MEDSKILL: INTERNALIZING CLINICALLY GROUNDED VISUAL SKILLS FOR MEDICAL MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

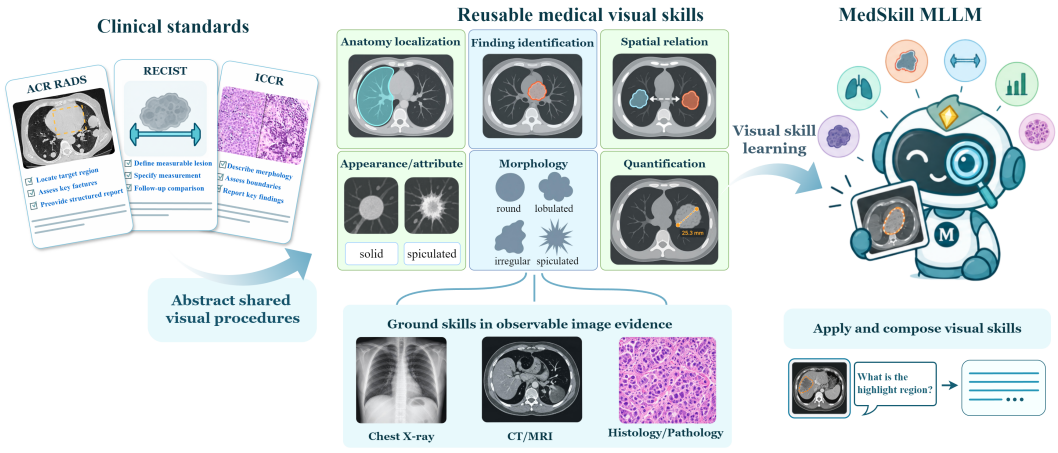

Figure 1: Overview of the MedSkill paradigm. MedSkill abstracts reusable visual procedures from clinical standards, grounds them in observable medical image evidence, and internalizes the resulting visual skills into medical MLLMs for direct visual reasoning.

ABSTRACT

Medical multimodal large language models (MLLMs) have shown strong performance in medical visual question answering, yet their visual competence is still largely acquired from image–text correlations and task-specific supervision. In clinical practice, however, image interpretation relies on reusable visual procedures for locating anatomy, recognizing findings, assessing spatial relations, and characterizing visual appearance and morphology. These procedures are explicitly reflected in standardized reporting systems and assessment guidelines, but remain largely implicit in current medical MLLMs. We introduce MedSkill, a framework for internalizing clinically grounded visual skills into medical MLLMs. MedSkill derives reusable visual skills from authoritative clinical standards and grounds them in diverse medical images. Through structured supervision, these skills are internalized as persistent visual capabilities, enabling the model to apply clinical visual expertise directly during reasoning. Experiments on multiple medical VQA benchmarks show that MedSkill improves overall question-answering performance and strengthens the corresponding visual skills measured through targeted capability evaluation. Further analyses demonstrate that effective transfer depends on how skills are represented, favoring concise, operation-oriented supervision over generic image descriptions or explicit reasoning traces. These results support visual skill internalization as a complementary direction for medical multimodal learning, in which clinically structured visual expertise is transformed from external knowledge into reusable model capabilities.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Multimodal large language models (MLLMs) have recently emerged as a powerful paradigm for integrating visual perception with language reasoning. By extending large language models with visual understanding capabilities, MLLMs have demonstrated strong performance on image description, visual question answering, visual grounding, and multimodal reasoning across a wide range of domains (Alayrac et al., 2022; Liu et al., 2023; Team, 2025). Medicine represents one of the most important application scenarios for these models, where clinical decision-making often requires jointly interpreting visual evidence and medical knowledge. Recent medical MLLMs, including LLaVA-Med, Med-Flamingo, RadFM, and HuatuoGPT-Vision, have adapted general multimodal architectures to biomedical images, radiological examinations, pathological specimens, and medical conversations (Moor et al., 2023; Wu et al., 2025; Chen et al., 2024). These studies demonstrate the potential of MLLMs to assist medical image understanding and clinical reasoning.

Recent research has explored multiple directions to further enhance medical MLLMs beyond conventional instruction tuning. Knowledge augmentation methods incorporate external medical information to improve factual grounding and clinical reasoning (Wang et al., 2025; Li et al., 2026). Reasoning-oriented approaches introduce explicit rationales, self-refinement mechanisms, or reinforcement learning objectives to improve the quality and consistency of model responses (Gai et al., 2025; Pan & Zheng, 2026; Jiang et al., 2026a). Another emerging direction focuses on visual interaction, where models are equipped with additional capabilities such as image cropping, zooming, localization, and segmentation to perform more active visual analysis (Zheng et al., 2025; Jiang et al., 2026b). Together, these approaches have significantly expanded the capabilities of medical MLLMs by improving access to information, reasoning processes, and visual interaction.

Despite this progress, a fundamental question remains: how should medical MLLMs acquire the visual expertise required for clinical interpretation? Medical image understanding is not only determined by recognizing disease concepts or generating clinically plausible answers. Expert interpretation relies on a set of recurring visual operations, including identifying relevant anatomy, detecting abnormalities, establishing spatial relationships, characterizing visual appearance, assessing measurable properties, and comparing changes across observations. However, current medical multimodal training paradigms primarily optimize image-text alignment or task-specific answer generation, leaving these underlying visual procedures largely implicit. As a result, models may learn associations between medical concepts and images without explicitly acquiring the reusable visual capabilities that support clinical interpretation.

Clinical practice provides a natural foundation for defining such capabilities. Medical reporting systems and assessment guidelines developed by clinical organizations encode standardized procedures for observing and describing image evidence. Although these standards were originally designed for human interpretation and communication, they reveal a structured space of recurring visual operations shared across different diseases, organs, and imaging modalities. This motivates a new perspective for medical multimodal learning: instead of treating clinical expertise only as external knowledge to retrieve or apply during inference, can these standardized visual procedures be transformed into capabilities that are internalized by the model itself?

To investigate this question, we introduce **MedSkill**, a framework for **internalizing clinically grounded visual skills for medical multimodal reasoning**. MedSkill first abstracts recurring visual operations from authoritative clinical standards into a unified and benchmark-independent skill space. These skills are subsequently grounded using heterogeneous external medical image collections, where source-grounded annotations are transformed into skill-aligned supervision for model training. Through this process, MedSkill converts clinically defined visual expertise into learnable capabilities rather than additional contextual information or inference-time assistance. The learned model can therefore benefit from clinically grounded visual knowledge while preserving the general multimodal reasoning ability acquired from conventional medical instruction tuning.

