Contrastive Answer Calibration for Medical Visual Question Answering

Alireza Kheyrkhah[1], Amir Hossein Saleknia[1], Reza Azad[2], Ulas Bagci[3], and Alaa Sulaiman[4]

[1]Independent Researcher, [2]University of Regensburg, Germany, [3]Northwestern University, USA, [4]Abdora INC, USA
alaa@abdora.ai

Abstract. Generative medical visual question answering (VQA) models built on masked-pretrained encoder-decoder backbones are trained to maximise the likelihood of a single ground-truth answer, yet at inference they are deployed as discriminators that rank a fixed list of candidate answers by conditional generation likelihood. This mismatch between the training and inference objectives is rarely examined, even though it is precisely where probability mass leaks onto plausible but incorrect answers. Maximum-likelihood training raises the score of the gold answer but leaves the scores of confusable competitors unconstrained, so a fluent decoder can rank a clinically incorrect alternative above the truth. We introduce Answer Contrastive Calibration (ACC), an auxiliary fine-tuning objective that closes this gap directly in the ranking score space. For every question in a mini-batch, ACC scores all batch answers under the decoder using the same summed token log-likelihood employed by the inference-time ranker and applies a contrastive loss that drives the gold answer above its in-batch competitors, which serve as semantically meaningful hard negatives. The objective adds no data, no parameters, and no external answer bank; it is strictly additive and weight-gated, so a weight of zero recovers the unmodified backbone exactly, and every reported gain is a controlled effect of the objective alone. We evaluate ACC on three standard medical VQA benchmarks, building on a faithful re-implementation of a strong masked-pretrained backbone. Across all three, it reaches the state-of-the-art level, with its clearest gains on VQA-RAD and PathVQA, whose large and confusable answer spaces make calibration the bottleneck. https://github.com/abdora-ai/ACC-MedVQA.

Keywords: Medical visual question answering · Answer calibration · Contrastive learning · Ranking-based inference · Vision-language models

1 Introduction

Medical VQA requires a clinically correct answer to a free-form question about a medical image, a demanding testbed for vision-language understanding in healthcare [1]. The setting is defined by scarcity: VQA-RAD [2], SLAKE [3], and PathVQA [4] contain only a few thousand image-question-answer triples

across radiograph and pathology, mixing closed yes/no decisions with a long tail of open-ended findings. This favors transfer learning, and the dominant recipe – pretraining a vision-language encoder on medical image-text pairs with masked and contrastive objectives, then fine-tuning per dataset – defines the strongest baselines on all three, from early meta-learning against data scarcity [5,6] to modern contrastively aligned transformer encoders [7].

Pretraining has received most attention: inspired by ALBEF [8] and CLIP [9], medical systems invest in the pretraining corpus, masking strategy, and unimodal and multimodal contrastive losses aligning the encoders [10]; foundation models such as BLIP-2 [11], LLaVA-Med [12], and HuatuoGPT-Vision [13] extend this to instruction-tuned multimodal LLMs, with calibration and preference alignment [14,15], retrieval augmentation [16], RL tuning [17], and lightweight task adaptation [18]. Fine-tuning, by contrast, is usually an afterthought: the encoder is paired with a generative text decoder and optimised with standard maximum likelihood on the gold answer, exactly as in image captioning [19] – leaving unexamined whether it suits how they are evaluated.

We argue it does not, for the dominant class of generative medical VQA systems built on masked-pretrained encoder-decoder backbones and evaluated by candidate ranking [10,7]. They are trained as language models, maximising the single reference answer's likelihood token by token, yet nearly all are evaluated as rankers, not open-ended generators, scoring every answer in a predefined candidate list by conditional generation likelihood and returning the most likely one [10,7], so training and inference optimise different objectives. Maximum likelihood only pushes probability *up* on the gold answer; it never tells the model that a clinically incorrect but fluent alternative – "left" instead of "right", a neighbouring organ, or a co-occurring finding – should receive *lower* likelihood, so a strong model can score several competing answers highly and rank the wrong one first. We call this the training-inference gap: reproducing a strong masked-pretrained backbone with the original code, checkpoint, and splits, its largest deficit falls exactly where the gap predicts while closed binary questions are handled well, so the residual error is largely a calibration rather than representation problem [20], addressable downstream without revisiting pretraining.

