Synthesizing High-Quality Visual Question Answering from Medical Documents with Generator-Verifier LMMs

Xiaoke Huang[1*], **Ningsen Wang**[1,2*], **Hui Liu**[3], **Xianfeng Tang**[3], **Yuyin Zhou**[1]
[1] UC Santa Cruz [2] Fudan University [3] Amazon Research
* indicates equal contribution
https://github.com/UCSC-VLAA/MedVLSynther

Abstract

Large Multimodal Models (LMMs) are increasingly capable of answering medical questions that require joint reasoning over images and text, yet training general medical VQA systems is impeded by the lack of large, openly usable, high-quality corpora. We present **MedVLSynther**, a rubric-guided generator-verifier framework that synthesizes high-quality multiple-choice VQA items directly from open biomedical literature by conditioning on figures, captions, and in-text references. The generator produces self-contained stems and parallel, mutually exclusive options under a machine-checkable JSON schema; a multi-stage verifier enforces essential gates (self-containment, single correct answer, clinical validity, image-text consistency), awards fine-grained positive points, and penalizes common failure modes before acceptance. Applying this pipeline to PubMed Central yields *MedSynVQA*: 13,087 audited questions over 14,803 images spanning 13 imaging modalities and 28 anatomical regions. Training open-weight LMMs with reinforcement learning using verifiable rewards improves accuracy across six medical VQA benchmarks, achieving averages of 55.85 (3B) and 58.15 (7B), with up to 77.57 on VQA-RAD and 67.76 on PathVQA, outperforming strong medical LMMs. Ablations verify that both generation and verification are necessary and that more verified data consistently helps, and a targeted contamination analysis detects no leakage from evaluation suites. By operating entirely on open literature and open-weight models, MedVLSynther offers an auditable, reproducible, and privacy-preserving path to scalable medical VQA training data.

1 Introduction

Large Multimodal Models (LMMs) are rapidly becoming capable assistants for biomedical discovery and clinical education, where questions must be answered by jointly interpreting medical images (e.g., X-ray, CT, microscopy) and the surrounding textual context (e.g., figure captions, narrative descriptions, etc.). Despite fast progress, training *general* medical VQA systems remains difficult because the community lacks large, openly usable, and *high-quality* training corpora.

On the evaluation side, recent benchmark (Hu et al., 2024; Ye et al., 2024) provide broad and challenging test suites, but they are designed for *assessment* rather than training and therefore offer no training splits. On the training side, existing datasets fall into three categories, each with a limitation. 1) **Manually curated** sets (Lau et al., 2018; Liu et al., 2021; He et al., 2020) are carefully annotated but are either small or bound to narrow modalities and topics, limiting coverage and transfer. 2) **Automatically generated** sets (Zhang et al., 2023b; Chen et al., 2024c) scale more easily but are typically produced by text-only LLMs that ignore visual evidence and figure–text relations, yielding noisy stems, ambiguous options, and medically dubious answers that can impede model learning. 3) **Closed, large-scale** resources (Li et al., 2024) exist but are not publicly shareable due to patient privacy, licensing, and institutional agreements, which slows open research and reproducibility. Collectively, these constraints lead to a practical bottleneck: we can *evaluate* medical VQA systems comprehensively, but we cannot *train* them broadly and transparently.
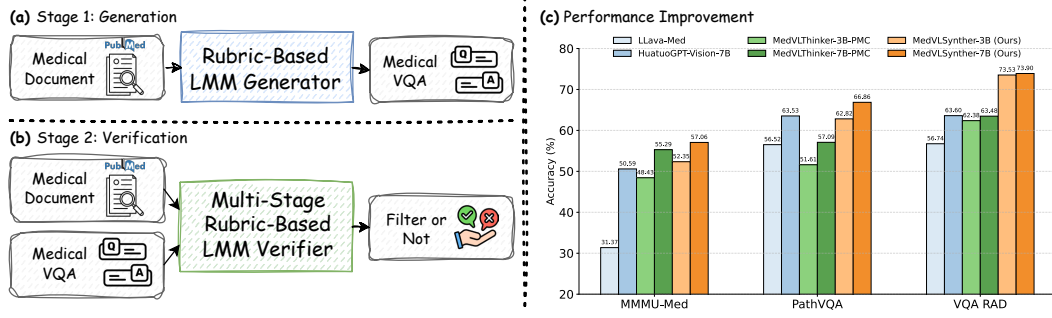
Figure 1: (a) Stage-1 generation: a rubric-guided LMM converts PubMed figures and captions into multiple-choice VQA items. (b) Stage-2 verification: a multi-stage, rubric-based LMM verifier screens items and filters low-quality ones. (c) Training open-weight students (3B/7B) on MedSyn-VQA yields consistent gains over strong medical LMM baselines.

This paper asks a simple question: *can we synthesize high-quality, auditable medical VQA data directly from open biomedical literature?* Our answer is **MedVLSynther**, a generator–verifier framework that leverages state-of-the-art open-weight LMMs (Zeng et al., 2025; Wang et al., 2025; Bai et al., 2025) to produce and automatically vet VQA triplets from figures and surrounding text in PubMed articles (Lozano et al., 2025; Roberts, 2001). The key design choice is to make both generation and verification **explicitly rubric-driven** and **context-aware**.