We evaluate MedSkill on VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020; 2021), covering diverse radiological and pathological visual reasoning scenarios. Beyond overall question answering performance, we analyze whether clinically grounded skill supervision improves the corresponding visual capabilities, transfers across different medical contexts, and preserves previously acquired model behavior. Our experiments demonstrate that clinical visual expertise can be transformed into a learnable capability space for medical MLLMs.

Our contributions are summarized as follows:

- We formulate **clinically grounded visual skills** as reusable visual operations derived from standardized clinical interpretation procedures, and organize them into five core skill families spanning localization, finding recognition, spatial reasoning, appearance, and morphology.
- We build a **grounded visual skill corpus** of 18,888 instances from PathCap, ROCOv2, and MultiCaRe, and systematically study alternative skill representations ranging from captions and coarse skills to atomic visual states, explicit traces, evidence supervision, and answer-aligned targets.
- We propose a **progressive skill internalization framework** that integrates visual skill supervision with downstream medical VQA training. MedSkill improves both task performance and skill-specific visual capability while retaining direct inference without retrieval, external tools, or explicit skill prompting.

2 Related Work

Recent work on medical multimodal reasoning has improved model performance from several complementary directions. A first line of research strengthens medical vision–language representations and task adaptation. Methods such as AMF improve cross-modal alignment and selectively integrate answer-related information (Zou & Yin, 2025), while lightweight architectures such as MedFG-VQA incorporate memory and graph-aware interaction to enhance medical visual question answering (Gu et al., 2026). A second line introduces additional medical knowledge or external evidence. AMANDA decomposes medical questions and retrieves biomedical knowledge for answer generation (Wang et al., 2025), MR-RAG learns multimodal relevance for retrieval-augmented medical reasoning (Li et al., 2026), and scene-graph-based systems such as SMR-Agents organize structured medical information and coordinate multiple expert agents (Wang et al., 2026). Other approaches focus on improving the reasoning process itself. MedThink supervises explicit medical reasoning traces (Gai et al., 2025), DuCoR refines uncertain predictions through confidence-guided self-refinement (Pan & Zheng, 2026), and Act as You Think optimizes visual and reasoning consistency with reinforcement learning (Jiang et al., 2026a). More recently, medical and general-purpose multimodal agents have incorporated explicit visual interaction. DeepEyes interleaves language reasoning with image operations such as cropping and zooming (Zheng et al., 2025), while IBISAgent trains biomedical MLLMs to perform fine-grained visual interaction and grounding (Jiang et al., 2026b). Collectively, these studies show that medical MLLMs can benefit from better representations, additional knowledge, stronger reasoning supervision, and richer visual interaction. MEDSKILL explores a complementary direction. Instead of providing external cases, procedures, or visual operations primarily at inference time, we use external medical visual experience as supervision for acquiring reusable visual capabilities and subsequently align these capabilities with the target medical reasoning task. The resulting model preserves direct image–question–answer inference while benefiting from medical experience beyond the original task-specific training data.

3 From Augmentation to Visual Capability Internalization

Recent medical MLLMs have achieved strong performance on medical visual question answering, but answer correctness alone offers limited insight into how a model reaches its decision. A correct answer may reflect genuine recognition of the relevant visual evidence, or it may arise from contextual cues and learned answer patterns. *When a model answers correctly, has it actually identified the visual findings that make the answer warranted?* This distinction matters in clinical settings, where reliable reasoning requires the model to locate the relevant structures, assess their visual properties, and connect those observations to its conclusion. We therefore study visual evidence acquisition directly, using zoom-based interaction to probe what the model chooses to inspect and how its answer changes after inspection.

We use the general-purpose Qwen3-VL-8B-Instruct (Team, 2025) as a fixed backbone to probe whether visual access translates into effective evidence acquisition. Direct answering establishes performance without tool assistance; autonomous zooming tests whether the model can decide to inspect and select a relevant region; forced zooming removes the decision to invoke the tool;

and multi-region inspection tests whether broader coverage helps when a single region may miss the evidence. The resulting behavior reveals a gap between answer accuracy and visual evidence acquisition. Zooming rescues some errors, yet the model also selects regions that do not support the question and sometimes changes an initially correct answer to an incorrect one. Thus, a correct answer does not necessarily imply that the model can reliably locate and interpret the evidence on which that answer should rest. This gap motivates MedSkill: learning reusable clinical visual skills that strengthen evidence acquisition within the model, rather than relying on additional visual access at inference time.

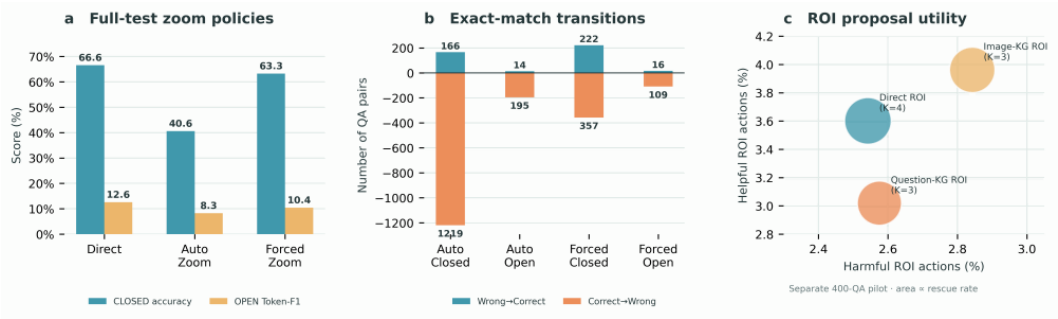

Figure 2: Zoom-based visual evidence diagnostics: full-test performance (a), paired answer transitions (b), and ROI proposal utility in a separate 400-question study (c).

Figure 2 summarizes the results on the combined VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2021) test set. Panel (a) compares direct answering with autonomous and forced zooming across 8,231 questions. Panel (b) tracks changes in the answers to those same questions after tool use. Zooming corrects some initial errors, but changes more correct answers into incorrect ones, especially when the model invokes the tool autonomously. Panel (c) examines candidate crops in a separate 400-question study. Image-KG proposals yield approximately 4.0% helpful ROI actions and 2.8% harmful actions, indicating that potentially useful regions are available even though zooming does not consistently improve the final answer. These results support the distinction raised above. A correct answer alone does not establish that the model can reliably locate and interpret the visual evidence needed to justify it.