We therefore propose Answer Contrastive Calibration (ACC), a lightweight fine-tuning objective aligning training with how the model is scored: for each question we score every mini-batch answer under the decoder using the evaluator's own test-time summed token log-likelihood, and a contrastive loss drives the gold answer above the other in-batch answers, semantically meaningful hard negatives [21]. It is discriminative where maximum likelihood is purely generative, suppressing confusable competitors, not just rewarding the truth; it is *added* to the generative loss under a single weight (zero recovers the backbone exactly), needs no extra data, and generalises across datasets. Across three benchmarks, it improves where calibration is the bottleneck and stays at parity elsewhere, with a single peaked weight ablation confirming a genuine calibration effect.

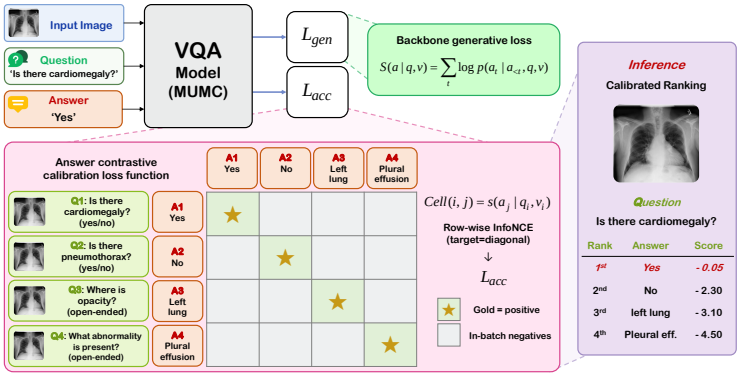

Fig. 1: Overview of Answer Contrastive Calibration (ACC).

2 Method

We first describe the generative VQA backbone and its inference-time ranking procedure, the precise target our objective matches. We then formalize the training inference gap and introduce the ACC loss.

2.1 Generative VQA backbone and inference time ranking

We build on the standard masked pretrained encoder decoder architecture for medical VQA [10,8], the architecture class for which the training-inference mismatch described above is most consequential, since candidate ranking Eq. (3) is its native inference protocol; the architecture is illustrated in Fig. 1. An image v is encoded by a Vision Transformer into a sequence of patch embeddings, and a question q is encoded by a multimodal text encoder (a BERT model [22] whose upper layers cross attend to the image patches), producing a fused question representation $H_{q,v} \in \mathbb{R}^{L_q \times d}$. A separate text decoder g, initialized from the upper layers of the encoder, generates the answer autoregressively while cross attending to $H_{q,v}$. The encoder is initialized from a checkpoint pretrained with masked and contrastive image text objectives; we use this checkpoint exactly as released and never modify it. Given an answer $a = (a_1, \ldots, a_T)$ tokenized with a terminal separator, the decoder defines a conditional likelihood

$$\log p(a \mid q, v) \;=\; \sum_{t=1}^{T} \log p_\theta(a_t \mid a_{<t},\, H_{q,v})\,. \tag{1}$$

The standard fine tuning objective is the per example negative log likelihood of the gold answer $a^\star$, optionally regularized by momentum distillation,

$$\mathcal{L}_{\text{gen}} \;=\; -\frac{1}{B}\sum_{i=1}^{B} \log p(a_i^\star \mid q_i, v_i)\,, \tag{2}$$

where B is the mini batch size. This is exactly the captioning objective: it raises the likelihood of the reference answer and is otherwise indifferent to how probability is distributed over other answers.

