ClinFusion: A Vision-Centric Multimodal LLM System for Holistic Medical Understanding

Hangjie Yuan[1,3,4,⋆], **Yichen Qian**[1,3,⋆], **Zhiwei Tang**[1,3,⋆], **Xianzhe Xu**[2,3,⋆], **Lirong Wu**[1,3], **Sicheng Yang**[1], **Jinwang Wang**[1,3], **Pengju Wang**[1,3], **Zhitao Zeng**[2], **Yizeng Han**[2,3], **Yan Xing**[1,3], **Shengxuan Luo**[2,3], **Tao Feng**[5], **Qing Xie**[6], **Weigen Yao**[6], **Yi Yang**[4], **Zuozhu Liu**[4,7], **Jiasheng Tang**[1,3], **Shaocheng Wang**[8], **Jitao Wang**[8,†], **Jiahong Dong**[8,†], **Weihua Chen**[2,3,†], **Feng Xu**[9,10,†], **Fan Wang**[1,†]

[1]DAMO Academy, Alibaba Group, Hangzhou, China [2]DAMO Academy, Alibaba Group, Beijing, China [3]Hupan Laboratory, Hangzhou, China [4]College of Computer Science and Technology, Zhejiang University, Hangzhou, China [5]Department of Computer Science and Technology, Tsinghua University, Beijing, China [6]Department of Radiology, The Affiliated Yangming Hospital of Ningbo University, Yuyao, China [7]Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University, Haining, China [8]Hepato-Pancreato-Biliary Center, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China [9]School of Software, Tsinghua University, Beijing, China [10]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China

[⋆]Equal contribution, [†]Corresponding authors

Multimodal large language models (MLLMs) hold immense potential to revolutionize clinical practice, yet deploying them in the medical domain is fundamentally a vision-centric challenge: models must absorb knowledge from heterogeneous 2D and 3D medical images, and evaluation protocols must align with radiologists' clinical practice and provide an accurate, fine-grained and factualness-driven assessment. In this paper, we introduce ClinFusion, a vision-centric MLLM designed for holistic medical understanding that systematically addresses these limitations. We propose a compositional and cascaded vision encoder architecture featuring a Cascade Spatial-Aware Locality Fusion operator that unifies diverse 2D and native 3D medical image understanding within a fused encoder. We further introduce a vision-grounded evaluation framework, including MedIF-Bench for instruction-following assessment and a region-of-interest-grounded method for clinically aligned and factualness-driven report generation evaluation. We show that ClinFusion sets a new state-of-the-art across a comprehensive suite of 2D and 3D multimodal medical benchmarks—spanning visual question answering, report generation, and instruction following—as well as textual medical tasks, outperforming leading open-source medical MLLMs (*e.g.*, Hulu-Med, Lingshu) on 20 out of 24 benchmarks and demonstrating multimodal capabilities better than powerful proprietary models such as GPT-5.2 and Gemini-3-Flash on 13 out of 16 benchmarks, and can be further augmented with agentic tool use for retrieval-augmented and tool-assisted clinical workflows. A blinded evaluation by board-certified radiologists confirms that ClinFusion produces the highest-ranked reports, and validates our RoI-grounded metric as achieving the strongest correlation with expert judgment among all automatic evaluation metrics examined.

Github: https://github.com/alibaba-damo-academy/ClinFusion
Huggingface: https://huggingface.co/collections/Alibaba-DAMO-Academy/clinfusion

DAMO
TECH TO THE FUTURE

1 Introduction

The recent proliferation of Multimodal Large Language Models (MLLMs) has unlocked unprecedented capabilities in integrating and reasoning over diverse data modalities [1, 2, 3, 4, 5]. Consequently, the medical domain has emerged as a fertile ground for MLLM-powered intelligent systems, which hold immense potential to revolutionize tasks such as automated diagnostics, clinical report generation, and interactive decision

support.

Because clinical decisions are predominantly driven by visual evidence—from radiographs and pathology slides to volumetric CT and MRI scans—successfully deploying MLLMs in the high-stakes medical field is fundamentally a **vision-centric** challenge. This challenge manifests along two tightly coupled dimensions: **1)** *Visual knowledge absorption*: the architecture must natively process and internalize the heterogeneous visual knowledge unique to medicine, spanning diverse 2D and 3D imaging modalities; and **2)** *Vision-grounded evaluation*: the evaluation system must go beyond surface-level text matching and instead assess a model's visual understanding through clinically meaningful, region-level analysis that reflects real-world diagnostic practice. Underpinning both dimensions is a principled data-curation pipeline that provides the diverse, high-quality visual training signal necessary for robust medical perception. Indeed, recent advancements in Medical MLLMs have made strides in addressing these aspects [6, 7, 8, 9, 10, 11, 12, 13].

Despite promising results on established benchmarks for multimodal [14, 15, 16] and text-based [17, 18, 19] medical tasks, critical gaps remain in both vision-centric dimensions identified above—visual knowledge absorption and vision-grounded evaluation—that hinder the transition of MLLMs to in-the-wild clinical practice. The first critical challenge lies in **Heterogeneous Visual Knowledge Absorption**: A single, monolithic vision encoder is often inadequate for capturing the diverse knowledge embedded in heterogeneous medical data. Medical imaging is fundamentally multimodal, spanning a wide array of 2D (*e.g.*, X-rays, pathology slides, fundus images) and 3D (*e.g.*, CT and MRI scans) modalities. This diversity presents a far greater challenge for visual feature extraction and understanding than that of general-domain applications. The prevailing approach in medical MLLMs has been predominantly *data-centric*: models such as Lingshu [6], Hulu-Med [7], LLaVA-Med [9], and HuatuoGPT-Vision [12] adapt general-purpose vision-language backbones through stage-wise fine-tuning on meticulously curated image-text datasets, yet their underlying vision architectures remain largely unchanged—typically a single, monolithic encoder. Existing architectural strategies can be broadly categorized into two groups: a 2D-centric approach that treats all visual data as 2D images by sampling dense slices from 3D volumes [20, 6, 7, 8], which inevitably sacrifices critical 3D structural information; a 3D-centric approach that employs representative encoders for 3D data, requiring extensive and complex alignment procedures [21, 22, 23]. While multi-encoder designs have been explored in the general vision-language domain [24, 25, 26, 27, 28, 29], they target multi-resolution inputs or stronger visual acuity rather than the unique challenges of medical imaging. The most related work, Cambrian-1 [26], uses a parallel design to aggregate visual information from multiple vision encoders, but lacks a native 3D encoder and the guided fusion mechanism we propose.

The second critical challenge lies in **Vision-Grounded Evaluation**. Current evaluation protocols for medical MLLMs have two critical gaps. The first is that existing benchmarks overlook instruction-following capability, which is often degraded by domain-specific fine-tuning yet remains a prerequisite for any meaningful downstream evaluation and clinical deployment. The second centers on report generation. Report generation is among the most clinically important applications of medical MLLMs, yet its evaluation remains fundamentally flawed. A common practice to evaluate report generation capability is by 1) providing a general instruction for MLLMs to generate a comprehensive report and then 2) comparing the generated report against a reference report. However, this paradigm is misaligned with clinical practice: radiologists interpret images guided by the patient's clinical context (*e.g.*, referral indication, medical history), which directs their attention to specific anatomical regions and shapes the structure of their reports—yet current protocols force models to produce comprehensive reports without any focus on the patient. Moreover, existing metrics compare generated reports against references through either surface-level lexical or semantic matching (*e.g.*, BLEU, ROUGE [30], METEOR [31], CIDEr [32], BERTScore [33]), entity-level extraction and matching that relies on rules or small-scale models with limited accuracy (*e.g.*, RadGraph-F1 [34], RaTEScore [35]), or holistic LLM-based scoring that conflates accuracy and completeness (*e.g.*, GREEN [36])—none of which provide an accurate, fine-grained and factualness-driven decomposition of matched, missed, and hallucinated clinical findings. Beyond these two core vision-centric challenges, a further practical consideration is that MLLMs are inherently passive and isolated systems: their parametric knowledge is static, their reasoning lacks traceability, and they face perception bottlenecks on specialized tasks where dedicated AI models remain superior [37, 38, 39, 40].