We examined images and accompanying descriptions from PathCap (Sun et al., 2024), ROCOv2 (Rückert et al., 2024), and MultiCaRe (Nievas Offidani et al., 2025) to look for visual judgments that recur across medical imaging domains. Despite differences in modality and clinical context, these sources repeatedly involve foundational skills such as identifying structures, locating findings, and assessing visual appearance. Across pathology, radiology, and other clinical images, the descriptions repeatedly involve recognizing structures, locating findings, and assessing their morphology or appearance. We organized such observations into visual-query and visual-state instances, then sampled 6,000 for the analysis in Figure 3, including 2,584 from PathCap, 1,656 from ROCOv2, and 1,760 from MultiCaRe. The figure reveals dense semantic regions associated with different visual operations, as well as regions in which several operations occur together. For the visualization, we mean-pool and L2-normalize MiniLM (Reimers & Gurevych, 2019) representations of the visual-query and visual-state text. Principal component analysis (PCA) reduces the embeddings to 50 dimensions before t-SNE projects them into two dimensions. Dense regions are identified with DBSCAN using $\varepsilon = 2.5$ and $\texttt{min_samples} = 20$, and clusters with at least 75 instances are highlighted.

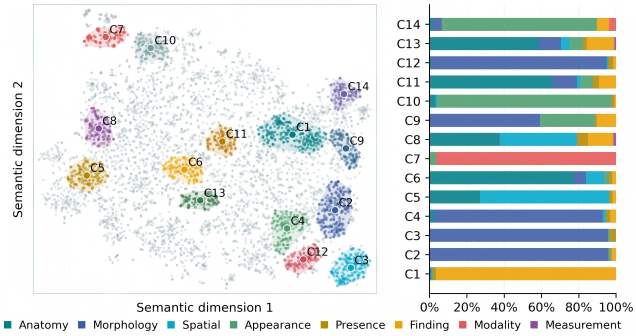

Figure 3: Semantic distribution and visual-operation composition of the sampled medical image instances.

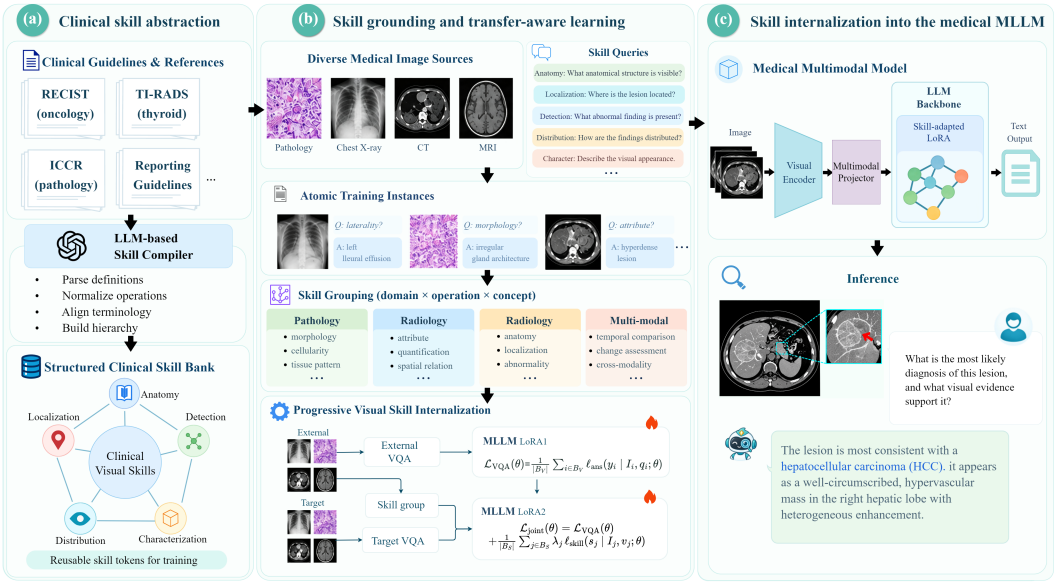

Figure 4: Overview of MedSkill. Clinically grounded visual skills are acquired according to their downstream transfer utility and internalized into medical MLLMs through adaptive skill learning.

4 MedSkill: Internalizing Clinically Grounded Visual Skills

4.1 Clinical Visual Skill Discovery

We derive clinical visual skills from standardized reporting systems and assessment guidelines, including ACR reporting systems, ICCR pathology datasets, and RECIST 1.1 D'Orsi et al. (2018); Tessler et al. (2017); Lung; Ellis et al. (2022); Eisenhauer et al. (2009). We identify fields that require direct visual assessment and decompose each into three elements: an observation target, a visual operation, and an observable state. Operations recurring across organs and imaging modalities are then aligned while preserving the anatomical context and modality-specific terminology required for their execution.

This process reveals a common set of visual procedures, including anatomical localization, finding recognition, spatial reasoning, appearance characterization, morphology assessment, quantitative measurement, and temporal comparison. Based on the visual evidence available in the image sources used in this work, we instantiate five core skill families: *anatomical localization*, *finding recognition*, *spatial relation*, *appearance characterization*, and *morphology description*. Quantitative and temporal procedures remain part of the broader clinical procedure space but are not included in the current empirical skill set.

We focus on operations that can be directly grounded in observable image evidence. Clinical scores, diagnostic categories, and management decisions are therefore not treated as individual visual skills, since they generally combine multiple observations with additional clinical criteria. Each retained skill specifies what should be inspected and the visual state that should be recognized, providing a reusable supervision unit for subsequent skill grounding and model training.

4.2 Constructing the External Visual Skill Corpus

We instantiate five visual skill families from heterogeneous medical image sources: anatomical localization, finding recognition, spatial relation, appearance characterization, and morphology description. Fine-grained operations such as presence detection and finding identification are grouped under finding recognition. Each skill definition specifies an applicable imaging context, an observation target, and an expected visual state. We construct candidate instances from PathCap (Sun et al., 2024),

ROCOv2 (Rückert et al., 2024), and MultiCaRe (Nievas Offidani et al., 2025) by aligning source annotations with the corresponding skill definitions.

