answer using the masked inputs. For latency-sensitive use, a single-pass variant without masking is available, but the two-pass variant better enforces longitudinal consistency.

3 Results

We adopt common generation metrics in VQA, BLEU-1/2/3/4 [32] (n-gram precision with a brevity penalty), METEOR [33] (stem matching with an emphasis on recall), ROUGE-L [34] (overlap and longest common subsequence), and CIDEr [35] (a TF-IDF based consensus metric) to evaluate different aspects such as surface-level matching, semantic alignment, and consistency with human references. To emphasize the medical keyword and semantic meaning, also considering the scale of the metrics, we combine METEOR and CIDEr as

$$0.6 \cdot \frac{\text{CIDEr}}{1 + \text{CIDEr}} + 0.4 \cdot \text{METEOR},$$

to select the model that performs the best at the end of training.

Table 1: Evaluation on Medical-Diff-VQA comparing ours with prior works.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
MCCFormers [36]	0.214	0.190	0.170	0.153	0.319	0.340	0
IDCPCL [37, 10]	0.614	0.541	0.474	0.414	0.303	0.582	0.703
EKAID [5, 10]	0.628	0.553	0.491	0.434	0.339	0.557	1.027
RegioMix [11]	0.705	0.633	0.572	0.517	0.381	0.651	1.804
PLURAL [9]	0.704	0.633	0.575	0.520	0.381	0.653	1.832
ReAl [10]	0.710	0.636	0.580	0.530	0.395	0.736	2.409
VED [38]	0.716	0.647	0.590	0.537	0.389	0.670	2.119
Ours	0.628	0.510	0.418	0.341	0.651	0.627	1.263

Table 1 shows that, although our BLEU-1/2/3/4 scores are modest, they still indicate non-trivial lexical precision: the model reliably reproduces key clinical tokens (e.g., anatomy, laterality, lesion attributes) across paired studies rather than collapsing to generic templates. Conversely, METEOR is notably high (0.651), evidencing strong synonym and inflectional coverage, while ROUGE-L (0.627) and CIDEr (1.263) suggest that the generated answers preserve sentence-level coverage and place appropriate emphasis on TF-IDF-informative, clinically salient phrases. Together, these metrics indicate that saliency guidance and keyword-conditioned targets help the decoder to provide a focus on disease-bearing regions/terms, resulting in good semantic adequacy with more conservative n-gram precision.

The overall room for improvement is understandable. First, none of our components were adapted to medical data: the general-purpose image backbone and text decoder were used off-the-shelf without further pretraining on radiology images and texts, and the keyword extractor relied on a general-purpose LLM rather than a domain-specialized model. Second, the current checkpoint was trained for 16 epochs, which likely limited convergence of our model; longer training (with an extended warm-up before enabling saliency) could benefit n-gram precision. Third, our registration is intentionally lightweight and near-identity, which may under-correct inter-visit motion differences; such residual misalignment can depress BLEU-n while still allowing METEOR/CIDEr to reflect correct semantics.

4 Acknowledgment

This work was partially supported by NIH R21EB034911.

References

[1] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. 143:102611.

[2] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3922–3931, 2021.

[3] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023.

[4] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering, 2024.

[5] Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4156–4165, New York, NY, USA, 2023. Association for Computing Machinery.

[6] Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, liangchen liu, Kazuma Kobayashi, Tatsuya Harada, Ronald Summers, and Yingying Zhu. Medical-Diff-VQA: A Large-Scale Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images.

[7] Dong Yul Oh, Jihang Kim, and Kyong Joon Lee. Longitudinal change detection on chest x-rays using geometric correlation maps. page 748–756, Berlin, Heidelberg, 2019. Springer-Verlag.

[8] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. A clinically accessible small multimodal radiology model and evaluation metric for chest X-ray findings. 16(1):3108.

[9] Yeongjae Cho, Taehee Kim, Heejun Shin, Sungzoon Cho, and Dongmyung Shin. Pretraining vision-language model for difference visual question answering in longitudinal chest x-rays, 2024.

[10] Zilin Lu, Yutong Xie, Qingjie Zeng, Mengkang Lu, Qi Wu, and Yong Xia. Spot the Difference: Difference Visual Question Answering with Residual Alignment . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland, October 2024.

[11] Ka-Wai Yung, Jayaram Sivaraj, Danail Stoyanov, Stavros Loukogeorgakis, and Evangelos B. Mazomenos. Region-Specific Retrieval Augmentation for Longitudinal Visual Question Answering: A Mix-and-Match Paradigm . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland, October 2024.