Rubric-guided context-aware generation (Figure 1 (a)). Given a figure, its caption, and the figure's in-text reference paragraph when available, the generator LMM is instructed to propose a VQA item, including question stem, multiple-choice options, and the correct answer, under a comprehensive rubric. The rubric enforces that stems are self-contained and anchored in the provided visual–textual context, that options are parallel and mutually exclusive, and that the answer can be justified from the figure and caption, not from world knowledge alone. The rubric also specifies a set of accepted *question archetypes* (e.g., recognition, localization, comparative, reasoning) and a JSON schema/format that simplifies downstream filtering and training.

Multi-stage rubric-based verification (Figure 1 (b)). To ensure quality, we feed the same context and the generated VQA to a verifier LMM and score it in three stages: 1) **Essential criteria** form strict pass/fail gates. Any failure discards the item. 2) **Fine-grained criteria** award positive points with justifications, and allow the verifier to surface additional criteria opportunistically. 3) **Penalty criteria** investigate common failure modes and subtract points when detected. We sum the fine-grained and penalty scores and apply a threshold to filter surviving items. This verifier is model-agnostic and can be instantiated with any open-weight LMM; in practice we find that a verifier different from the generator improves robustness.

The generator–verifier loop yields a *data pipeline* whose rules are transparent and auditable end-to-end. Because we build on open literature rather than protected clinical data, the entire pipeline, including prompts, rubric, and metadata, can be inspected and reproduced. At the same time, recent open-weight LMMs rival proprietary systems on many multimodal tasks (Zeng et al., 2025), allowing us to benefit from strong perception and reasoning while staying fully open.

The resulting medical VQA dataset, **MedSynVQA**, covering diverse modalities, subspecialties, and question archetypes. Models trained on this data with Reinforcement Learning with Verifiable Rewards (RLVR) (Guo et al., 2025; Shao et al., 2024) outperform counterparts trained on PMC-VQA (Zhang et al., 2023b), as well as the strong baseline trained on text-only medical corpora (Huang et al., 2025b). As summarized in Figure 1 (c), our training improves accuracy on MMMU-Med (Yue et al., 2024), PathVQA (He et al., 2020), and VQA-RAD (Lau et al., 2018) over strong baselines (Alshibli et al., 2025; Li et al., 2023). Meanwhile, ablations reveal that (i) both *generation* and *verification* are necessary: their synergy yields the best accuracy, and (ii) scale matters: more verified data consistently helps. We analyze topic coverage, modality distribution, and question types, and most importantly, conduct a **contamination analysis** tailored for synthetic medical VQA; we find **no** detectable leakage from the evaluation sets.

Table 1: Comparison among medical VQA datasets. MedSynVQA is open and reproducible, covering 13 modalities and 28 anatomical regions, with 13,087 questions over 14,803 images. "N/A" indicates missing statistics. "# Rate" denotes ratio of images/questions.

Dataset	# Questions	# Images	# Rate	# Modality	# Anatomy	Annotation	Data Availability	General QA	Training Set
MedXpertQA-MM	2,000	2,852	1.43	10	11	Expert	Open access	Yes	No
GMAI-MMBench	25,831	25,831	1.00	38	N/A	Automatic	Open access	Yes	No
OmniMedVQA	127,995	118,010	0.92	12	26	Automatic	Open access	Yes	No
SLAKE	14,028	642	0.05	3	5	Expert	Open access	No	Yes
VQA-RAD	3,515	315	0.09	3	3	Expert	Open access	No	Yes
PathVQA	32,799	4,998	0.15	2	N/A	Automatic	Open access	No	Yes
PMC-VQA	226,946	149,075	0.66	N/A	N/A	Automatic	Open access	Yes	Yes
GMAI-VL-5.5M	≈ 5,500,000	N/A	N/A	13	N/A	Automatic	Not Open	Yes	Yes
MedSynVQA	13,087	14,803	1.13	13	28	Automatic	Open access	Yes	Yes

(a) Question type

(b) Modality

(c) Anatomy

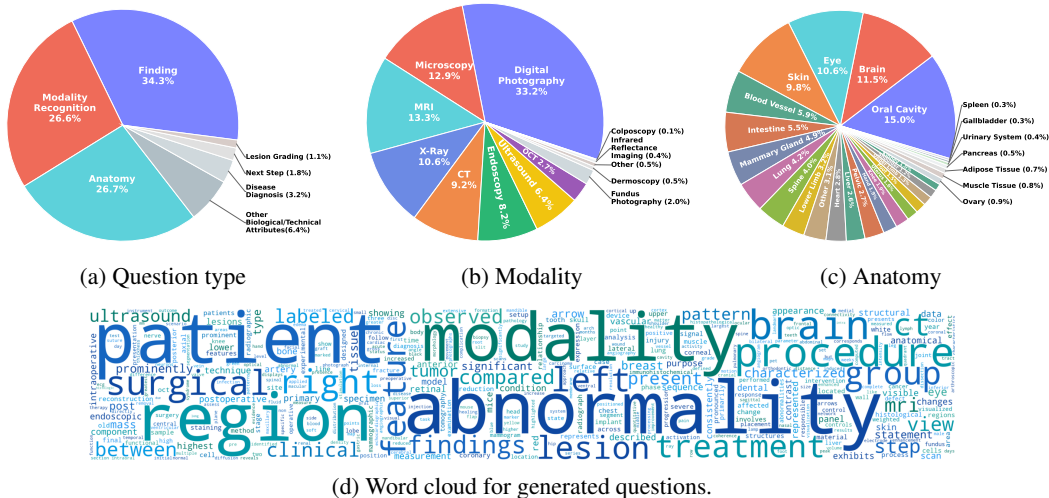(d) Word cloud for generated questions.