For each candidate, Qwen3.8-Max (Qwen Team, 2026) receives the image, its source annotation, and the corresponding skill definition. It produces a focused visual question, a concise answer, and the supporting annotation span.

We retain instances whose visual state is directly supported by the image and source annotation, and remove unsupported, duplicated, or answer-leaking examples. The resulting corpus contains 18,888 instances over 18,887 unique images, including 8,108 from PathCap, 5,448 from ROCOv2, and 5,332 from MultiCaRe.

4.3 Progressive Visual Skill Internalization

We progressively internalize visual skills in two stages. We first adapt Qwen3-VL-8B-Instruct (Qwen Team, 2026) to the medical visual domain using PathCap, ROCOv2, and MultiCaRe. This stage exposes the model to diverse radiological and pathological image–text data and establishes a broad medical visual representation before downstream VQA alignment. Starting from this medically adapted model, we jointly optimize target-domain VQA supervision from VQA-RAD, SLAKE, and PathVQA together with the grounded visual-skill instances constructed in Section 4.2.

For a mini-batch B_V of target VQA examples, each instance consists of an image I_i, a question q_i, and a short answer y_i. We optimize

$$\mathcal{L}_{\mathrm{VQA}}(\theta) = \frac{1}{|B_V|} \sum_{i \in B_V} \ell_{\mathrm{ans}}(y_i \mid I_i, q_i; \theta), \tag{1}$$

where θ denotes the trainable model parameters, $|B_V|$ is the number of VQA examples in the mini-batch, and ℓ_{ans} is the token-level cross-entropy computed over the assistant answer. For a mini-batch B_S of visual-skill instances, each example consists of an image I_j, a focused visual query v_j, and its observable visual state s_j. Joint training uses

$$\mathcal{L}_{\mathrm{joint}}(\theta) = \mathcal{L}_{\mathrm{VQA}}(\theta) + \frac{1}{|B_S|} \sum_{j \in B_S} \lambda_j \, \ell_{\mathrm{skill}}(s_j \mid I_j, v_j; \theta), \tag{2}$$

where $|B_S|$ is the number of skill instances in the mini-batch, ℓ_{skill} is the token-level cross-entropy over the target visual state, and λ_j controls the contribution of the corresponding skill supervision.

We begin with broad coverage of the external skill corpus so that the model acquires reusable visual capabilities across different skill families. The sampling distribution is then refined according to the effect of each skill group on the answer loss of the target VQA data, emphasizing skills that transfer more effectively to downstream reasoning. After this broad skill acquisition stage, we perform VQA-only consolidation and continue training from the selected checkpoint with a newly initialized optimizer. In the final stage, skill examples are further aligned with the target task using answer concepts extracted from the VQA training data. These selected skill instances are expressed in the target short-answer interface and assigned lower weights in Eq. 2. A final VQA-only consolidation step preserves direct answering behavior. At inference time, MedSkill receives only an image and a question, without external retrieval, visual tools, or explicit skill prompts.

5 Experiments

We evaluate MedSkill through four questions concerning answer quality, visual capability, training design, and consistency across medical VQA settings.

RQ1: Does MedSkill improve medical VQA performance?

RQ2: Does MedSkill strengthen the visual skills underlying correct answers?

RQ3: How do the stages of skill internalization contribute to performance?

RQ4: How should visual skills be represented for effective transfer?

Table 1: Performance comparison with previous state-of-the-art methods on VQA-RAD, SLAKE, and PathVQA datasets. For generative methods, recall is reported for open-ended questions and accuracy for closed-ended questions. For discriminative methods, accuracy is used for both question types. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	VQA-RAD		SLAKE		PathVQA		Overall
	Open	Closed	Open	Closed	Open	Closed	
Discriminative methods							
MEVF-BAN (Zhan et al., 2020)	49.20	77.20	77.80	79.80	–	–	71.00
MMQ (Do et al., 2021)	53.70	75.80	–	–	13.40	84.00	56.73
CP+BAN+CR (Liu et al., 2022)	60.50	80.40	80.50	84.10	–	–	76.38
PubMedCLIP (Eslami et al., 2023)	60.10	80.00	78.40	82.50	–	–	75.25
BiomedCLIP (Zhang et al., 2023)	56.13	71.32	85.96	87.14	37.54	88.16	71.04
MUMC (Li et al., 2023b)	61.45	83.46	71.19	87.10	37.86	89.14	71.70
Generative methods							
Medical LM (Van Sonsbeek et al., 2023)	–	–	84.30	82.10	40.00	87.00	73.35
LLaVA-Med (Li et al., 2023a)	31.59	73.32	75.99	79.49	29.87	75.99	61.04
Med-MoE (Jiang et al., 2024)	58.55	82.72	85.06	85.58	34.74	91.98	73.11
AMANDA (Wang et al., 2025)	41.73	63.97	54.41	73.56	–	–	58.42
MedKInstruct (Wu et al., 2026)	57.70	80.88	62.02	81.01	–	–	70.40
DuCoR (Pan & Zheng, 2026)	**68.87**	87.13	**88.87**	87.52	42.95	89.84	77.53
Skill internalization variants							
MedSkill(LLaVA-Med)	64.72	87.35	86.97	87.89	40.15	91.30	76.40
MedSkill(Qwen3-VL)	65.75	**89.85**	88.12	**89.20**	**43.94**	**92.58**	**78.24**

5.1 Experimental Setup

Datasets and metrics. We evaluate on VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2021). We report normalized exact-match accuracy for closed-ended questions and recall for open-ended questions. The external skill corpus described in Section 4.2 is used for training only.

Baselines. We compare with discriminative medical VQA methods, including MEVF-BAN (Zhan et al., 2020), MMQ (Do et al., 2021), CP+BAN+CR (Liu et al., 2022), PubMedCLIP (Eslami et al., 2023), M3AE (Chen et al., 2022), BiomedCLIP (Zhang et al., 2023), and MUMC (Li et al., 2023b). Generative baselines include Medical LM (Van Sonsbeek et al., 2023), LLaVA-Med (Li et al., 2023a), Med-MoE (Jiang et al., 2024), MedKInstruct (Wu et al., 2026), DuCoR (Pan & Zheng, 2026), and AMANDA (Wang et al., 2025). To isolate the effect of skill internalization, we also evaluate three controls with the same backbone: answer-only VQA fine-tuning, skill-only fine-tuning, and uniform skill sampling during joint training.