To address these fundamental challenges, we present ClinFusion, a vision-centric multimodal large language model system designed for holistic medical understanding. As illustrated in Fig. 1, our work introduces a principled framework that systematically tackles each of the aforementioned limitations through a contribution

spanning architecture, evaluation, and system-level integration.

First, to overcome the challenge of heterogeneous visual knowledge absorption, we propose a Compositional Vision Encoder. Instead of relying on a monolithic encoder, ClinFusion integrates a foundational and well-aligned vision transformer with an ensemble of specialist 2D encoders. More critically, we introduce Cascade Spatial-Aware Locality (CaSL[1]) Fusion, a hierarchical operator that enables a cascaded interplay among the 2D encoders, which progressively enriches a foundational visual representation while preserving its alignment. To natively handle volumetric data, we further incorporate a dedicated 3D encoder and design a 2D-anchored depth-aware CaSL Fusion, where well-aligned 2D feature representations guide the alignment of 3D features. In contrast to the parallel aggregation designs of prior work [26], ClinFusion uniquely incorporates a native 3D vision encoder alongside a compositional ensemble of 2D encoders, coordinated by the CaSL Fusion operator that enables both a cascaded interplay among the 2D encoders for progressive feature enrichment and a unique 2D-anchored interplay between the 2D and 3D modalities.

Second, to bridge the gap between evaluation and clinical practice, we introduce a vision-grounded evaluation framework centered on two key innovations. We propose MedIF-Bench, a comprehensive benchmark that addresses the critical yet overlooked issue of instruction-following in medical contexts, assessing whether MLLMs can accurately execute complex, practical medical instructions—a prerequisite for reliable clinical interaction. More fundamentally, for vision-intensive tasks like report generation, we introduce a Region-of-Interest (RoI)-grounded evaluation methodology that aligns generation with clinical practice by conditioning on patient-specific clinical context (*i.e.*, clinically salient anatomical areas) and employs an LLM-as-a-judge to decompose diagnostic claims into matched, missed, and hallucinated findings, which provides an accurate and factualness-driven assessment.

To validate both our model and our evaluation methodology against expert clinical judgment, we conduct a blinded expert evaluation in which six board-certified radiologists independently rank reports generated by ClinFusion, Gemini-3-Flash, and Hulu-Med across 300 clinical cases spanning CT and X-ray modalities. This study serves a dual purpose: it confirms that ClinFusion produces the highest-ranked reports across factual accuracy, completeness, and clinical utility, and it demonstrates that our RoI-grounded metric achieves the strongest correlation with expert judgment among all eleven automatic metrics examined—providing direct clinical evidence that our evaluation framework measures what clinicians value.

Furthermore, to enhance the practical deployment of ClinFusion in real-world clinical settings, we develop an agentic tool use system that extends the model's capabilities beyond its parametric knowledge. This system equips ClinFusion with retrieval-augmented generation (RAG) for grounding answers in current, citable medical literature, and a suite of perception expert tools that enable it to invoke specialized AI models for high-precision analysis like organ segmentation or disease classification. While the core technical contributions of this work lie in the vision architecture and evaluation framework described above, this agentic extension demonstrates how ClinFusion can be practically deployed as an active clinical assistant with dynamic knowledge access and verifiable, tool-augmented reasoning.

In summary, our main contributions are:

- **A compositional and cascaded vision architecture** that natively unifies heterogeneous 2D and 3D medical imaging through CaSL Fusion, a cascaded and 2D-anchored interplay mechanism, and progressively enriches a foundational and well-aligned visual representation.
- **A new vision-grounded evaluation framework**, featuring the MedIF-Bench for assessing instruction-following in medical scenarios and an RoI-grounded methodology for more clinically aligned and precise report generation evaluation.
- **Leading performance validated by both automatic benchmarks and expert radiologists**. ClinFusion achieves state-of-the-art results across 2D and 3D multimodal medical benchmarks alongside textual medical tasks, outperforming both general-purpose and medical MLLMs, with multimodal performance competitive with leading proprietary models such as GPT-5.2 and Gemini-3-Flash. A blinded evaluation by six board-certified radiologists on 300 clinical cases further confirms that ClinFusion produces the highest-ranked reports and that our RoI-grounded metric best reflects expert clinical judgment. We additionally

[1]CaSL is pronounced as "castle" (UK: /ˈkɑːsəl/, US: /ˈkæsəl/).

demonstrate that equipping ClinFusion with an agentic tool use system—integrating dynamic knowledge retrieval and external specialist models—yields consistent improvements in both text-only and multimodal clinical scenarios.

2 Results

This section presents the results of our study, beginning with an overview of the proposed framework, **ClinFusion**, and the evaluation settings, followed by a detailed analysis of the main findings.

2.1 Overview

2.1.1 System Overview

ClinFusion is a vision-centric multimodal large language model designed for holistic medical understanding. As illustrated in Fig. 2, the system is built upon two core contributions: 1) a **compositional and cascaded vision encoder** featuring the CaSL Fusion operator, which unifies the processing of diverse 2D and native 3D medical data by progressively enriching a foundational Qwen ViT representation with features from specialist encoders; and 2) a **vision-grounded evaluation framework**, including the MedIF-Bench for assessing instruction-following, and an RoI-grounded methodology for report generation evaluation. To further enhance its practical deployment, ClinFusion is complemented by an **agentic tool use extension** that equips it with retrieval-augmented generation and the ability to invoke external specialized models as tools, demonstrating consistent improvements in both text-only and multimodal clinical scenarios. We present results for several variants of our proposed architecture. Our primary models, ClinFusion-8B and ClinFusion-32B, are built upon the Qwen3-VL-8B and Qwen3-VL-32B foundations, respectively.

2.1.2 Evaluation Settings

Models for comparison. To comprehensively evaluate the performance of ClinFusion, we conduct extensive comparisons against three categories of state-of-the-art models:

- **Leading generalist MLLMs**: We include leading open-source MLLMs to establish a general-domain baseline. This includes Qwen2.5-VL series [1], Qwen3-VL series [2] and InternVL3 series [41].
- **Leading medical MLLMs**: To situate ClinFusion within the current medical AI landscape, we compare it against a wide array of specialized medical MLLMs. This includes recent models such as Lingshu [6], Hulu-Med [7], MedGemma [8], HuatuoGPT-V [12], LLaVA-Med [9], BioMediX2 [42], HealthGPT [43], MedDr [44], BiomedGPT [13], Med-R1 [11] and MedVLM-R1 [10].
- **Leading proprietary models**: We also include results from leading proprietary models, including GPT-5.2, Gemini-3-Flash, Claude-Sonnet-4.5, to benchmark against the current frontier of multimodal AI.

To ensure a fair and reproducible comparison, we utilize a unified evaluation kit for leading open-source models. For models where official checkpoints are available, we re-evaluate their performance on our full benchmark suite using this standardized protocol. For 3D volumetric inputs, we by default sample 32 slices per volume, each resized to 256×256, unless a model's official inference code prescribes a different setting, in which case we follow its native configuration (*e.g.*, Hulu-Med [7], which processes all slices). For 2D benchmarks, all available images in each sample are provided to the model without subsampling.

Benchmarking suite. We assess all models across a diverse suite of benchmarks (detailed in Section 4.2 and Extended Data Table 2) categorized as follows:

- **2D multimodal medical benchmarks:** Standard VQA and report generation tasks on 2D medical images.
- **3D multimodal medical benchmarks:** Tasks requiring understanding of volumetric data like CT/MRI scans.
- **Textual medical benchmarks:** Standard medical question-answering datasets to evaluate the underlying language and knowledge capabilities of the LLM component.

- **Medical instruction following benchmark (our MedIF-Bench):** Our proposed benchmark to specifically measure the model's ability to follow complex, clinically-relevant instructions.