Figure 2: **MedSynVQA** statistics: 1) Dataset distributions for question type, imaging modality, and anatomy. 2) Word cloud for generated questions.

Our contributions are summarized as follows:

- **MedVLSynther**, a **rubric-guided, context-aware generator–verifier pipeline** that synthesizes reliable medical VQA from open biomedical articles.

- A **comprehensive rubric** for medical VQA quality, spanning essential gates, fine-grained positive criteria, and penalty criteria, together with a machine-checkable schema that supports automatic filtering and auditing.

- A **synthetic medical VQA training set** (MedSynVQA) that substantially improves medical LMMs on multiple medical VQA benchmarks and complements existing resources without relying on private patient data.

- **Transparency and reproducibility:** our pipeline operates entirely on open data and open models, enabling the community to inspect prompts, scoring rules, and filtering decisions end-to-end.

While synthetic data *cannot* replace carefully curated clinical datasets, our results indicate that *high-quality, auditable synthesis is both feasible and useful* for medical VQA. We hope **MedVLSynther** provides a practical path to scalable training data that respects privacy, encourages openness, and accelerates progress in multimodal medical intelligence.

2 RELATED WORKS

Multimodal medical VQA. Early, expert-curated datasets (Lau et al., 2018; Liu et al., 2021; He et al., 2020) established Med-VQA but remain small or modality-restricted, limiting transfer. Later,
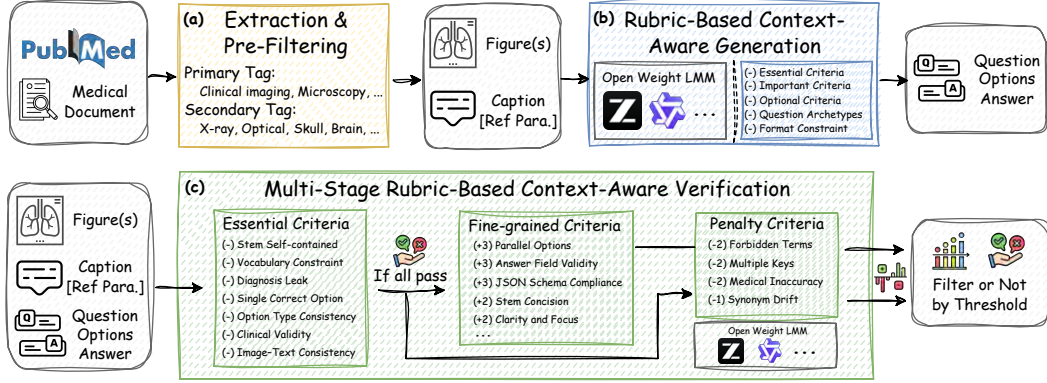
Figure 3: From PubMed documents we extract figures and reference text, then apply (a) extraction and pre-filtering by primary/secondary tags; (b) rubric-based, context-aware generation with format constraints and question archetypes; (c) multi-stage verification with essential, fine-grained, and penalty criteria. Items are retained if their rubric score exceeds a threshold.

broad benchmarks (Hu et al., 2024; Ye et al., 2024; Zuo et al., 2025) consolidated evaluation across many modalities and anatomies yet offer little or no training data, creating a supervision bottleneck. In contrast, large literature-derived corpora (Subramanian et al., 2020; Rückert et al., 2024) and especially Lozano et al. (2025)'s 24M image–caption pairs provide open, scalable raw material. Our work converts this open substrate into exam-quality VQA by coupling context-aware generation with rigorous verification, bridging the gap between expansive evaluation suites and accessible training data.

Synthetic data generation for multimodal medical VQA. Prior synthetic pipelines scale supervision but suffer quality issues: Li et al. (2023)'s self-instruct approach and Zhang et al. (2023b)'s 227k auto-generated pairs (largely from text-only LLMs) can omit modality cues, produce ambiguous stems, and yield visually ungrounded answers; broader compilations like Li et al. (2024) are closed, while modality-specific (Hu et al.) sets remain narrow. These limitations motivate a quality-first strategy: we condition on figures, captions, and in-text references and enforce a rubric-guided generator plus a multi-stage verifier to filter low-quality items, yielding reliable, open data suitable for training medical LMMs without relying on private images.

Multimodal models, medical adaptation, and reasoning. General LVLMs (Hurst et al., 2024; Comanici et al., 2025; Wang et al., 2025; Bai et al., 2025; Liu et al., 2024; An et al., 2025; An et al.) acquire instruction following via visual SFT, while medical variants (Tu et al., 2024; Luo et al., 2023; Alshibli et al., 2025; Liu et al., 2023; Wu et al., 2025; Chen et al., 2024b; Zhou et al., 2024; Wu et al., 2023) add in-domain pretraining and SFT/RL for clinical competence. Recent deliberate-reasoning models (Jaech et al., 2024) show that reinforcement learning with verifiable rewards (e.g., GRPO (Guo et al., 2025)) can surpass SFT-only methods on multi-step problems, and early medical efforts point the same way but lack open, high-quality multimodal supervision. Our rubric-verified VQA corpus supplies that missing signal and pairs naturally work for RLVR, contributing auditable and trustworthy visual reasoning in open-weight medical LMMs (Chen et al., 2024c; Su et al., 2025).