Implementation details. We adapt Qwen3-VL-8B-Instruct (Team, 2025) with LoRA (Hu et al., 2022), using rank 16, scaling factor 32, and dropout 0.05 while freezing the visual encoder. Training uses AdamW and bfloat16 precision. The initial VQA adaptation uses a learning rate of 5×10^{-6} with cosine decay, a micro-batch size of 1, and 16-step gradient accumulation. Subsequent skill training follows the progressive schedule in Section 4.3. All methods are evaluated by answering directly from the image and question, without skill prompts or visual tools.

5.2 Medical VQA Performance (RQ1)

Table 1 compares MedSkill with previous methods on VQA-RAD, SLAKE, and PathVQA. MedSkill with Qwen3-VL achieves the best overall performance of 78.24, exceeding the strongest competing method, DuCoR, by 0.71 points. The gains are especially clear on closed-ended questions. MedSkill reaches 89.85 on VQA-RAD, 89.20 on SLAKE, and 92.58 on PathVQA, demonstrating stable performance across the three benchmarks. For open-ended questions, MedSkill achieves 65.75 on VQA-RAD, 88.12 on SLAKE, and the best PathVQA result of 43.94, remaining competitive with the strongest existing methods across different answer formats. The benefit of skill internalization

Table 2: Visual-skill execution on 1,927 instances, evaluated by an anonymous multimodal judge. Scores range from 1 to 5. Brackets give 95% bootstrap confidence intervals; the final column reports responses scoring at least 4.

Model stage	Overall	Anatomy	Spatial	Presence	Appearance	Morphology	Score ≥ 4 (%)
Initial VQA	2.743 [2.664, 2.819]	2.406 [2.245, 2.570]	2.263 [2.102, 2.427]	3.893 [3.711, 4.072]	2.856 [2.687, 3.026]	2.264 [2.128, 2.400]	39.60
VQA continuation	2.624 [2.550, 2.705]	2.339 [2.172, 2.503]	2.331 [2.161, 2.495]	3.706 [3.522, 3.890]	2.633 [2.469, 2.795]	2.090 [1.962, 2.223]	36.38
MedSkill	**3.086** [3.006, 3.162]	**2.681** [2.506, 2.853]	**2.699** [2.527, 2.877]	**4.146** [3.972, 4.307]	**3.205** [3.033, 3.374]	**2.672** [2.536, 2.805]	**48.21**

Table 3: Ablation of QA supervision and visual signals. Avg. denotes the unweighted mean of the six reported scores.

Method	QA supervision		Visual signal	VQA-RAD		SLAKE		PathVQA		Avg.
	External	Target		Open	Closed	Open	Closed	Open	Closed	
Effect of QA supervision										
Vanilla	–	–	–	37.88	70.59	55.39	76.92	16.95	64.99	53.79
VQA only	✓	–	–	42.15	78.30	79.56	82.45	25.67	86.23	65.73
VQA only	–	✓	–	44.03	80.62	81.74	84.12	28.09	88.80	67.90
Answer SFT	✓	✓	–	47.84	83.39	83.30	86.97	32.68	90.72	70.82
Interaction between Skill and QA supervision										
Skill only	–	–	Skill	57.49	84.20	83.85	83.67	38.77	89.64	72.94
Skill + External QA	✓	–	Skill	58.76	84.85	84.23	84.21	39.56	90.12	73.62
Skill + Target QA	–	✓	Skill	59.23	85.12	84.67	84.90	40.15	90.89	74.16
MedSkill	✓	✓	Skill	**65.75**	**89.85**	**88.12**	**89.20**	**43.94**	**92.58**	**78.24**

also transfers across model backbones. Applying MedSkill to LLaVA-Med yields an overall score of 76.40, substantially higher than the reported LLaVA-Med baseline of 61.04 and comparable to strong specialized VQA methods. These results show that MedSkill provides consistent gains across radiology and pathology benchmarks while retaining direct image–question–answer inference.

5.3 Visual Skill Acquisition (RQ2)

We assess visual skill acquisition with focused questions covering anatomy localization, spatial relations, presence detection, appearance characterization, and morphology description. The comparison follows three model stages. Initial VQA is the answer-adapted model, VQA continuation further trains it on VQA-RAD, SLAKE, and PathVQA without external skills, and MedSkill subsequently incorporates the external visual-skill supervision. All three models receive the same image and visual question. Their anonymized answers are scored from 1 to 5 by a Qwen3.8-Max (Qwen Team, 2026) judge using the image, reference visual state, and source annotation. As shown in Table 2, VQA-only continuation is accompanied by a decrease in the overall visual-skill score from 2.743 to 2.624. Its scores also fall in anatomy localization, presence detection, appearance characterization, and morphology description. After external skill training, MedSkill reaches 3.086 overall, improving by 0.462 points over VQA continuation and 0.343 points over Initial VQA. The proportion of responses scoring at least 4 rises from 36.38% to 48.21% relative to VQA continuation. The largest gains over that stage occur in morphology description and appearance characterization, at 0.582 and 0.572 points, respectively. These results indicate that the improvement in VQA answers is accompanied by stronger performance on the visual observations probed here.

5.4 Contribution of Visual Skill Supervision (RQ3)

Table 3 first examines how conventional QA supervision affects medical VQA adaptation. Starting from the vanilla model, external QA increases the average score from 53.79 to 65.73, while target-domain QA reaches 67.90. Combining both sources further improves the score to 70.82. These results show that QA supervision provides an effective basis for medical adaptation, with external QA contributing broader medical knowledge and target-domain QA aligning the model more closely with the downstream answer space.

We next study how visual skill learning interacts with the QA-adapted model. Skill-only training reaches an average score of 72.94, indicating that structured visual supervision provides an additional source of learning beyond QA adaptation. Adding external or target QA to skill training further improves the score to 73.62 and 74.16, respectively, while MedSkill reaches 78.24 when both QA sources are combined with visual skill supervision. This pattern suggests that QA supervision and skill learning serve complementary roles. QA training establishes medical and task-specific answering ability, whereas skill supervision encourages the model to acquire reusable visual operations that can transfer across questions and datasets.