[12] Yusuf Brima and Marcellin Atemkeng. Saliency-driven explainable deep learning in medical imaging: Bridging visual explainability and statistical quantitative analysis. 17(1):18.

[13] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. One map does not fit all: Evaluating saliency map explanation on multi-modal medical images, 2021.

[14] Ricardo Bigolin Lanfredi, Ambuj Arora, Trafton Drew, Joyce D. Schroeder, and Tolga Tasdizen. Comparing radiologists' gaze and saliency maps generated by interpretability methods for chest x-rays, 2023.

[15] Jianxun Lou, Huasheng Wang, Xinbo Wu, John Cho Hui Ng, Richard White, Kaveri A. Thakoor, Padraig Corcoran, Ying Chen, and Hantao Liu. Chest x-ray visual saliency modeling: Eye-tracking dataset and saliency prediction model. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):16920–16930, 2025.

[16] Qizao Wang, Xuelin Qian, Yanwei Fu, and Xiangyang Xue. Co-attention aligned mutual cross-attention for cloth-changing person re-identification. page 351–368, Berlin, Heidelberg, 2022. Springer-Verlag.

[17] Guangshuai Gao, Wenting Zhao, Qingjie Liu, and Yunhong Wang. Co-saliency detection with co-attention fully convolutional network, 2020.

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks, 2016.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[20] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. 6(1):317.

[21] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019.

[22] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220, 2000.

[23] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019.

[24] Jacob Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

[25] AI@Meta. Llama 3 model card. `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`, 2024.

[26] Ollama/ollama: Get up and running with OpenAI gpt-oss, DeepSeek-R1, Gemma 3 and other models.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. pages 770–778.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[31] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[33] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[34] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[35] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015.

[36] Kodai Nakashima Ryota Suzuki Kenji Iwata Hirokatsu Kataoka Yue Qiu, Shintaro Yamamoto and Yutaka Satoh. Describing and localizing multiple changes with transformers, 2021.

[37] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning, 2018.

[38] Luis-Jesus Marhuenda, Miquel Obrador-Reina, Mohamed Aas-Alas, Alberto Albiol, and Roberto Paredes. Unveiling differences: A vision encoder-decoder model for difference medical visual question answering. In *Medical Imaging with Deep Learning*, 2025.

A Appendix and Discussions

This paper proposes a generative framework for longitudinal medical image differential question-answering. It combines near-identity affine registration, dual-image encoding, and a text-image generation decoder to establish a closed-loop process from keywords to saliency (CAM) to feature weighting. This ensures consistent alignment between linguistic cues and visual evidence during both training and inference phases, thereby enhancing the localizability of lesion-related regions and the interpretability of answers. It is crucial to emphasize that all backbone models and decoders in this work are pre-trained on general-purpose datasets without further training on medical data. The LLM used for keyword extraction is also a general-purpose model, further highlighting the method's transferability and engineering feasibility.

Despite the aforementioned progress, this study has limitations. The additional saliency extraction step introudces large time complexity. When keyword extraction relies on general-purpose LLMs rather than medically specialized pre-trained models, it may omit critical lesion terminology or introduce overly generic vocabulary, thereby weakening visual alignment effectiveness. Additionally, the currently employed affine registration is relatively simple and may struggle to accommodate drastic variations in pose and imaging conditions. Apart from that, since current data and tasks focus on the difference problem within the MIMIC framework, cross-dataset generalization remains to be validated. Furthermore, the absence of medical retraining may limit the upper performance bound. Finally, as this is an ongoing study, further exploration is needed regarding the impact of additional ablation studies, keyword extraction and saliency method selection, and larger encoders on model performance.

Despite these limitations, this study holds substantial significance and value. We pioneered the conversion of medically relevant keywords automatically extracted by general-purpose LLMs into saliency targets for feature weighting. This establishes a closed-loop mechanism where language supervision directly constrains visual attention, systematically enhancing evidence utilization and model interpretability without requiring medical pre-training. Given that medical question-answering is often driven by lesion nouns and key anatomical locations, these keywords provide sparse yet strongly constrained supervisory signals, demonstrating both necessity and novelty. The comprehensive workflow provides a reusable baseline and engineering paradigm for future extensions to stronger visual backbones, more sophisticated deformation models, and larger language models.