Evaluation metrics. For VQA benchmarks (both 2D and 3D), we report accuracy for multiple-choice questions (MCQ) and open-ended questions. For report generation benchmarks, we employ our proposed RoI-grounded evaluation methodology (Section 4.2.3), reporting Precision, Recall, and F1 scores computed via an LLM-as-a-Judge that assesses clinical correctness within specified anatomical regions. For textual medical benchmarks, we report accuracy. For MedIF-Bench, we report an Overall Instruction-Following (Overall-IF) score that measures the percentage of responses conforming to the prescribed output format via regex-based verification, isolating format compliance as a prerequisite for reliable clinical deployment.

2.2 Benchmark Evaluation

2.2.1 Evaluation on 2D Multimodal Medical Benchmarks

Leading VQA capabilities among medical MLLMs, competitive with proprietary models. ClinFusion demonstrates state-of-the-art performance across a comprehensive suite of 2D medical benchmarks, substantially outperforming both leading open-source models and strong proprietary APIs in report generation and challenging VQA tasks. Our evaluation is divided into two key areas: a diverse set of VQA benchmarks and the complex generative task of clinical report generation.

On VQA tasks, ClinFusion consistently ranks first or second across all eight benchmarks, showcasing its robust and versatile visual reasoning capabilities. Fig. 3a compares ClinFusion against leading proprietary, general-purpose, and medical MLLMs on eight 2D benchmarks (detailed scores in Supplementary Table 1). Our ClinFusion-8B model achieves top scores among all medical MLLMs on seven of the eight benchmarks, including the challenging multi-modality datasets that require a comprehensive understanding of all visual modalities and the reasoning-focused MedXpertQA. When compared to proprietary models, ClinFusion-8B and ClinFusion-32B outperform the powerful Gemini-3-Flash on six benchmarks (OmniMedVQA, PMC-VQA, MedFrameQA, VQA-RAD, SLAKE, PathVQA), demonstrating that a specialized, well-designed open-source model can surpass even the most advanced generalist APIs on domain-specific tasks. For extremely hard reasoning benchmarks like MedXpertQA, we observe a margin between our model and the proprietary model. We anticipate that this gap can be closed through scaling the model since we observe a clear performance boost by scaling ClinFusion-8B to ClinFusion-32B.

Leading 2D clinical report generation capabilities among medical MLLMs and proprietary models. In clinical report generation, ClinFusion also sets a new state of the art. As shown in Fig. 3b (detailed scores in Supplementary Table 2), our ClinFusion-8B achieves an F1 score of 37.8 on CheXpert-Plus and a remarkable 57.3 on IU-XRAY. This represents a substantial improvement over the previous small open-source medical MLLM, Hulu-Med-7B (31.9 and 46.5, respectively), showcasing our model's superior ability to generate clinically accurate and coherent text using the fundamentally enhanced vision capabilities. When compared with the proprietary models, all versions of our model are on the same level or better than these proprietary models, especially on the CheXpert-Plus dataset, which ClinFusion clearly outperforms them.

A qualitative example in Extended Data Fig. 1a further illustrates ClinFusion's 2D diagnostic capability. Despite the clear presence of cardiomegaly and pulmonary vascular prominence in the chest X-ray,multiple strong baselines—including Hulu-Med, Lingshu, and Gemini-3-Flash—erroneously generate "normal" reports, likely due to over-reliance on global features. In contrast, ClinFusion accurately identifies both pathologies, correctly describing the "cardiac silhouette" as "mildly enlarged" and noting the "mild prominence of the pulmonary vasculature"—a fine-grained perception enabled by its compositional vision encoder architecture.

2.2.2 Evaluation on 3D Multimodal Medical Benchmarks

Strong 3D multimodal VQA capabilities among medical MLLMs and proprietary models. ClinFusion achieves the best or near-best results in native 3D medical understanding across volumetric VQA benchmarks, as shown in Fig. 4a (detailed scores in Supplementary Table 4). This performance stems from our vision-centric architecture, featuring a dedicated 3D encoder and the 2D-anchored CaSL Fusion mechanism for volumetric reasoning.

Our ClinFusion-8B model, despite its modest size, substantially outperforms much larger models. On AMOS-MCQ, it achieves 80.2, surpassing the previous best 7B model, Hulu-Med-7B (65.7), by 14.5 points, and even exceeding the 32B Hulu-Med (73.9). When compared to proprietary APIs, ClinFusion-8B outperforms Gemini-3-Flash by 16 points on AMOS-MCQ. Our ClinFusion-32B model further extends this advantage, achieving 81.7 on AMOS-MCQ and 89.0 on CT-Rate MCQ, while remaining competitive with Hulu-Med-32B on 3D-RAD.

Strong 3D report generation capabilities among medical MLLMs and proprietary models. Fig. 4b reports Precision, Recall and F1 on 3D report generation under the RoI-Grounded protocol (detailed scores in Supplementary Table 4). On CT-Rate Report, ClinFusion-32B achieves the best F1 (23.9) across all evaluated models, surpassing the strongest medical baseline Hulu-Med-32B (23.4) and all proprietary APIs (Gemini-3-Flash 20.2; GPT-5.2 14.1). On AMOS Report, ClinFusion-32B obtains the best F1 (16.1) among open-source medical MLLMs, while Gemini-3-Flash (23.4) leads overall, indicating headroom on CT reporting that we leave to future work. At the smaller scale, ClinFusion-8B (F1: 12.0 / 21.6) consistently outperforms its 7B medical counterpart Hulu-Med-7B (10.7 / 18.3) on both datasets.

Beyond quantitative metrics, a qualitative analysis on the native CT case in Extended Data Fig. 1b further illustrates how ClinFusion excels in both diagnostic accuracy and clinical workflow adherence. For *diagnostic accuracy*, while the proprietary Gemini-3-Flash generates a generic report incorrectly stating the liver is "normal in size and attenuation" and missing the key diagnosis—a common failure of generalist MLLMs on subtle, diffuse density changes in volumetric data—ClinFusion correctly identifies the decreased liver density and reaches the final "Impression: Mild fatty liver". For *clinical workflow adherence*, ClinFusion's report is meticulously structured according to the requested "Areas of Focus", systematically evaluating each organ system. It further provides specific pertinent negatives, such as "The intrahepatic and extrahepatic bile ducts are not dilated" and "No abnormalities are observed in the gallbladder", rather than the ground-truth (GT) report's general statement that "upper abdominal organs are normal". This combination of pinpointing pathology and systematically confirming normalcy yields reports that are accurate, structured, and clinically coherent.

2.2.3 Evaluation on MedIF-Bench

Strong instruction-following preserved through medical adaptation. To address the critical gap between benchmark performance and practical utility, we evaluated models on our newly proposed MedIF-Bench, which is designed to test complex clinical instructions. As shown in Fig. 4c (detailed scores in Supplementary Table 5), ClinFusion achieves Overall-IF scores of 98.1 (8B) and 98.9 (32B), the best results among medical MLLMs and surpassing all proprietary systems including GPT-5.2 (96.0), Gemini-3-Flash (96.6), and Claude-Sonnet-4.5 (86.5). More importantly, while general-purpose MLLMs such as Qwen2.5-VL and Qwen3-VL exhibit strong instruction-following, existing medical adaptations of these backbones (e.g., Lingshu-7B: 80.4; Lingshu-32B: 82.6) suffer substantial degradation. ClinFusion is a medical-adapted model that retains near-backbone instruction-following capability,with consistently high scores across diverse tasks (detailed scores in Supplementary Table 5) including challenging ones such as "Seer" (prognosis based on specific conditions) and structured report generation where most other medical models fail.