5.5 Effect of Visual Skill Representation (RQ4)

Visual skills can be represented through substantially different supervision interfaces. We therefore explored a series of alternatives, including *raw captions* that describe the entire image, *coarse skills* that organize visual information around broad clinical concepts, *atomic visual states* that associate a specific visual operation with a concise observable outcome, *direct skill answers* that express the visual state using the same answer interface as VQA, and explicit *state–answer traces* that require the model to verbalize the intermediate visual state before producing the final answer. We further examined answer-centric weighting and staged skill-to-VQA alignment to determine whether the main difficulty lies in output formatting or in transferring acquired visual states to the downstream answer space.

The controlled comparison in Table 4 shows that representation strongly affects transfer. Raw captions cause a clear performance drop, indicating that global image descriptions provide overly diffuse supervision. Converting the same visual information into structured skills largely recovers this degradation, while adding atomic visual states further improves open-ended transfer. Additional experiments in Appendix B show that visual-state supervision alone does not establish the downstream answer interface, whereas explicitly generating state–answer traces introduces additional decoding burden and degrades direct VQA performance. These findings motivate our final design: MedSkill represents visual knowledge as concise, operation-oriented observable states, uses them as auxiliary supervision during training, preserves direct-answer alignment through VQA supervision, and requires only the final answer at inference.

Table 4: Effect of visual-skill representation.

Representation	C	O	Avg.
Direct VQA	89.03	59.61	74.32
Raw caption	89.52	56.29	72.90
Coarse skill	90.10	59.49	74.80
Coarse + atomic	**90.54**	**65.94**	**78.24**

6 Conclusion

We introduced MedSkill to study whether clinically grounded visual skills can be internalized by medical MLLMs as reusable capabilities. By transforming standardized clinical visual procedures into grounded skill supervision, MedSkill improves both medical VQA performance and the execution of corresponding visual skills, suggesting that clinical visual expertise can be transferred into the model rather than used only as external context. Our experiments also show that skill representation matters: concise operation-oriented visual states provide a more effective interface than generic descriptions or explicit reasoning traces. Our study remains limited in scale. The current evaluation covers only several medical VQA benchmarks, a restricted set of external image sources, and a relatively small clinically grounded skill space. Extending this framework to larger skill corpora, more imaging modalities, broader clinical workflows, and real-world clinical evaluation will be important for determining how far visual skill internalization can generalize beyond benchmark settings.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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*, pp. 7346–7370, 2024.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 679–689. Springer, 2022.

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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 64–74. Springer, 2021.

Carl D'Orsi, L Bassett, and S Feig. Breast imaging reporting and data system (bi-rads). *Breast Imaging*, pp. 29–47, 2018.

Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, Danielle Sargent, Robert Ford, Janet Dancey, Stephen Arbuck, Steve Gwyther, Margaret Mooney, et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

Ian Ellis, KH Allison, Chau Dang, H Gobbi, J Kulka, SR Lakhani, T Moriya, CM Quinn, A Sapino, S Schnitt, et al. Invasive carcinoma of the breast histopathology reporting guide. *International Collaboration on Cancer Reporting: Sydney*, 2022.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the association for computational linguistics: EACL 2023*, pp. 1181–1193, 2023.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. Medthink: A rationale-guided framework for explaining medical visual question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7453–7465, 2025. doi: 10.18653/v1/2025.findings-naacl.415.

Haowen Gu, Gensheng Pei, Zeren Sun, Mingwu Ren, Xiangbo Shu, Yazhou Yao, and Fumin Shen. Medfg-vqa: Low-frequency memory and graph attention for lightweight medical vqa. *arXiv preprint arXiv:2608.26848*, 2026.

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.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers)*, pp. 708–718, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 3843–3860, 2024.

Songtao Jiang, Yuan Wang, Ruizhe Chen, Yan Zhang, Ruilin Luo, Bohan Lei, Yeying Jin, Sibo Song, ZhiBo Yang, Jimeng Sun, Jian Wu, and Zuozhu Liu. Act as you think: Reinforcing consistent reasoning in medical visual question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1788–1805, 2026a.

Yankai Jiang, Qiaoru Li, Binlu Xu, Haoran Sun, Chao Ding, Junting Dong, Yuxiang Cai, Xuhong Zhang, and Jianwei Yin. Ibisagent: Reinforcing pixel-level visual reasoning in mllms for universal biomedical object referring and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20996–21005, 2026b.

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*, 5:180251, 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. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023a.

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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 374–383. Springer, 2023b.

Xuze Li, Haozhao Wang, Zhenyu Huang, Zhongxu Wang, Jinghua Zhang, and Ruixuan Li. Mr-rag: Multimodal relevance-aware retrieval-augmented generation for medical visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15010–15019, 2026.

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, 2021.

Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE transactions on medical imaging*, 42(5):1532–1545, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

RADS Lung. Lung ct screening reporting & data system (lung-rads). *American College of Radiology Committee on Lung-RADS®. Lung-RADS Assessment Categories version*, 1.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023.

Mauro Nievas Offidani, Facundo Roffet, María Carolina González Galtier, Miguel Massiris, and Claudio Delrieux. An open-source clinical case dataset for medical image classification and multimodal ai applications. *Data*, 10(8):123, 2025.

Meihong Pan and Yefeng Zheng. Dual-level confidence based implicit self-refinement for medical visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17215–17225, 2026.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

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.

Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5034–5042, 2024.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Franklin N Tessler, William D Middleton, Edward G Grant, Jenny K Hoang, Lincoln L Berland, Sharlene A Teefey, John J Cronan, Michael D Beland, Terry S Desser, Mary C Frates, et al. Acr thyroid imaging, reporting and data system (ti-rads): white paper of the acr ti-rads committee. *Journal of the American college of radiology*, 14(5):587–595, 2017.

Tom Van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. Open-ended medical visual question answering through prefix tuning of language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 726–736. Springer, 2023.

Dujuan Wang, Tao Cheng, Sutong Wang, Youhua Chen, and Yunqiang Yin. Smr-agents: Synergistic medical reasoning agents for zero-shot medical visual question answering with mllms. *Information Processing & Management*, 63(1):104297, 2026. doi: 10.1016/j.ipm.2025.104297.