3 MEDVLSYNTHER AND MEDSYNVQA

Our goal is to synthesize high-quality, clinically valid multiple-choice VQA (MC-VQA) examples directly from biomedical papers (Lozano et al., 2025). We cast the task as a **Generator–Verifier** pipeline driven by Large Multimodal Models (LMMs): a *rubric-guided generator* produces MC-VQA items from figures and text, and a *multi-stage rubric-guided verifier* performs automatic quality control before data are admitted to the final corpus (Figure 3).

3.1 Data source, extraction, and pre-filtering

Source. We build on Biomedica (Lozano et al., 2025), a large-scale extraction of figures and figure-level metadata from the PubMed Central Open-Access (PMC-OA) collection (Roberts, 2001). For each paper we ingest: 1) The figure image(s) (a single caption may reference up to 6 images), 2) The figure caption. 3) The corresponding *figure references* in the main text (when present).

Samples missing either images or a caption are discarded.

Pre-filtering. We retain items annotated by Lozano et al. (2025) with the **primary labels**: *Clinical imaging* and *Microscopy*, and 25 **secondary subtypes** (e.g., *x-ray radiography*, *optical coherence tomography*, *skull*, *brain*, etc.). After pre-filtering we obtain **23,788** figure-caption(-reference) triplets.

We denote each pre-filtered sample by

$$x = (\mathcal{I}, \mathcal{C}, \mathcal{R}), \tag{1}$$

where $\mathcal{I}$ is one or more images, $\mathcal{C}$ the caption, and $\mathcal{R}$ the in-text references.

Choice of generator and verifier LMMs. We use state-of-the-art *open-weight LMM* capable of long-context vision-language reasoning: GLM-4.5V-108B (Zeng et al., 2025), InterVL-3.5-38B (Wang et al., 2025), and Qwen2.5-VL-72B (Bai et al., 2025). Unless otherwise noted, GLM-4.5V-108B serves as the default generator due to its strong instruction-following and image-grounding performance. The rubric and strict JSON schema make the output predictable and machine-verifiable.

3.2 Rubric-based, context-aware VQA generation

Given x, the *generator LMM* G_θ produces a *5-option* MC-VQA instance in strict JSON format: $y = \{q, \text{options}\{A..E\}, \text{answer} \in \{A..E\}\}$. Generation is *context-aware* the model receives the image(s) together with $\mathcal{C}$ and $\mathcal{R}$. To ensure exam-quality items, the prompt instills the role of an *expert medical-education item writer* and enforces a *self-check rubric*.

- **Essential (must pass before output):** 1) *Stem self-contained* (no "caption/context" mentions); 2) *Image–content alignment* (requires inspecting specific visual features); 3) *Implicit use of caption facts without answer leakage*; 4) *Exactly one best answer*; 5) *Medical correctness* (modality, anatomy, terminology).
- **Important (strongly recommended):** cognitive level is over application; strong, parallel distractors; clear focus on a single concept.
- **Optional:** localization or quantitative details when clearly supported.

A small set of *question archetypes* (*i.e.*, finding identification, diagnosis, next step, localization, modality recognition) reduces prompt entropy and encourages clinically meaningful questions.

3.3 Multi-stage, rubric-based, context-aware verification

While the generator is reliable, automatic verification is essential for scale and precision. Given x and a candidate MC-VQA y, the *verifier LMM* V_ϕ is prompted to operate in two roles, *Referee* and *Critic*, and to return only a structured rubric with binary scores. Verification is also *context-aware*: V_ϕ sees the same images, caption, and references as G_θ plus the proposed MC-VQA.

Stage-1: Essential screening (hard gate). The Referee evaluates seven non-negotiable items; a sample *must pass all* to proceed: 1) Stem Self-contained; 2) Vocabulary Constraint (no unsupported clinical facts); 3) Diagnosis Leak (no verbatim restatement from sources); 4) Single Correct Option; 5) Option Type Consistency (same semantic type); 6) Clinical Validity (terminology/modality/anatomy); 7) Image–Text Consistency.

Items are scored $\{0, 5\}$ with a fair, rule-based mindset. During this stage, we remove instances that the verifier cannot grade (e.g., malformed JSON), leaving **23,635** candidates. After applying the essential filter, **22,903** remain.

Table 2: Generator–Verifier pipeline ablation. Rubric-guided generation outperforms text-only, and adding verification yields the best average accuracy. Cells are shaded by accuracy; darker is better.