Instruction-following ability is a prerequisite for reliable evaluation. MedIF-Bench also serves as a proxy for model "evaluability": a model's inability to adhere to instructional formats leads to inaccurate assessments of its actual knowledge. The case of MedGemma-1.5-4B-IT illustrates this—it achieves an Overall-IF of only 27.4, and manual inspection confirms that it consistently fails to produce answers in prescribed formats, rendering rule-based and LLM-as-a-judge evaluation unreliable. This deficiency directly manifests as artificially low scores on other benchmarks (e.g., PMC-VQA: 18.7, Supplementary Table 1; SuperGPQA: 2.25, Supplementary Table 3), demonstrating that instruction-following is a prerequisite for evaluating the latent knowledge of any MLLM. Notably, MedGemma-1.5-4B-IT is trained on diverse medical data yet lacks IF capability, suggesting that data curation alone is insufficient if the training methodology does not explicitly instill robust instruction-following.

2.2.4 Evaluation on Textual Medical Benchmarks

Vision-centric architecture need not sacrifice textual competence. Despite being a vision-centric architecture, ClinFusion's textual abilities match or surpass the best medical MLLMs under unified evaluation, as shown in Extended Data Fig. 2 (detailed scores in Supplementary Table 3). Our ClinFusion-32B achieves the best scores among all medical MLLMs on seven of eight benchmarks, with clear improvements over Hulu-Med-32B on MedXpertQA (26.7 vs. 19.8), Medbullets (74.8 vs. 67.5), and MedQA-MCMLE (93.8 vs. 87.0). At the smaller scale, ClinFusion-8B is competitive with Hulu-Med-7B, achieving the best scores on PubMedQA (77.6), MedQA-MCMLE (89.1), and MMLU-Med (81.3) while trailing on MedMCQA and MedQA-USMLE. When compared to general-purpose MLLMs, ClinFusion consistently outperforms its corresponding backbones, demonstrating that medical adaptation enhances rather than compromises textual knowledge. We note a remaining gap with the strongest proprietary models, particularly on complex reasoning benchmarks such as MedXpertQA and SuperGPQA, though this gap narrows substantially when scaling from 8B to 32B (e.g., MedXpertQA: 20.0 $\rightarrow$ 26.7; SuperGPQA: 32.6 $\rightarrow$ 41.8), suggesting that further scaling can close this gap.

2.3 Clinical Validation with Expert Radiologists

To complement the automatic evaluation presented above, and following methodological precedents for expert validation of AI-generated reports [45], we conducted a blinded expert evaluation with six board-certified radiologists (each with $\geq$6 years of clinical experience or longer). We randomly sampled 300 cases from our report generation benchmarks (100 Chest X-ray from [46], 100 Chest CT from 3D-RAD [47] and 100 Abdominal CT from AMOS-MM [48]). Then we collected reports from four systems corresponding to these cases: ClinFusion with agentic tools, ClinFusion (standalone), Gemini-3-Flash, and Hulu-Med. For each case, the evaluators examined the original medical image alongside the four anonymized model outputs, then ranked all four reports along three clinically grounded dimensions: *Factual Accuracy* (how few hallucinated findings), *Completeness* (how few missed abnormalities), and *Clinical Utility* (overall value for downstream decision-making).

Experts validate ClinFusion's clinical advantage and the benefit of agentic augmentation. Fig. 5a summarizes the aggregated Borda scores across all annotators. ClinFusion with agentic tools ranks first in every dimension—Overall, Accuracy, Completeness, and Operability—with a statistically significant margin over Hulu-Med ($p < 0.001$) and Gemini-3-Flash ($p < 0.001$). Notably, the standalone ClinFusion already outperforms both Gemini-3-Flash and Hulu-Med, and adding the agentic tools yields a further consistent uplift, corroborating the quantitative gains reported in Tables 2 to 4.

The RoI-grounded report generation metric achieves the highest correlation with expert judgment. We assessed how faithfully each automatic metric reflects the clinical quality perceived by radiologists, as shown in Fig. 5b. Among eleven metrics—including METEOR [31], BLEU-4 [49], ROUGE-L [30], RadGraph-F1 [50], BERTScore [33], CIDEr [32], and GREEN [36]—our proposed metric achieves the highest rank correlation with expert consensus on all three agreement measures (Kendall's $\tau = 0.511$, Spearman's $\rho = 0.572$, Top-1 Accuracy $= 55.5\%$). At the individual-case level, as shown in Fig. 5c, the per-case Kendall's τ distribution of our metric exhibits a higher mean and lower variance than all alternatives, indicating that its advantage is systematic rather than driven by outliers.

Inter-annotator agreement confirms the reliability of the evaluation. To verify the scientific rigor of the annotation study, we assessed inter-annotator agreement using Kendall's coefficient of concordance (W). As shown in Fig. 5d, the overall W across all dimensions is 0.665 ± 0.275 ($n = 300$), indicating substantial agreement among the radiologists. As shown in Fig. 5e, agreement is consistent across modalities (CT: 0.678 ± 0.260; X-ray: 0.641 ± 0.300) and across individual dimensions, with Accuracy reaching $W = 0.687 \pm 0.272$. These values confirm that the expert rankings are reproducible and that the observed model differences reflect genuine quality gaps rather than annotator noise.

2.4 Design Analysis

This section validates our two core design contributions: the RoI-grounded evaluation methodology and the compositional vision architecture.

2.4.1 Analysis of RoI-Grounded Report Generation

To validate the design choices behind our RoI-Grounded Report Generation evaluation—which achieves the highest correlation with expert judgment among eleven metrics (Fig. 5b)—we present a detailed analysis comparing our methodology against RadGraph-F1 [50], a widely adopted clinical metric for report generation, as employed in recent work [8]. We identify two key limitations of this conventional approach and show how our protocol addresses them.

Clinical context ensures meaningful report generation comparisons. In a typical prior evaluation setup, such as the one employed by MedGemma [8], the model is prompted to generate a report from an image without any clinical context. To demonstrate why this is problematic, we ablate within our RoI-Grounded pipeline (Section 4.2.3) on whether the extracted clinical context ("Clinical Indication" and "Area of Focus") is provided during generation. We compare two models (ClinFusion-8B and Lingshu-7B [6]) across report generation benchmarks.

As shown in Fig. 6a (detailed scores in Supplementary Table 6), the context-free setting reveals two issues: **1)** metrics drop substantially for both models, and **2)** the performance gap between the two models is severely compressed—under the context-aware setting ClinFusion-8B consistently outperforms Lingshu-7B, yet under the context-free setting both produce nearly identical scores. The root cause is that a single medical image contains many anatomical structures, but a radiologist's report sometimes focuses only on areas relevant to a specific clinical indication. Without this context, the model may describe regions absent from the GT report, and there is no way to determine whether such claims are hallucinations or legitimate findings that the original radiologist chose not to document—yet all are penalized as false positives, leading to low report generation scores and narrow performance gaps. By introducing the extracted clinical context, our RoI-Grounded protocol constrains the model to the anatomical areas covered by the GT, ensuring that every claim can be categorized as a True Positive, False Negative, or False Positive.

A representative case is shown in Extended Data Fig. 3: With clinical context (left), ClinFusion's output stays within the anatomical scope of the GT (pleural effusion, pericardial effusion, lung consolidations consistent with Covid-19 pneumonia, and liver findings), so each claim is verifiable. Without context (middle), the same model produces additional findings on regions the GT never addresses—*e.g.*, mediastinal shift, cardiomegaly, splenomegaly, anasarca, osseous lesions, and a diagnosis of decompensated congestive heart failure—which cannot be distinguished from hallucinations and are uniformly penalized as false positives, deflating the score.