Ziqing Wang, Chengsheng Mao, Xiaole Wen, Yuan Luo, and Kaize Ding. Amanda: Agentic medical knowledge augmentation for data-efficient medical visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24810–24832, 2025.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

Yinan Wu, Jihang Jin, Xuhao Bao, Weiyan Zhang, Hanjing Yan, Tong Ruan, and ChunMing Wang. Medkinstruct: A multimodal knowledge graph based framework for multi-hop and hard-negative instruction data synthesis in medvqa. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27935–27947, 2026.

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2345–2354, 2020.

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*, 2023.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Yuanhao Zou and Zhaozheng Yin. Alignment, mining and fusion: Representation alignment with hard negative mining and selective knowledge fusion for medical visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29623–29633, 2025.

A Data Sources and Skill Corpus Construction

A.1 Dataset Overview

We use six medical image datasets with complementary roles. PathCap, ROCOv2, and MultiCaRe provide image–caption pairs for constructing the external visual-skill corpus, while VQA-RAD, SLAKE, and PathVQA provide image–question–answer pairs for downstream training and evaluation.

Table 5: Overview of the datasets used in MedSkill.

Dataset	Role	Local scale	Unique images	Primary annotation
PathCap	External skill construction	223,169	223,169	Pathology captions
ROCOv2	External skill construction	79,793	79,793	Radiology captions / concepts
MultiCaRe	External skill construction	161,702	161,702	Captions / anatomical metadata
VQA-RAD	VQA training / evaluation	3,515 QA	314	Questions / answers / modality
SLAKE	VQA training / evaluation	14,028 QA	642	Bilingual QA / semantic labels
PathVQA	VQA training / evaluation	32,632 QA	4,289	Pathology questions / answers

A.2 Filtering Criteria

We apply filtering at both the source-record and generated-instance levels.

Source-record filtering. A candidate must contain a readable image, a nonempty source annotation, and sufficient lexical evidence for at least one supported visual operation. For MultiCaRe, only radiology and pathology images are retained. Annotations with invalid length or unsupported image domains are discarded before teacher generation.

Generated-instance filtering. A generated instance is retained only when the assigned visual operation is applicable, the visual state is grounded in the source annotation, and the query–state pair is consistent with the operation definition. We additionally enforce length constraints on the visual target, query, and state, remove unsupported quantitative or nonvisual content, and deduplicate repeated image–operation–state combinations.

Table 6: Major reasons for rejecting generated skill instances.

Filtering reason	Initial	Expansion
Operation not correctly executed	1,504	6,507
Insufficient source grounding	1,407	4,959
Supporting span not verbatim	291	2,581
Operation marked inapplicable	899	2,141
Low-confidence output	962	2,302

A.3 Clinical Visual Skill Definitions

The visual procedures identified from clinical standards are consolidated into five core skill families used throughout this work: *anatomical localization*, *finding recognition*, *spatial relation*, *appearance characterization*, and *morphology description*. Each family describes a reusable visual operation rather than a disease category or a downstream VQA question type.

B Additional Analysis of Visual Skill Representation

The final MedSkill representation was selected after exploring multiple alternatives for encoding and transferring visual skills. These designs differ along three dimensions: the semantic granularity of the supervision target, the sequence generated by the student, and the way skill information is connected to downstream VQA training. This section describes the explored formats in detail.

Table 7: The five clinically grounded visual skill families used in MedSkill. Each skill specifies a reusable visual operation and its corresponding observable state.

ID	Skill family	Core visual question	Representative clinical descriptors
S1	Anatomical Localization	Where is the relevant anatomy or visual finding located?	organ, tissue, anatomical subregion, laterality, lobe, quadrant, zone, compartment, anatomical boundary
S2	Finding Recognition	What clinically relevant visual finding is present?	mass, nodule, effusion, hemorrhage, calcification, necrosis, inflammation, abnormal cell population, lesion presence
S3	Spatial Relation	How is the finding spatially related to the surrounding anatomy or other findings?	adjacent to, within, surrounding, crossing a boundary, invading, overlapping, bilateral, focal, diffuse, clustered, distributed
S4	Appearance Characterization	What image-visible attributes characterize the finding?	density, attenuation, signal intensity, echogenicity, enhancement, opacity, color, staining pattern, vascularity, texture
S5	Morphology Description	What structural or morphological properties characterize the finding?	shape, margin, contour, architecture, composition, cellular arrangement, glandular pattern, spindle morphology, irregularity

B.1 Raw Caption Supervision

The simplest form of external supervision directly uses the source image caption as the training target:

$$I \rightarrow c,$$

where I denotes the medical image and c is its source caption. The student is trained to reproduce a global description using the standard answer interface:

> <answer>
> Computed tomography shows a small hematoma with perifocal edema ...
> </answer>

This representation preserves the original medical information but provides little control over which visual operation should be learned. A caption may simultaneously contain anatomy, morphology, diagnosis, and clinical context, and is typically much longer than a downstream VQA answer. We use this setting to test whether additional medical image–text supervision alone accounts for the benefit of MedSkill.

B.2 Coarse Skill Supervision

We next transform external captions into coarse skill-oriented question–answer pairs:

$$(I, q^c) \rightarrow s^c,$$

where q^c is a broad skill question and s^c is a coarse visual state.

> Question: What abnormality is shown in this image?
> Answer: pulmonary lesion with surrounding opacity

Compared with raw caption generation, this format introduces an explicit question–answer interface and a skill-family label. However, a single example may still combine several visual operations, such as localization, characterization, and diagnosis. Coarse skill supervision therefore represents an intermediate stage between global caption learning and operation-specific skill supervision.

B.3 Atomic Visual-State Supervision

We then decompose each external observation into an atomic visual operation:

$$\mathcal{S} = (I, o, t, q^s, s), \tag{3}$$

where o denotes the visual operation, t the observation target, q^s the focused visual query, and s the observable visual state.

For example,

> Operation: Morphology description
> Target: polygonal cells
> Question: What is the morphology of some cells?
> State: Some cells were polygonal.

Atomic supervision isolates a single visual judgment in each training instance. It also enables grouping and sampling by operation, visual concept, and data source.

State-specific serialization. The visual state can be generated using a dedicated output tag,

> <visual_state>
> Some cells were polygonal.
> </visual_state>

which explicitly separates skill execution from conventional VQA answering.

Answer-aligned serialization. Alternatively, the same visual state can be generated using the VQA answer interface,

> <answer>
> Some cells were polygonal.
> </answer>

which reduces the difference between skill supervision and downstream VQA.