Inference is ranking, not generation. At test time the model is not a free form generator. Following [10,7], a fixed candidate set $\mathcal{A} = \{\tilde{a}^{(1)}, \ldots, \tilde{a}^{(N)}\}$ of admissible answers is assembled from the dataset, and each is scored by its conditional generation likelihood under Eq. (1); the prediction is

$$\hat{a} \;=\; \arg\max_{\tilde{a} \in \mathcal{A}} \; \log p(\tilde{a} \mid q, v). \tag{3}$$

(For efficiency the candidates are first pruned by the first token probability and the survivors re scored with the full summed log likelihood, but the quantity that decides the ranking is Eq. (1).) The model is thus trained as a generator (Eq. (2)) but evaluated as a discriminator over $\mathcal{A}$ (Eq. (3)).

2.2 The training inference gap

The mismatch between Eq. (2) and Eq. (3) has a concrete failure mode. Maximum likelihood constrains only the gold answer's score; it places no constraint on the scores of the other candidates in $\mathcal{A}$. A decoder that has become a fluent medical language model can therefore assign high likelihood to several clinically plausible answers at once. Whenever a competing candidate, such as an adjacent anatomical structure, the opposite laterality, or a co occurring finding, receives a summed log likelihood at least as large as the gold answer's, Eq. (3) returns the wrong answer, even though the generative loss in Eq. (2) may already be small. This is most damaging when $\mathcal{A}$ is large and densely populated with near synonyms, that is, for open ended questions, and least damaging for closed binary questions whose two candidates are trivially separable, precisely the error profile we observe in practice (Sec. 3). The remedy does not require a better representation; it requires telling the model, during training, to separate the gold answer from its competitors in the same score space used at inference.

2.3 Answer Contrastive Calibration

Answer Contrastive Calibration adds a discriminative term to the fine tuning objective that operates directly in the ranking score space of Eq. (1). The construction reuses the answers already present in each mini batch as candidates, so it needs no external answer bank, no negative sampling heuristics, and no additional forward passes through the image encoder.

In batch answer scoring. Consider a mini batch of B triples $\{(v_i, q_i, a_i^\star)\}_{i=1}^{B}$. After encoding each question image pair into H_{q_i,v_i}, we score every batch answer $a_j^\star$ against every question i by teacher forcing $a_j^\star$ through the decoder conditioned on H_{q_i,v_i} and summing its token log likelihoods,

$$s_{ij} \;=\; \log p\big(a_j^\star \,\big|\, q_i, v_i\big) \;=\; \sum_t \log p_\theta\big(a_{j,t}^\star \mid a_{j,<t}^\star, H_{q_i,v_i}\big)\,. \tag{4}$$

The diagonal s_{ii} is the gold score for question i; the off diagonal s_{ij} $(j \neq i)$ are the scores of competing real answers drawn from the same dataset, which serve as semantically meaningful hard negatives. Critically, s_{ij} is computed with the *same* summed log likelihood that Eq. (3) uses, so the loss optimises the test criterion itself rather than a surrogate.

Contrastive objective. The per question scores act as logits, and we minimize the temperature scaled cross entropy that ranks the gold answer first in the batch,

$$\mathcal{L}_{\text{acc}} \;=\; -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(s_{ii}/\tau)}{\displaystyle\sum_{j=1}^{B}\mathbb{1}[j\in\mathcal{V}_i]\,\exp(s_{ij}/\tau)}, \tag{5}$$

where τ is a temperature and $\mathcal{V}_i$ is the set of valid columns for row i. Because a mini batch can contain repeated answers (for example two "yes" questions), we mask any off diagonal entry whose answer string is identical to the gold answer, so that a correct answer is never penalized as its own negative; formally $\mathcal{V}_i = \{i\} \cup \{j : a_j^\star \neq a_i^\star\}$. This InfoNCE style loss [23] is discriminative: it raises s_{ii} and suppresses the confusable s_{ij}, directly reducing the ranking errors described above. We deliberately use the unnormalized summed log likelihood, matching the inference ranker, rather than a length normalized score, so that training and testing remain consistent.