Model	MMMU	MedX-M	PathVQA	PMC	SLAKE	VQA-RAD	Avg.
Qwen2.5-VL-3B-Instruct	44.12	20.69	61.96	44.77	61.30	62.01	49.14
PMC-Style Text-only Generation	48.82	20.40	63.38	51.08	73.08	72.06	54.80
Rubric Context-Aware Generation	52.35	20.60	62.49	51.83	70.43	70.59	54.72
+ Rubric Context-Aware Verification	52.35	21.40	62.82	50.23	74.76	73.53	55.85
Qwen2.5-VL-7B-Instruct	52.94	18.89	65.39	49.30	65.71	68.75	53.50
PMC-Style Text-only Generation	51.76	21.70	64.31	53.43	68.03	71.69	55.15
Rubric Context-Aware Generation	58.24	23.50	65.41	53.83	68.03	75.00	57.33
+ Rubric Context-Aware Verification	57.06	23.15	66.36	53.78	67.79	77.21	57.56

Stage-2: Fine-grained positive criteria (bonus points). The Critic now assumes *the item is not excellent* and awards points *only on irrefutable evidence* We query 4–8 bonus criteria (binary, with weights *Important* = 3 or 4, *Optional* = 1 or 2), including: 1) Plausible Distractors (every distractor is a strong near-miss); 2) Parallel Options (length/structure uniformity); 3) Stem Concision (less than two sentences and concise); 4) Clarity and Focus (single, unambiguous question); 5) Answer-field Validity (answer exists, matches an option); 6) JSON Schema Compliance (exact keys, no extras).

The Critic denies a criterion if it can imagine a slightly better wording or distractor, pushing precision over recall.

Stage-3: Penalty criteria (error hunting). Finally, the Critic actively searches for pitfalls (negative weights): 1) Forbidden Terms (-2; stem contains "caption/context"); 2) Synonym Drift (-1; introduces unsupported specific facts); 3) Multiple Keys (-2), and Medical Inaccuracy (-2).

Each pitfall is triggered only with a concrete reason.

3.4 AGGREGATION AND ACCEPTANCE RULE

Let $\mathcal{P}$ be the set of positive (Important $\cup$ Optional) criteria with weights $w_i > 0$ and binary scores $s_i \in \{0, w_i\}$. Let $\mathcal{N}$ be the pitfalls with $w_j < 0$ and scores $p_j \in \{0, w_j\}$. We compute a *normalized quality score*:

$$S(x,y) = \text{clip}_{[0,1]}\left(\frac{\sum_{i\in\mathcal{P}} s_i + \sum_{j\in\mathcal{N}} p_j}{\sum_{i\in\mathcal{P}} w_i}\right). \quad (2)$$

Candidates passing Stage-1 are *accepted* if $S(x,y) \geq \tau$ with $\tau = 0.9670$. This high threshold emphasizes precision while keeping a useful yield; it results in **13,087** MC-VQA items, which we call **MedSynVQA**.

3.5 TRAINING MEDICAL LMMs WITH MEDSYNVQA

We use our synthesized corpus to train medical LMMs with two LMM finetuning approaches.

Supervised Fine-Tuning (SFT). Following MedVLThinker, we elicit *thinking traces* with GLM-4.5V-108B and perform SFT on (thinking trace, answer) pairs. The supervision emphasizes clinically grounded reasoning paths while preserving the strict answer format.

RL with Verbal Rewards (RLVR). We then apply GRPO on *answers only* (no trace optimization), again mirroring hyper-parameters from Huang et al. (2025b). The reward promotes exact-match accuracy and adherence to the schema without over-fitting to any single imaging modality.

4 EXPERIMENTS

4.1 SETUP

Models. Unless otherwise stated, we finetune two open-weight LMMs Qwen2.5-VL 3B and 7B Instruct (Bai et al., 2025), using the same training schedule, image resolution, tokenization, and

Table 3: Dataset scale ablation. Effect of the number of MedSynVQA training items (1k–13k) on downstream accuracy. Performance improves with scale, with diminishing returns beyond 5k examples. "N/A" denotes zero-shot (no additional training). Cells are shaded by accuracy.

Model	Scale	MMMU	MedX-M	PathVQA	PMC	SLAKE	VQA-Rad	Avg.
	N/A	44.12	20.69	61.96	44.77	61.30	62.01	49.14
	1000	50.59	20.20	63.18	48.37	65.87	67.65	52.64
Qwen2.5-VL	2000	47.06	19.95	64.07	47.27	74.04	76.84	54.87
3B-Instruct	5000	52.35	21.40	62.82	50.23	74.76	73.53	55.85
	10000	48.82	20.55	63.44	49.87	72.84	74.63	55.03
	Full	51.76	22.30	63.03	48.92	72.60	72.43	55.17
	N/A	52.94	18.89	65.39	49.30	65.71	68.75	53.50
	1000	57.65	21.60	65.53	50.93	68.27	65.81	54.96
Qwen2.5-VL	2000	60.00	22.35	67.76	51.18	67.31	73.16	56.96
7B-Instruct	5000	57.06	23.15	66.36	53.78	67.79	77.21	57.56
	10000	57.06	22.45	66.86	52.73	71.88	73.90	57.48
	Full	55.88	22.10	65.56	55.43	72.36	77.57	58.15

Table 4: Choice of generator and verifier LMMs. We vary the generator and verifier. Higher-capacity generator/verifier pairs produce higher-quality data and consistently improve the final average accuracy. "N/A" indicates the zero-shot performance. Cells are shaded by accuracy.