B.4 Full Skill Trace and Answer

To make the skill execution process explicit, we also explored structured state-to-answer traces. The target contains the visual operation, observation target, assessability, evidence state, execution steps, and an explicit link between the visual evidence and the final answer:

> <skill_trace>
> operation=detection_presence;
> target=acute brain infarct;
> evidence=Bright wedge-like diffusion hyperintense foci ...
> steps=...
> answer_link=...
> </skill_trace>
> <answer>Yes</answer>

This representation directly supervises the complete path from visual operation to observable evidence and downstream answer. However, the trace is substantially longer than the final VQA response and contains teacher-generated language that is not required during direct inference.

B.5 Compact State-and-Answer Supervision

We therefore reduce the full trace to a single evidence state followed by the short answer:

> <skill_trace>
> Bright diffusion-restricted foci are visible in the right frontal lobe.
> </skill_trace>
> <answer>Yes</answer>

The compact representation removes operation names, explicit execution steps, assessability labels, and answer-link explanations while retaining an intermediate visual state. This format reduces sequence length, but still requires the model to generate additional text before the direct answer.

B.6 Teacher Evidence and Answer

Another representation replaces structured traces with a natural-language visual evidence description:

```
<evidence>
The frontal chest radiograph shows bilateral nodular opacities ...
</evidence>
<answer>Lung</answer>
```

Unlike the structured skill trace, this target does not contain explicit execution steps. Instead, it asks the student to reproduce a teacher-generated image rationale before answering. This experiment tests whether richer visual evidence generation itself improves downstream VQA.

B.7 Oracle Visual Evidence as a Diagnostic Upper Bound

We additionally evaluate whether correct visual evidence can improve answering when it is provided directly to the model. Two levels of evidence are used.

Cue-only evidence. A concise visual cue describes the discriminative observation without directly revealing the answer.

> Focal wedge-shaped diffusion restriction in the right frontal cortex.

Full visual state. A more explicit state may contain the complete clinically relevant observation.

> An acute infarct is visible in the right frontal lobe.

In both settings the evidence is provided as additional input, while the model still produces only the final answer. These experiments are diagnostic rather than deployable MedSkill configurations and estimate how much downstream performance could improve when the required visual evidence is already available.

B.8 Privileged Evidence Distillation

We also study a training interface in which visual evidence is available only to a teacher model:

$$\text{Teacher:} \quad (I, q, e) \rightarrow p_T(a),$$

$$\text{Student:} \quad (I, q) \rightarrow p_S(a),$$

where e denotes privileged atomic visual evidence. The student is optimized to match the teacher answer distribution while never receiving the evidence at inference time.

Unlike explicit visual-state or trace supervision, this design does not require the student to generate a skill representation. It therefore tests whether skill information can be transferred implicitly through answer-distribution distillation.

B.9 Grounded Direct Skill QA

The final MedSkill design moves the visual skill from the generated output to the construction of the training example itself. A grounded skill instance contains a visual operation, a focused skill-oriented question, a supported short answer, and provenance information used for filtering and auditing. However, the student receives only the image and question:

$$(I, q^s) \rightarrow a^s.$$

For example,

Question: What is the hyperdense round lesion identified in this CT scan?
<answer>small hematoma</answer>

The visual operation, source annotation, supporting span, and auxiliary visual evidence are not exposed to the student. They are used for data construction, quality control, grouping, and sampling.

This design aligns both the output protocol and answer granularity of skill supervision with downstream VQA. The skill remains encoded in what the question asks the model to observe and in the evidence required for an instance to be accepted, rather than in an additional sequence that must be generated at inference time.

Table 8: Summary of the visual-skill representations explored in this work.

Design	Skill information	Student target
Raw Caption	Global image description	Caption
Coarse Skill	Broad skill concept	Coarse state
Atomic Visual State	Single visual operation	Visual state
Full Skill Trace	Operation + evidence + steps	Trace + answer
Compact State + Answer	Evidence state	State + answer
Teacher Evidence	Free-form visual rationale	Evidence + answer
Privileged Distillation	Teacher-only evidence	Answer distribution
Grounded Direct Skill QA	Grounded operation encoded in question	Short answer

B.10 Visual-Skill Evaluation and Judge Prompt

We construct 2,000 candidate questions from the external skill corpus, with 400 examples from each of five operations: anatomy localization, spatial relation, detection/presence, appearance characterization, and morphology description. The candidates comprise 709 examples from ROCOv2, 673 from MultiCaRe, and 618 from PathCap. We retain one question per image, verify that each image can be decoded, and sample without inspecting the evaluated models' predictions. Existing visual questions and reference states are used without modification.

All model stages receive the same image and visual question under greedy decoding with a maximum of 64 new tokens. The operation label, reference visual state, and source annotation are withheld from the evaluated models. For judging, model responses are randomly assigned anonymous identifiers. Qwen3.8-Max receives the image, question, operation label, reference state, source annotation, and anonymous responses. It first assesses reference validity and image assessability; only instances marked `yes` on both checks enter the primary analysis. Of the 2,000 candidates, 1,927 satisfy these criteria. Mean scores and confidence intervals are computed over these instances, with paired bootstrap intervals used for differences between model stages.

The judge receives the following system prompt. The original judging run included a fourth anonymous response; the analyses in Section 5.3 use the three model stages reported in Table 2.

Judge system prompt

You are grading execution of an atomic medical visual skill. Inspect the image and use the source annotation and reference visual state as factual anchors. Grade semantic correctness, not verbosity or exact wording. Do not favor longer answers. The four responses are anonymous.

Score every response independently:

5 Fully correct visual state, clinically consistent, and directly answers the query.
4 Essentially correct, with only a minor omission or harmless imprecision.
3 Partly correct and relevant, but incomplete, ambiguous, or containing a material imprecision.
2 Mostly incorrect or unsupported, although it contains a small relevant element.
1 Incorrect, contradictory, hallucinated, empty, or does not answer the visual query.

Also determine whether the supplied reference itself is supported and whether the image is assessable.

Return JSON only with `reference_valid`, `image_assessable`, a score for each anonymous response, `best_responses`, and a brief `reason`.

> Do not reveal reasoning or infer model identity.

The user message supplies the image and a structured payload containing `visual_operation`, `visual_query`, `reference_visual_state`, `source_annotation`, and the anonymous responses `R0`–`R3`. The source annotation is truncated to at most 1,600 characters.