Total objective. ACC is added to the generative objective with one weight $\lambda \geq 0$,

$$\mathcal{L} \;=\; \mathcal{L}_{\text{gen}} \;+\; \lambda\,\mathcal{L}_{\text{acc}}. \tag{6}$$

Setting $\lambda = 0$ recovers the unmodified backbone exactly, both numerically and computationally, since the ACC branch is then skipped. Every improvement we report is therefore a controlled effect attributable solely to $\mathcal{L}_{\text{acc}}$, with the data, checkpoint, optimizer, schedule, seed, and evaluation held fixed. The weight λ trades off generative fluency against ranking discrimination, and Sec. 4 shows this trade off has a single, interpretable optimum.

Cost and implementation. ACC reuses the question encodings already computed for $\mathcal{L}_{\text{gen}}$ and adds only the $B \times B$ decoder scoring of Eq. (4); the answers are short, so for the small batch sizes used in medical VQA this is a minor overhead, and the image encoder, the most expensive component, is run once per image. The method requires only a few lines on top of a standard generative VQA loop.

3 Experiments

3.1 Datasets, evaluation and implementation

We evaluate on three standard medical VQA benchmarks. **VQA-RAD** [2] has 3,064 training and 451 test radiology pairs, an even closed/open split, and the

Table 1: Comparison with published methods on three standard medical VQA benchmarks (%). †Our faithful reproduction under identical training conditions.

	VQA-RAD			PathVQA			SLAKE
Method	Open	Closed	Overall	Open	Closed	Overall	Overall
MEVF [6]	43.9	75.1	62.6	8.1	81.4	44.8	78.6
MMQ [5]	52.0	72.4	64.3	11.8	82.1	47.1	–
VQAMix [26]	56.6	79.6	70.4	13.4	83.5	48.6	–
AMAM [27]	**63.8**	80.3	73.3	18.2	84.4	50.4	–
CPRD [28]	61.1	80.4	72.7	–	–	–	82.1
PubMedCLIP [29]	60.1	80.0	72.1	–	–	–	80.1
MUMC†[10]	62.0	82.0	74.06	38.31	90.10	64.22	**86.15**
Ours (ACC)	62.6	**83.5**	**75.17**	**38.32**	**90.51**	**64.44**	86.05

largest open vocabulary for its size. **SLAKE** [3] (English) has 4,919 training, 1,053 validation and 1,061 test pairs. **PathVQA** [4] is the largest, with tens of thousands of pathology pairs and by far the largest answer space, roughly half open ended. Following the generative backbone [10], we rank a fixed candidate set (Eq. (3)); accuracy is exact match after standard normalization, micro averaged over all test questions, with open/closed reported separately.

To isolate the objective's effect, we fix the backbone and training recipe to a faithful reimplementation of [10], changing only the loss. The image encoder is ViT B/16 [24] at 384^2, the text encoder/decoder BERT base [22], initialized from the released pretrained checkpoint (no missing parameters, never modified). We fine tune for 40 epochs with AdamW [25] (lr 2×10^{-5}, weight decay 0.05, cosine schedule), batch size 8, the backbone's augmentation, and momentum distillation. ACC uses one global setting ($\lambda = 0.5$, $\tau = 0.5$) with no dataset specific tuning. All runs share seed and hardware, making the comparison against the $\lambda = 0$ baseline fully controlled.

3.2 Main results

Table 1 compares ACC with published medical VQA methods. For a controlled comparison, we retrain the backbone (MUMC†)[10] under identical code, checkpoint, data and training settings and evaluate ACC under the same conditions; published methods provide external context.