Model	Generator	Verifier	MMMU	MedX-M	PathVQA	PMC	SLAKE	VQA-Rad	Avg.
	N/A	N/A	44.12	20.69	61.96	44.77	61.30	62.01	49.14
Qwen2.5-VL	GLM-4.5V 108B	Qwen2.5-VL 72B	52.35	21.40	62.82	50.23	74.76	73.53	55.85
3B-Instruct	GLM-4.5V 108B	GLM-4.5V 108B	51.18	20.30	63.56	50.63	71.63	70.22	54.59
	Qwen2.5-VL 72B	GLM-4.5V 108B	47.65	21.50	62.37	48.87	73.32	69.85	53.93
	InternVL3.5 38B	GLM-4.5V 108B	49.41	21.90	61.81	51.98	74.76	71.32	55.20
	N/A	N/A	52.94	18.89	65.39	49.30	65.71	68.75	53.50
Qwen2.5-VL	GLM-4.5V 108B	Qwen2.5-VL 72B	57.06	23.15	66.36	53.78	67.79	77.21	57.56
7B-Instruct	GLM-4.5V 108B	GLM-4.5V 108B	58.82	23.65	67.22	54.48	71.15	73.16	58.08
	Qwen2.5-VL 72B	GLM-4.5V 108B	56.47	22.55	67.25	52.38	67.07	72.79	56.42
	InternVL3.5 38B	GLM-4.5V 108B	57.65	23.30	66.12	53.58	70.67	75.37	57.78

optimization hyper-parameters as Huang et al. (2025b). We use our rubric-guided generator–verifier pipeline to synthesize training items from PubMed figures and captions (Figure 3), and we train students either with SFT or RLVR. Unless otherwise noted, experiments use **5K** samples.

Benchmarks and metric. We follow Huang et al. (2025b) evaluation suite and scripts, reporting multiple-choice accuracy on six medical VQA benchmarks: MMMU medical split (MMMU-Med) (Yue et al., 2024), (MedX-M) (Zuo et al., 2025), PathVQA (He et al., 2020), PMC-VQA (Zhang et al., 2023b), SLAKE (Liu et al., 2021), and VQA-RAD (Lau et al., 2018).

Baselines. We compare against strong general-purpose and medical LMMs used in Huang et al. (2025b), including Gemma3 4B (Team et al., 2025), Qwen2.5-VL-3B/7B-Instruct, MedGemma 4B (Sellergren et al., 2025), LLaVA-Med (Li et al., 2023), HuatouGPT-Vision-7B (Chen et al., 2024c), and MedVLThinker (Huang et al., 2025b), strong baselines trained solely on text-only data.

4.2 RESULTS

Ablation on the Generator–Verifier pipeline. Table 2 studies each stage of our pipeline. We begin from zero-shot Qwen2.5-VL students and add 1) PMC-style text-only question generation, 2) rubric-guided context-aware generation, and 3) rubric-aware verification. For 3B student, the base model averages 49.14. Text-only generation lifts the average to 54.80. Switching to rubric-guided, context-aware generation performs similarly on average (54.72). Adding verification yields the best average, 55.85, with large gains on clinically grounded datasets. For 7B student, the base model averages 53.50. Text-only generation yields 55.15, rubric-guided generation 57.33, and with verification we obtain 57.56 and again improving across benchmarks. Overall, rubric guidance already outperforms a PMC-style text-only recipe, and multi-stage verification supplies the remaining headroom, producing the best average in both scales (Table 2). The trend aligns with the high-level improvements visualized in Figure 1.

Table 5: Training approach and data source ablation. Comparing SFT and RLVR using three sources: PMC (image-text), m23k (text-only), MedSynVQA. RL consistently outperforms SFT, and MedSynVQA leads to the highest averages across all benchmarks. Cells are shaded by accuracy.

Model	MMMU	MedX-M	PathVQA	PMC	SLAKE	VQA-RAD	Avg.
Qwen2.5-VL-3B-Instruct	44.12	20.69	61.96	44.77	61.30	62.01	49.14
SFT (PMC)	47.84	21.46	52.76	54.55	65.79	58.58	50.16
SFT (m23k)	32.55	16.00	42.74	28.53	43.91	33.09	32.80
SFT (MedSynVQA)	48.82	20.90	63.12	47.57	54.33	59.93	49.11
RL (PMC)	48.43	21.51	51.61	54.22	75.56	62.38	52.28
RL (m23k)	52.16	22.90	62.28	47.32	63.38	71.08	53.19
RL (MedSynVQA)	52.35	21.40	62.82	50.23	74.76	73.53	55.85
Qwen2.5-VL-7B-Instruct	52.94	18.89	65.39	49.30	65.71	68.75	53.50
SFT (PMC)	49.80	21.39	53.02	54.67	67.71	57.72	50.72
SFT (m23k)	46.86	16.40	56.35	34.58	54.97	53.80	43.83
SFT (MedSynVQA)	49.41	20.90	64.81	50.08	59.62	66.54	51.89
RL (PMC)	55.29	24.11	57.09	55.38	66.59	63.48	53.66
RL (m23k)	56.86	24.43	66.83	50.67	65.79	64.71	54.88
RL (MedSynVQA)	57.06	23.15	66.36	53.78	67.79	77.21	57.56

Table 6: Comparison to baselines. Average and per-benchmark accuracy of general-purpose and medical LMMs versus models trained with MedSynVQA. Both MedVLSynther 3B and 7B achieve the best average across benchmarks, demonstrating strong gains at small and medium scales.