LLM-as-a-Judge evaluation overcomes the synonymy insensitivity and length bias. As an alternative to text-matching metrics such as RadGraph-F1 [34, 8], our RoI-Grounded protocol adopts an LLM-as-a-Judge for report evaluation. RadGraph-F1 first parses both the predicted and GT reports through RadGraph-XL [50]—a transformer trained to extract clinical entities (*e.g.*, anatomical structures, observations) and their relations (*e.g.*, *suggestive of*, *located at*) from radiology text—and then computes an F1 score over the matched entity and relation sets. Critically, RadGraph-XL operates in a *closed-form* fashion: it maps input text to a fixed,pre-defined schema, unlike LLM-based evaluators that accept and produce open-form natural language;furthermore, as a relatively small extractor, it has limited accuracy on out-of-distribution clinical phrasings. This rigidity enables an LLM-free, computationally cheap pipeline, but introduces two failure modes. **1) Synonymy insensitivity:** Semantically equivalent but differently worded findings receive no credit—for example, *"the liver appears enlarged"* fails to match *"hepatomegaly is present"* with a score of zero, despite being clinically identical, as they map to different schema entries with no token overlap. **2) Length bias:** Because RadGraph-XL extracts entities from the *entire* input text, longer outputs have a greater statistical chance of overlapping with the GT entities, so even purely verbose language that casually mentions anatomical terms can inflate the score without reflecting clinical accuracy.

Fig. 6b (detailed scores in Supplementary Table 7) provides quantitative evidence of both failure modes. First, RadGraph-F1 scores cluster within a narrow range across the first three models (11.8–21.0 on CheXpert-Plus; 16.3–36.6 on IU-XRAY), making it difficult to distinguish model quality, whereas our LLM-as-a-Judge spans a wider and more discriminative range (16.8–37.8 on CheXpert-Plus; 36.7–57.3 on IU-XRAY). Second, the Long Output variant substantially inflates RadGraph-F1 (11.8→18.8 on CheXpert-Plus; 16.3→33.0 on IU-XRAY) despite producing lower-quality outputs, while our LLM-as-a-Judge correctly penalizes verbosity (16.8→15.1 on CheXpert-Plus; 36.7→33.6 on IU-XRAY).

2.4.2 Ablation Studies on Architectural Design

All ablation studies on architectural designs (*e.g.*, the contribution of different encoders or fusion strategies) are conducted on the 2D and 3D multimodal medical benchmarks, as these tasks are most sensitive to changes in visual perception. To enable rapid iteration, we train on a subset of the full training data and, given the high computational cost of 3D evaluation, also sample a subset of the 3D evaluation data. We ensure that every comparison within the same table or figure is performed under identical settings.

Local cross-attention is a better fusion mechanism than other alternatives. We compare our proposed CaSL Fusion operator against three alternatives: **1)** channel-wise concatenation, which concatenates feature maps along the channel dimension; **2)** global cross-attention, where each query token attends to the entirety of the key/value feature map, in contrast to our localized approach; and **3)** a mixture-of-vision-experts approach, which computes a token-wise weighted combination of features via a learned gating mechanism [51]. The model is configured with Qwen ViT as the foundational encoder and DINOv2 as the additional vision encoder. To facilitate efficient validation, the model is trained for the first three training stages on a subset of the full corpus. As shown in Fig. 6c (detailed scores in Supplementary Table 8), all evaluated fusion strategies improve over the single-encoder baseline (Qwen ViT only), supporting our hypothesis that a well-aligned foundational encoder can be enhanced by complementary features from a specialist model. Among the four mechanisms, local cross-attention attains the highest average performance, and we adopt it as the default in all subsequent ablation studies.

DINOv2 augments the Qwen ViT the most. We next identify the most effective single specialist encoder to augment the foundational Qwen ViT, comparing DINOv2, ConvNext, and Medsiglip. As shown in Fig. 6d (detailed scores in Supplementary Table 9), no single encoder dominates both tasks: Medsiglip leads on 2D VQA (64.8 vs. 64.5 / 64.1) while ConvNext slightly leads on report generation (27.9 vs. 27.8 / 26.6). However, DINOv2 is the only encoder that ranks within 0.3 of the best on both tasks, while ConvNext underperforms on VQA and Medsiglip lags on report generation. We hypothesize that DINOv2's self-supervised pre-training yields more general-purpose visual features that transfer well across both task types, rather than specializing in either. We therefore adopt DINOv2 as the specialist encoder in all subsequent multi-encoder experiments.

Additional vision encoders complement Qwen ViT, with diminishing returns beyond one additional encoder. Building upon our finding that DINOv2 serves as the most effective single specialist encoder in addition to Qwen ViT, we further explore the optimal combinations of two additional encoders, evaluating ensembles of DINOv2 paired with either ConvNeXt or Medsiglip under different cascade orders. As shown in Fig. 6e (detailed scores in Supplementary Table 10), the DINOv2→Medsiglip cascade attains the highest average VQA score (65.2), while the DINOv2→ConvNeXt cascade attains the highest average score on report generation (28.8). Compared with the best single-encoder configuration (Fig. 6d), the second specialist encoder adds at most 0.4 on VQA and 0.9 on report generation, indicating diminishing returns beyond one additional specialist encoder. Since report generation places higher demands on visual perception—requiring both global understanding and localized findings—we adopt the Qwen ViT + DINOv2 + ConvNeXt combination as the specialist ensemble in our final ClinFusion architecture.

Cascade fusion outperforms parallel fusion on report generation. We additionally compare two ways of combining the same DINOv2 + ConvNeXt encoder pair: our Cascade fusion (DINOv2→ConvNeXt) and a Parallel fusion in which each specialist encoder is independently fused into the Qwen ViT query. As shown in Fig. 6e (detailed scores in Supplementary Table 10), the two strategies perform similarly on 2D VQA (64.4 vs. 64.6), but Cascade fusion clearly outperforms Parallel fusion on report generation (28.8 vs. 27.3). We hypothesize that the cascaded design allows the representation to be progressively enriched across specialist encoders, which benefits the open-ended generation setting more than the closed-form VQA setting. We therefore use Cascade fusion as the default in our final architecture.

The native 3D encoder is essential for 3D volumetric understanding. We quantify the contribution of our dedicated 3D perception module by deactivating the 3D encoder at inference time, forcing the model to rely solely on sparsely sampled 2D slices. As shown in Fig. 6f (detailed scores in Supplementary Table 11), removing the 3D encoder degrades performance consistently across all evaluated tasks: the average 3D VQA score drops by 3.0 points (65.8 → 62.8), and the average 3D report-generation F1 drops by 4.0 points (15.7 → 11.7), with the largest single-task degradations on CT-Rate Open VQA (63.6 → 56.4) and CT-Rate Report F1 (20.3 → 15.5). These results indicate that volumetric features extracted by the native 3D encoder cannot be substituted by

sparse 2D slices alone, supporting the inclusion of a dedicated 3D perception module in ClinFusion.

2.5 Practical Extension with Agentic Tool Use

We extend ClinFusion with an agentic tool-use framework that combines a RAG tool and perception expert tools within a plan–act workflow. We evaluate this extension on text-only medical benchmarks and 2D/3D multimodal benchmarks at both the 8B and 32B scales, and illustrate the workflow on representative text-only and multimodal cases.

Agentic tools improve 2D and 3D multimodal evaluations, with the largest gains on descriptive tasks. On multimodal benchmarks (Fig. 7, left column; detailed scores in Supplementary Table 2 and Supplementary Table 4), agentic tools improve ClinFusion across both 2D report generation and 3D VQA / report tasks at both scales. For ClinFusion-8B, the largest gains appear on tasks requiring descriptive output, including CheXpert-Plus 2D report (+12.4), 3D-RAD Open VQA (+8.4), AMOS Report (+4.6), and CT-Rate Report (+3.7), while gains on closed-form 3D MCQ tasks are smaller (AMOS MCQ +5.5, 3D-RAD MCQ +2.2, CT-Rate MCQ +2.8). ClinFusion-32B follows a similar pattern, with the largest gains again concentrated on descriptive-output tasks (3D-RAD Open VQA +9.4, CheXpert-Plus +6.8, CT-Rate Report +3.7, AMOS Report +3.0) rather than on closed-form MCQ tasks tasks that selects among predefined options.