Three observations emerge. On **SLAKE**, both MUMC†[10] and ACC outperform all previously published methods, reaching state-of-the-art performance, and their difference is negligible: the baseline is already well calibrated there. On **PathVQA**, ACC improves modestly over the reproduced backbone while staying substantially ahead of published methods; as the largest and most data rich benchmark, it leaves little room for calibration gains. On **VQA-RAD**, ACC gains most over the reproduced backbone, raising overall accuracy by 1.11 per-

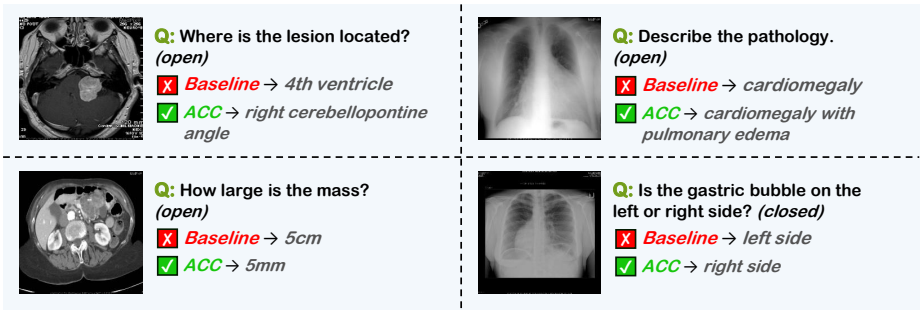

Fig. 2: Qualitative VQA-RAD examples where ACC corrects baseline errors.

centage points, as its small training set and large, confusable answer space make ranking calibration harder.

Overall, the results support our hypothesis: ACC helps most when ranking calibration is the limiting factor and stays neutral when the backbone is already well calibrated, with no extra data, parameters, or architectural changes. Fig. 2 shows qualitative examples of improved answer ranking.

4 Ablation Studies and Analysis

ACC's single hyperparameter is the weight λ balancing the generative and contrastive terms in Eq. (6). We sweep it on VQA-RAD and plot the open, closed, and overall accuracy in Fig. 3. The behaviour follows a single peaked curve: at $\lambda = 0.25$ the contrastive signal is too weak to alter the ranking and the model reproduces the baseline (74.06); at $\lambda = 0.5$ the overall accuracy peaks at 75.17; increasing the weight to $\lambda = 1.0$ reverses the gain (73.84), falling below the baseline. Excessive contrastive pressure sharpens rankings at the expense of generative likelihood, indicating that ACC is most effective when the two objectives are properly balanced. We therefore use $\lambda = 0.5$ across all datasets.

The two largest answer space benchmarks (VQA-RAD and PathVQA) exhibit a large open versus closed gap, whereas SLAKE shows little given its smaller answer vocabulary and larger training set. ACC therefore helps most with many plausible candidate answers. Even on SLAKE, open ended accuracy improves from 82.48 to 82.95 (+0.47). Fig. 4 shows training convergence on VQA-RAD over 40 epochs. ACC consistently reaches a higher accuracy plateau, improving overall, open, and closed accuracy over the baseline throughout training.

Finally, test sets are small (VQA-RAD: 451 questions), so single-run accuracies carry non-trivial variance. We therefore emphasize the controlled, identical-seed comparison against the λ=0 baseline, which exactly recovers the backbone.

To validate the design choice of in batch negatives, we compare three settings: (i) no contrastive loss (baseline), (ii) ACC with in batch negatives (our method), and (iii) ACC with negatives randomly drawn from the full 509 answer

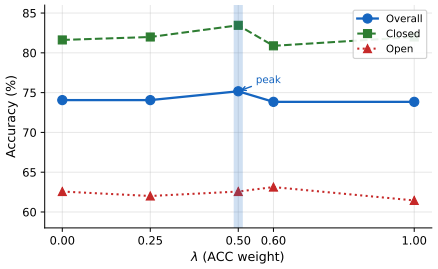

Fig. 3: VQA-RAD accuracy versus λ, split by answer type.

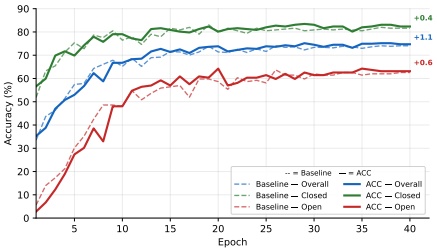

Fig. 4: VQA-RAD training curves (40 epochs): ACC consistently outperforms the baseline on all three metrics.