Model	MMMU	MedX-M	PathVQA	PMC	SLAKE	VQA-Rad	Avg.
General LLM							
Gemme 3 4B	46.67	21.89	59.24	44.42	66.59	56.86	49.28
Qwen2.5-VL-3B-Instruct	44.12	20.69	61.96	44.77	61.30	62.01	49.14
Qwen2.5-VL-7B-Instruct	52.94	18.89	65.39	49.30	65.71	68.75	53.50
Medical LLM							
MedGemma 4B	32.55	8.17	59.64	42.73	83.49	78.55	50.86
MedGemma 27B	35.88	12.13	62.09	36.75	77.40	72.67	49.49
Llava Med V1.5 7B	31.37	22.56	56.52	34.28	62.82	56.74	44.05
HuatuoGPT-Vision-7B	50.59	22.00	63.53	53.39	75.00	63.60	54.69
MedVLThinker-3B	52.16	22.90	62.28	47.32	63.38	71.08	53.19
MedVLThinker-7B	56.86	24.43	66.83	50.67	65.79	64.71	54.88
MedVLSynther-3B	52.35	21.40	62.82	50.23	74.76	73.53	55.85
MedVLSynther-7B	55.88	22.10	65.56	55.43	72.36	77.57	58.15

How much synthesized data do we need? We vary the number of MedSynVQA training items from 1K to 13K (Table 3). For 3B student. Accuracy increases from 52.64 (1K) to 55.85 (5K), then plateaus at 55.03 (10K) and 55.17 (Full). For 7B student. The curve is similar: 54.96 (1K), 56.96 (2K), 57.56 (5K) with a slight dip to 57.48 at 10K, and a peak of 58.15 with the full dataset. This tendency suggests the potential for further refinement of the filtering method. Moreover, to reduce computational cost, we use 5K items as the default scale in subsequent experiments.

Which generator and verifier LMMs should we use? We next vary the capacity and identity of the generator and verifier LMMs used during data synthesis (Table 4). For the 3B student, pairing a GLM-4.5V-108B generator with a Qwen2.5-VL-72B verifier yields the best average 55.85; other high-capacity pairs are close. For the 7B student, the same open-weight verifier gives 57.56 with a GLM-108B generator, while using GLM-108B as both generator and verifier further nudges the average to 58.08. We keep the Qwen2.5-VL-72B verifier for the main results to maximize reproducibility with open weights, but Table 4 indicates that stronger verifier capacity translates to higher downstream accuracy.

Training approach and data source ablation. Table 5 compares SFT vs RL from verification reward (RL) across three data sources: PMC-VQA (image–text pairs) (Zhang et al., 2023b), m23k

---

Case 1: Context-Aware Generation

Question: Which of the following best describes the vertebral anomaly shown in this image?
Options: (A) Six lumbar vertebrae with incomplete sacral integration (B) Five lumbar vertebrae with complete sacral fusion ...
Answer: (A)

Caption: Lumbar transitional vertebrae
Context: The subject was diagnosed as having CLBP and a lumbar transitional vertebra. Computer tomography (CT) showed six lumbar vertebrae, which is one more lumbar vertebra than a normal person ... This indicates the first sacrum is not completely integrated...
Pass Verification: **True**

Case 2: Leakage Rejection by Verifier

Question: A 37-year-old woman presents with a palpable mass in the upper outer quadrant of the right breast. Imaging and biopsy reveal a circumscribed hypoechoic mass with flow, invasive ductal carcinoma, grade 1, with predominant tubular formation, mild nuclear atypia, and strong diffuse ER positivity. Which histopathological feature is most consistent with the tumor grade?
Options: ... (C) Well-differentiated with predominant tubular formation and mild nuclear atypia ...
Answer: (C)

Caption: ...invasive ductal carcinoma, grade 1, ...showing predominant tubular formation (arrows), mild nuclear atypia ... ER immunostain showing strong diffuse positivity ...
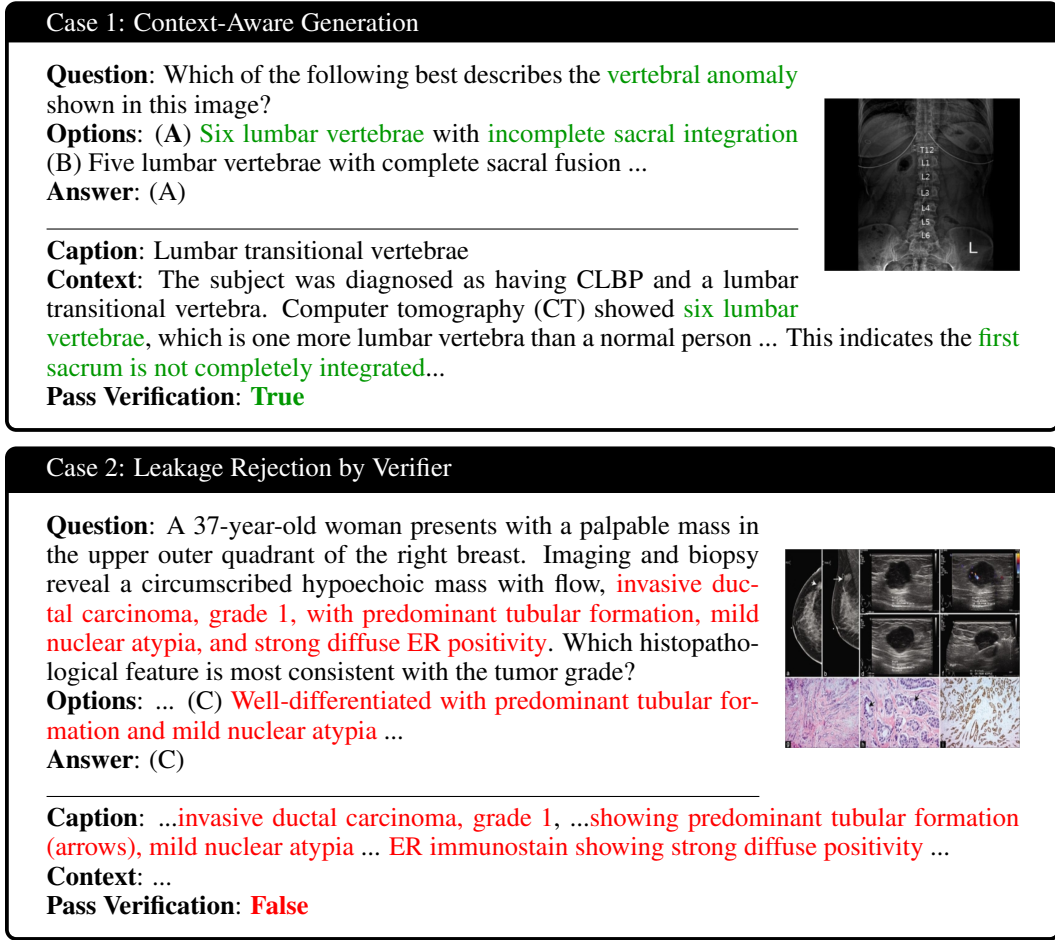**Context**: ...
Pass Verification: **False**