Agentic tools improve text-only evaluations, with substantial gains on knowledge-intensive benchmarks. On text-only medical benchmarks (Fig. 7, right column; detailed scores in Supplementary Table 3), agentic tools improve ClinFusion consistently at both scales. For ClinFusion-8B, the largest gains appear on the most knowledge-intensive benchmarks, including SuperGPQA (+12.8), MedQA-U (+11.1), MedXpertQA (+7.7), and MedMCQA (+7.3); the only exception is MedQA-M, where the score drops by 4.8 points. We attribute this to the limited Chinese-language coverage in our retrieval corpus, which restricts the utility of RAG on a Chinese-language exam benchmark. For ClinFusion-32B, gains follow a similar pattern (SuperGPQA +9.1, MedXpertQA +4.5, MedQA-U +4.2), and the MedQA-M deficit no longer appears (+0.7), consistent with the stronger multilingual capability of the larger backbone, which makes better use of the predominantly English retrieved evidence.

The agentic framework makes model decisions verifiable through tool-produced evidence. We illustrate the agentic workflow (shown in Extended Data Fig. 4c, detailed in Section 4.4) on two representative cases (Extended Data Fig. 4a,b). In the provided text-only case (Extended Data Fig. 4a), ClinFusion decomposes the clinical question into complementary retrieval queries, invokes the hybrid retrieval engine to collect provenance-aware evidence from heterogeneous sources (*e.g.*, PubMed, textbooks, StatPearls), summarizes the evidence into a traceable intermediate conclusion before generating the final answer. In the provided multimodal case (Extended Data Fig. 4b), ClinFusion identifies the input as a 3D chest CT volume, invokes a Multi-Organ Lesion Segmenter that returns structured findings (*e.g.*, lesion count and size), and integrates these verified outputs into a radiology-style report with *Findings* and *Impression* sections. In both scenarios, the intermediate tool outputs remain exposed and auditable, making the reasoning chain verifiable rather than relying solely on implicit parametric generation.

3 Discussion

The integration of MLLMs into clinical practice holds significant promise, yet this transition has been impeded by fundamental obstacles: the challenge of processing heterogeneous medical data, the gap between academic evaluation and real-world clinical needs, and the static nature of model knowledge. In this work, we introduced ClinFusion, a vision-centric MLLM that provides a systematic and holistic framework to address these critical limitations. Our work presents a systematic framework that addresses these limitations and demonstrates strong benchmark performance, suggesting a viable path toward deploying MLLMs in clinical environments.

At its core, ClinFusion's advancements are built on two principal contributions. Architecturally, we moved beyond monolithic vision encoders by proposing a compositional and cascaded architecture featuring the CaSL Fusion operator. This design unifies diverse 2D and native 3D data processing through a cascaded and 2D-anchored interplay mechanism, creating a rich and synergistic visual representation that consistently benefits nuanced diagnostic tasks, as confirmed by our ablation studies. To bridge the evaluation-practice gap,

we introduced a new vision-grounded evaluation framework. This includes the MedIF-Bench to enforce reliable instruction-following, and a novel RoI-grounded report generation methodology for clinically aligned and factualness-driven report generation evaluation. To further enhance practical deployment, we developed an agentic tool use extension that equips ClinFusion with retrieval-augmented generation and the ability to use external specialist models as tools, demonstrating consistent improvements in both text-only and multimodal clinical scenarios.

Our comprehensive experiments demonstrate that these innovations translate into state-of-the-art performance. ClinFusion not only outperforms leading open-source models (*e.g.*, Hulu-Med and Lingshu) across a wide array of 2D, 3D, and textual benchmarks but also surpasses powerful proprietary APIs on multimodal tasks. The superior performance, particularly in complex multimodal understanding, supports our architectural hypothesis that holistic medical perception benefits from a compositional vision system capable of natively processing heterogeneous 2D and 3D data, rather than relying on a single monolithic encoder. Independently, the strong correlation between our RoI-Grounded evaluation and expert radiologist judgment confirms that vision-grounded, factualness-driven metrics better reflect real clinical needs than conventional text-matching approaches. The consistent improvements observed when augmenting ClinFusion with agentic tool use further suggest that dynamic knowledge retrieval and specialist tool integration can complement the base model's capabilities. Furthermore, the strong results on MedIF-Bench underscore that strong instruction-following capabilities are preserved through medical adaptation—a critical factor for clinical usability and a property often degraded in other medical MLLMs.

The broader implication of our work lies in the paradigm shift it advocates for. Our results suggest that advancing medical AI requires equal attention to both perception architecture and evaluation methodology—building modular, verifiable systems while ensuring that evaluation faithfully reflects clinical needs. ClinFusion illustrates one instantiation of this modular design philosophy. Its compositional architecture allows for flexible integration of new specialist encoders as imaging technology evolves. Its agentic tool use extension demonstrates how continuous knowledge updates and delegation to auditable, high-precision tools can further enhance a strong foundation model for practical deployment. Concurrently, our proposed evaluation paradigm provides the rigorous assessment required to ensure these complex systems are not only capable but also reliable and aligned with clinical workflows.

Despite its strong performance, we acknowledge several limitations that pave the way for future research. First, while our agentic system demonstrates the power of tool use, the current toolset is limited. Future work should expand this ecosystem to encompass a broader range of modalities (*e.g.*, ultrasound, MRI) and functionalities (*e.g.*, treatment planning simulators). Second, despite competitive performance among open-source medical MLLMs, a notable gap remains between ClinFusion and the strong proprietary models (*e.g.*, Gemini-3-Flash) on knowledge-intensive text benchmarks. We attribute this primarily to the capacity and training data scale differences between our ClinFusion and the substantially larger proprietary systems, and note that our agentic tool use extension partially mitigates this gap through knowledge retrieval. Finally, although ClinFusion excels in offline benchmarks, its translation into clinical practice necessitates rigorous human-in-the-loop evaluation to validate its impact on diagnostic accuracy, efficiency, and patient outcomes, while navigating the complex ethical and regulatory landscape.

In conclusion, ClinFusion lays the groundwork for a perceptually powerful, rigorously evaluated and verifiable medical AI system. We hope this work motivates further research to build more clinically-aligned and actionable medical AI systems.

4 Methods

4.1 Architecture

ClinFusion is built upon the Qwen-VL series [2, 1], which is comprised of a Large Language Model (LLM), a redesigned Vision Transformer (ViT) architecture, which we name Qwen ViT in the paper, and a vision-language merger. The core of the perception capabilities stems from the Qwen ViT, pre-trained on vast general-domain data to develop a strong understanding of general visual concepts. However, a single architecture, even one as capable as Qwen ViT, is challenging for capturing the full spectrum of visual information present in medical imaging, which ranges from high-frequency textural details (crucial for pathology) to global structural semantics (for organ anatomy). We drew inspiration from recent works demonstrating that a composition of models can learn distributions more data-efficiently [52] and that a strategic composition of encoders can induce significantly stronger capabilities [26]. Building on this philosophy, ClinFusion extends the Qwen-VL architecture by introducing a compositional ensemble of specialized vision encoders that synergistically fuse diverse native 2D and 3D representations, enabling a more holistic and robust perception of medical imagery, as shown in Fig. 2a and Fig. 2b.

4.1.1 Perception with Compositional Vision Encoders

Enhancing 2D perception with additional 2D vision encoders. To construct a comprehensive 2D perception system, we create a compositional ensemble of vision encoders, where each model is selected to complement Qwen ViT's capability, addressing the inherent limitations of the generalist Qwen ViT. Our two adopted specialist encoders are as follows: **1)** *A specialist for local features*: We incorporate ConvNeXt [53], a modern convolutional network. Its strong inductive bias for spatial locality makes it exceptionally skilled at capturing high-frequency textural details and local patterns, which are critical for interpreting pathological slides or subtle radiological findings. **2)** *A specialist for robust semantics*: We include DINOv2 [54], a powerful self-supervised model. By learning from visual data without explicit language supervision, it develops robust, high-level semantic representations that are less prone to dataset biases, enhancing its ability to generalize across diverse medical imaging modalities. This choice is further motivated by recent evidence that self-supervised encoders can scale effectively for medical image understanding even without text supervision [55]. We additionally evaluated MedSigLIP [8] as a candidate domain-knowledge specialist, as its pre-training on large-scale medical image–text pairs (similar in spirit to BiomedCLIP [56]) can inject clinical domain knowledge directly. However, our ablations (Supplementary Table 10) show it does not offer a consistent advantage, and we ultimately adopt DINOv2 for its stronger overall balance across VQA and report generation. We note that clinical domain knowledge is instead injected into the model through progressive staged training (Section 4.5).