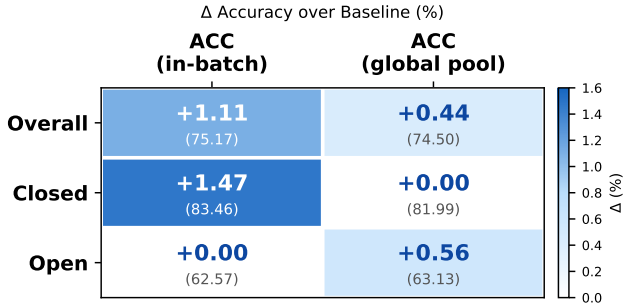

Fig. 5: Negative source on VQA-RAD: gain (Δ%) over the baseline. In batch negatives outperform the global pool on overall and closed accuracy; the pool gains marginally on open ended questions. Separately trained run from Table 1 (same seed and hyperparameters), so absolute values differ slightly.

pool (global pool). Fig. 5 shows the accuracy gain over the baseline. In batch negatives yield the largest overall improvement, driven primarily by closed question accuracy, while the global pool produces a slightly larger gain on open ended questions. Since in batch negatives require no pool pre encoding or external storage while achieving the best overall accuracy, we adopt them as the default.

5 Conclusion

We identified a training-inference gap in candidate-ranking generative medical VQA: maximum-likelihood training deploys models as rankers over a fixed candidate set, an objective it never shapes discriminatively [7]. ACC is a strictly additive, pretraining-free fix that reduces exactly to the baseline when disabled, improving accuracy where confusable answer spaces bottleneck calibration (VQA-RAD, PathVQA) while leaving the saturated SLAKE unharmed.

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. *Artificial Intelligence in Medicine*, 143:102611, 2023.
2. Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.
3. Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.
4. Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.
5. Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *MICCAI*, pages 64–74. Springer, 2021.
6. Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *MICCAI*, pages 522–530. Springer, 2019.
7. Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.
8. Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.
9. 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.
10. Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *MICCAI*, pages 374–383. Springer, 2023.
11. Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
12. 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. *Advances in neural information processing systems*, 36:28541–28564, 2023.
13. Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370, 2024.
14. Changdae Oh, Hyesu Lim, Mijoo Kim, Dongyoon Han, Sangdoo Yun, Jaegul Choo, Alexander Hauptmann, Zhi-Qi Cheng, and Kyungwoo Song. Towards calibrated

robust fine-tuning of vision-language models. *Advances in neural information processing systems*, 37:12677–12707, 2024.
15. Jinlong He, Pengfei Li, Gang Liu, and Shenjun Zhong. Parameter-efficient fine-tuning medical multimodal large language models for medical visual grounding. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025.
16. Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Y Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. In *International Conference on Learning Representations*, volume 2025, pages 66188–66217, 2025.
17. Wenhui Zhu, Xuanzhao Dong, Xin Li, Peijie Qiu, Xiwen Chen, Abolfazl Razi, Aris Sotiras, Yi Su, and Yalin Wang. Toward effective reinforcement learning fine-tuning for medical vqa in vision-language models. *arXiv preprint arXiv:2505.13973*, 2025.
18. Aditya Shourya, Michel Dumontier, and Chang Sun. Adapting lightweight vision language models for radiological visual question answering. *arXiv preprint arXiv:2506.14451*, 2025.
19. Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.
20. Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.
21. Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
22. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
23. Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
24. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
25. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
26. Haifan Gong, Guanqi Chen, Mingzhi Mao, Zhen Li, and Guanbin Li. Vqamix: Conditional triplet mixup for medical visual question answering. *IEEE Transactions on Medical Imaging*, 41(11):3332–3343, 2022.
27. Haiwei Pan, Shuning He, Kejia Zhang, Bo Qu, Chunling Chen, and Kun Shi. Amam: an attention-based multimodal alignment model for medical visual question answering. *Knowledge-Based Systems*, 255:109763, 2022.
28. Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *MICCAI*, pages 210–220. Springer, 2021.
29. Sedigheh Eslami, Gerard De Melo, and Christoph Meinel. 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.