Figure 4: Examples of context-aware generation and leakage rejection by the verifier.

(text-only) (Huang et al., 2025a), and MedSynVQA. 1) RL outperform SFT for both 3B and 7B models, across all data source. 2) Under RL the MedSynVQA signal is the strongest, giving the best average on both 3B (55.85) and 7B (57.56). The results indicate that rubric-based context-aware MedSynVQA dataset are more effective training source than the previous synthetic PMC-VQA (Zhang et al., 2023b) and the text-only one (Huang et al., 2025b).

Comparisons. Table 6 summarizes head-to-head results on the full benchmark suite. Our students trained with MedSynVQA achieve 55.85 (3B) and 58.15 (7B), state-of-the-art averages among open-weight models considered. Notably, our *3B* student surpasses MedVLThinker-*7B* by +0.97 and all other *3–7B* baselines; the *7B* student improves over the best prior MedVLThinker-7B by +3.27. Gains are consistent across datasets, with strong results on VQA-RAD (up to 77.57).

Case study. Figure 4 presents two cases, revealing deep comprehension with context for our generator and the leakage rejection by our verifier. Please refer to the **appendix** for more details.

Contamination analysis. We practice contamination analysis between MedSynVQA and the evaluation suites in the **appendix**. **No** overlaps were found under this protocol.

5 CONCLUSIONS

MedVLSynther shows that high-quality, auditable medical VQA data can be synthesized at scale from open biomedical literature by pairing rubric-guided, context-aware generation with a multi-stage verifier. The resulting MedSynVQA delivers consistent gains for open-weight LMMs across six benchmarks and ablations confirm that both the generator and verifier are necessary. Operating entirely on open data and models, the approach offers a reproducible, privacy-preserving, and transparent path to supervision for medical VQA.

REFERENCES

Ahmad Alshibli, Yakoub Bazi, Mohamad Mahmoud Al Rahhal, and Mansour Zuair. Vision-biollm: Large vision language model for visual dialogue in biomedical imagery. *Biomedical Signal Processing and Control*, 103:107437, 2025.

Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Min Wu, Ming-Ming Cheng, Ender Konukoglu, and Serge Belongie. Multimodality helps few-shot 3d point cloud semantic segmentation. In *The Thirteenth International Conference on Learning Representations*.

Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Junlin Han, Ender Konukoglu, and Serge Belongie. Generalized few-shot 3d point cloud segmentation with vision-language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16997–17007, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024a.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024b.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024c.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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.

Xinyue Hu et al. Medical-cxr-vqa dataset: A large-scale llm-enhanced medical dataset for visual question answering on chest x-ray images.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22170–22183, 2024.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *arXiv preprint arXiv:2504.00869*, 2025a.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. Medvlthinker: Simple baselines for multimodal medical reasoning. *arXiv preprint arXiv:2508.02669*, 2025b.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

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):1–10, 2018.

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.

Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, et al. Gmai-vl & gmai-vl-5.5 m: A large vision-language model and a comprehensive multimodal dataset towards general medical ai. *arXiv preprint arXiv:2411.14522*, 2024.

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)*, pp. 1650–1654. IEEE, 2021.

Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, et al. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19724–19735, 2025.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Richard J Roberts. Pubmed central: The genbank of the published literature, 2001.

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, et al. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin Ye, Pengcheng Chen, Ming Hu, et al. Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning. *arXiv preprint arXiv:2504.01886*, 2025.

Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine Van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*, 2020.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Siddharth Tumre, Sangameshwar Patil, and Alok Kumar. Improved near-duplicate detection for aggregated and paywalled news-feeds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 979–987, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*, 2023.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37: 94327–94427, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023a.

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. *arXiv preprint arXiv:2305.10415*, 2023b.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649, 2024.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.