Compressing representations with CaSL Fusion $\bowtie$. To effectively fuse features from the compositional encoder ensemble, we design our fusion strategy around two core principles: **1) Principle 1: Information density over token length**: We prioritize increasing the information density of the visual tokens fed into the LLM rather than increasing the number of tokens. This is motivated by findings that standard visual representations in MLLMs often contain significant redundancy, and compressing richer information into a fixed-length sequence is more efficient [57, 58]. **2) Principle 2: Guided alignment via the foundation encoder**: We should leverage the strong, pre-existing alignment between the Qwen ViT and the LLM decoder. Therefore, the representations from newly introduced encoders should be fused in a way that guides them towards this established vision-language space, rather than disrupting it.

Let the set of encoders be denoted by $\{\mathcal{E}_i\}_{i=1}^{M}$, where $\mathcal{E}_1$ is always the foundational Qwen ViT. For an input image $\boldsymbol{I}$, each encoder $\mathcal{E}_i$ extracts a feature map $\boldsymbol{X}_i$. To harmonize the features, which originate from different embedding spaces and pre-processing pipelines, we first project and resize them:

$$\boldsymbol{X}_i = \boldsymbol{\mathcal{E}}_i(\boldsymbol{I}); \quad \bar{\boldsymbol{X}}_i = \text{Resize}(\boldsymbol{W}_{\text{align}}\boldsymbol{X}_i) \in \mathbb{R}^{N\times D} \tag{1}$$

Here, $\boldsymbol{W}_{\text{align}}$ is a learnable projection matrix that maps the features from encoder $\mathcal{E}_i$ to the hidden dimension D of the Qwen ViT. The Resize operation aligns each feature map to Qwen ViT's spatial shape of $H \times W$, where $N = H \times W$. One exception is ConvNeXt, whose feature map is instead resized to a larger spatial resolution to better preserve high-frequency detail; we defer this encoder-specific treatment to Section 4.6.

To achieve this fusion, we propose the Cascade Spatial-Aware Locality Fusion (CaSL Fusion), denoted by the operator $\bowtie$. The core idea of CaSL is to incrementally enrich the foundational Qwen ViT representation with features from the specialist encoders in a spatially-aware and computationally efficient manner. This is realized through a cascaded application of a local cross-attention mechanism, as shown in Fig. 2c.

We exemplify the process with the fusion of two feature maps, $\widetilde{\boldsymbol{X}}_1 = \bar{\boldsymbol{X}}_1 \bowtie \bar{\boldsymbol{X}}_2$. We first define a spatial local field extractor, $\mathcal{N}_{2\mathrm{D}}$, which extracts a $k \times k$ patch of neighboring tokens (the window size k is detailed in Section 4.6). For the j-th feature token $\bar{\boldsymbol{X}}_1[j]$, we treat it as a query $\boldsymbol{q}_j = \bar{\boldsymbol{X}}_1[j] \in \mathbb{R}^{1\times D}$. The corresponding keys and values $\boldsymbol{K}_j$ and $\boldsymbol{V}_j$ are extracted from the local neighborhood centered at the same spatial location j within the second feature map $\bar{\boldsymbol{X}}_2$:

$$\boldsymbol{K}_j, \boldsymbol{V}_j = \mathcal{N}_{2\mathrm{D}}(\bar{\boldsymbol{X}}_2)_j = \left\{ \bar{\boldsymbol{X}}_2[p] \mid p \in \mathcal{B}_{2\mathrm{D}}(j) \right\} \in \mathbb{R}^{k^2\times D}, \quad \forall j \in \{1, \ldots, N\} \tag{2}$$

where $\mathcal{B}_{2\mathrm{D}}(j)$ contains the index set around the j-th feature in the spatial dimension. This local extraction enforces a strong spatial inductive bias and dramatically reduces the computational cost from quadratic (over N) to linear. The fusion is then performed via a standard cross-attention operation [59]:

$$\boldsymbol{A}_j = \mathrm{softmax}\left(\frac{\boldsymbol{q}_j \boldsymbol{K}_j^\top}{\sqrt{D}} \right) \in \mathbb{R}^{1\times k^2}; \quad \boldsymbol{o}_j = \boldsymbol{A}_j \boldsymbol{V}_j \in \mathbb{R}^{1\times D} \tag{3}$$

Note that we omit the multi-head formulation and the projections for queries, keys, values and outputs for clarity. The aggregated feature $\boldsymbol{o}_j$ is then integrated back into the query representation using a residual connection and a Feed-Forward Network (FFN), forming a complete fusion block:

$$\widetilde{\boldsymbol{X}}_1' = \bar{\boldsymbol{X}}_1 + \mathrm{2DLocalCrossAttn}(\bar{\boldsymbol{X}}_1, \bar{\boldsymbol{X}}_2; k); \quad \widetilde{\boldsymbol{X}}_1 = \widetilde{\boldsymbol{X}}_1' + \mathrm{FFN}(\widetilde{\boldsymbol{X}}_1') \tag{4}$$

where $\mathrm{2DLocalCrossAttn}(\cdot)$ denotes the process in equation (3) for all tokens.

For an ensemble of M encoders, the fusion is applied in a cascaded manner, sequentially enriching the base representation:

$$\boldsymbol{Z} = \left(\left(\bar{\boldsymbol{X}}_1 \bowtie \bar{\boldsymbol{X}}_2 \right) \ldots \bowtie \bar{\boldsymbol{X}}_M \right) \tag{5}$$

Crucially, the fusion operator $\bowtie$ is *asymmetric and left-associative*, where the left operand always serves as the query. This ensures that the well-aligned Qwen ViT representation remains the central backbone of the fusion process, which is progressively enhanced by the specialist encoders. The fused feature $\boldsymbol{Z}$ replaces the original $\bar{\boldsymbol{X}}_1$ and is fed into the LLM decoder.

CaSL Fusion with stochastic residual regularization. A potential challenge in our cascaded fusion design is the risk of optimization shortcuts. Since the Qwen ViT ($\bar{\boldsymbol{X}}_1$) is already well-aligned with the LLM, the model might learn to predominantly rely on this path during fine-tuning, effectively ignoring the contributions from less-aligned specialist encoders. To mitigate this, we introduce a stochastic regularization technique during training. Specifically, we modify the residual connection in the fusion block (equation (10)) by stochastically dropping the original query representation. The update rule becomes:

$$\tilde{\boldsymbol{X}}_1' = \lambda \cdot \bar{\boldsymbol{X}}_1 + \mathrm{2DLocalCrossAttn}(\bar{\boldsymbol{X}}_1, \bar{\boldsymbol{X}}_2; k); \quad \lambda \sim \mathrm{Bernoulli}(1 - p_{\mathrm{drop}}) \tag{6}$$

During training, this forces the model to learn meaningful representations from the specialist encoders' features. During inference, this stochasticity is disabled by setting $\lambda = 1$ and $p_{\mathrm{drop}} = 0$, ensuring stable and full utilization of all features. p_{drop} is set to 0.1 by default.

CaSL Fusion's compatibility with DeepStack. Qwen3-VL [2] introduced a DeepStack mechanism [60], which injects visual features into multiple, deeper layers of the LLM to avoid the dilution of visual information, enabling a more profound and persistent vision-language alignment. CaSL Fusion is fully compatible with and designed to leverage this deep fusion paradigm. We denote the feature map from the foundational Qwen ViT ($\mathcal{E}_1$) at the l-th DeepStack layer as $\bar{\boldsymbol{X}}_1^{(l)}$, where $l \in \{1, \ldots, L\}$. At each layer l, the layer-specific Qwen ViT feature, $\bar{\boldsymbol{X}}_1^{(l)}$, serves as the query for a new cascaded fusion. The specialist features are then sequentially fused with this query. This process yields a set of L enriched DeepStack features $\{\boldsymbol{Z}^{(l)}\}_{l=1}^{L}$. The computation can be formulated as:

$$\boldsymbol{Z}^{(l)} = \left(\left(\bar{\boldsymbol{X}}_1^{(l)} \bowtie \bar{\boldsymbol{X}}_2 \right) \ldots \bowtie \bar{\boldsymbol{X}}_M \right) \quad \forall l \in \{1, \ldots, L\} \tag{7}$$

$\boldsymbol{Z}^{(l)}$ then replaces the original DeepStack feature $\bar{\boldsymbol{X}}_1^{(l)}$ and is injected into the LLM. By enabling CaSL Fusion with DeepStack, we ensure that the LLM's reasoning remains continuously grounded in the visual evidence.

4.1.2 2D-Anchored Native 3D Volumetric Understanding

A limitation of existing medical MLLMs is their inability to natively process 3D volumetric data such as CT and MRI scans. Some methods [6, 7] treat 3D volumes as a sequence of independent 2D slices, leading to the loss of critical inter-slice spatial context and anatomical continuity. ClinFusion addresses this by integrating a dedicated 3D encoder and proposing a 2D-anchored 3D CaSL Fusion to enrich 2D anchor features with 3D volumetric features. Our key motivation behind this strategy is twofold: **1) Accelerated alignment**: By leveraging the strong language alignment of the 2D features as anchors, the 3D features can rapidly converge to the vision-language space without requiring extensive 3D-specific alignment data; **2) Knowledge transfer**: Even when processing 3D data, the model benefits from the rich pre-trained knowledge embedded in the 2D encoders, enabling robust generalization across diverse volumetric imaging modalities.

Dual-path processing with 2D anchors. For a given 3D volume, we employ a dual-path processing strategy.

Path 1: Native 3D encoding. We incorporate a 3D vision encoder $\mathcal{E}_{\text{3D}}(\cdot)$ [61] and we pre-align it with language through contrastive learning on 3D medical volumes and their corresponding radiology reports, following a paradigm that has proven effective for both 2D radiographs [62] and 3D oncology imaging [23]. This pre-alignment endows the encoder with an intrinsic understanding of 3D anatomical structures while establishing a basic connection to the language space. Specifically, we process the entire 3D volume using 3D encoder and obtain a 3D feature map $\boldsymbol{P} = \mathcal{E}_{\text{3D}}(\boldsymbol{I}_{\text{3D}})$, where $\boldsymbol{I}_{\text{3D}}$ denotes the 3D volume used.

Path 2: 2D anchor encoding. Simultaneously, a set of S representative 2D slices $\mathcal{S}$ are randomly sampled from the input volume $\boldsymbol{I}_{\text{3D}}$ (we set $S = 4$, used identically at training and inference). The anchor slices are thus denoted as $\{\boldsymbol{I}_{\text{3D}}[s] \mid s \in \mathcal{S}\}$. The 2D features (after projection and resize) for these sampled frames can be denoted as: $\{\bar{\boldsymbol{X}}_1[s], \bar{\boldsymbol{X}}_2[s], \ldots, \bar{\boldsymbol{X}}_M[s] \mid s \in \mathcal{S}\}$. Utilizing the CaSL Fusion operator in equation (5), we can obtain their fused 2D features, $\boldsymbol{Z}[s] \in \mathbb{R}^{N \times D}$, for all slices in S:

$$\boldsymbol{Z}[s] = \left(\left(\bar{\boldsymbol{X}}_1[s] \bowtie \bar{\boldsymbol{X}}_2[s]\right) \ldots \bowtie \bar{\boldsymbol{X}}_M[s]\right), \quad \forall s \in \mathcal{S} \tag{8}$$

This process generates a set of anchor feature maps, each rich with 2D-aligned semantics, ready to guide the fusion with the native 3D representation.

2D-anchored depth-aware CaSL Fusion $\underset{\text{3D}}{\bowtie}$. To effectively fuse the 3D features with the 2D anchor representations, we extend CaSL Fusion to a depth-aware variant. The key insight is that each 2D anchor slice should attend to the 3D features not only in its spatial neighborhood but also across the depth dimension, thereby capturing the volumetric context that surrounds each slice.

Following the projection and resize operation mentioned in equation (1), we also perform this operation for the 3D feature $\boldsymbol{P}$, obtaining $\bar{\boldsymbol{P}} \in \mathbb{R}^{D_{\text{vol}} \times N \times D}$, where its spatial shape aligns with the spatial size of $\boldsymbol{Z}[s] \in \mathbb{R}^{N \times D}$ and D_{vol} represents 3D feature volume's depth ($D_{\text{vol}} = 4$, obtained from a 32-frame input with a downsampling factor of 8 along the depth dimension). We then exemplify the process with the fusion of 2D anchor features with 3D features, $\widetilde{\boldsymbol{Z}}[\boldsymbol{s}] = \boldsymbol{Z}[\boldsymbol{s}] \underset{\text{3D}}{\bowtie} \bar{\boldsymbol{P}}$. We define a depth-aware spatial local field extractor, $\mathcal{N}_{\text{3D}}$, that extracts a patch spanning a $k \times k$ spatial area across the full depth D_{vol}. For the j-th feature token in $\boldsymbol{Z}[s]$, we treat it as a query $\boldsymbol{q}_j = \boldsymbol{Z}[s, j] \in \mathbb{R}^{1 \times D}$. The corresponding keys and values $\boldsymbol{K}_j$ and $\boldsymbol{V}_j$ are extracted using depth-aware spatial local field extractor:

$$\boldsymbol{K}_j, \boldsymbol{V}_j = \mathcal{N}_{\text{3D}}(\bar{\boldsymbol{P}})_j = \left\{\bar{\boldsymbol{P}}[p] \mid p \in \mathcal{B}_{\text{3D}}(j)\right\} \in \mathbb{R}^{D_{\text{vol}} \times k^2 \times D}, \quad \forall j \in \{1, \ldots, N\} \tag{9}$$

where $\mathcal{B}_{\text{3D}}(j)$ contains the index set around the j-th feature in the spatial and depth dimension. After the definition of query, keys and values, we can perform the attention operation defined in equation (3) and residual connections defined in equation (10):

$$\widetilde{\boldsymbol{Z}}[\boldsymbol{s}]' = \boldsymbol{Z}[s] + \text{3DLocalCrossAttn}(\boldsymbol{Z}[s], \bar{\boldsymbol{P}}; k, D_{\text{vol}}); \quad \widetilde{\boldsymbol{Z}}[\boldsymbol{s}] = \widetilde{\boldsymbol{Z}}[\boldsymbol{s}]' + \text{FFN}(\widetilde{\boldsymbol{Z}}[\boldsymbol{s}]') \tag{10}$$

The obtained features $\widetilde{\boldsymbol{Z}}[\boldsymbol{s}], \forall s \in \mathcal{S}$ are concatenated to compose a full sequence of length $S \times N$ and fed into the LLM decoder to perform inference. As with 2D CaSL Fusion $\bowtie$, $\underset{\text{3D}}{\bowtie}$ is also compatible with DeepStack and stochastic residual regularization to enable robust feature learning.

4.2 Vision-Grounded Evaluation System

To evaluate whether a medical multimodal LLM can function as a useful clinical tool, we introduce a comprehensive evaluation system that mirrors the integrated skills of a radiologist. As detailed in Extended Data Table 2, this system builds on prior frameworks like MedGemma [8], Lingshu [6], HuLu-Med [7], and HuatuoGPT [63] to assess five critical dimensions: foundational 2D and 3D visual understanding (evaluated by 2D and 3D VQA tasks), text-based medical knowledge (evaluated by QA tasks), clinical report generation, and medical instruction-following. It is worth noting that we extend prior frameworks by incorporating medical instruction-following evaluation, as this capability is overlooked in all existing benchmarks yet critical for determining model usability. In the following subsections, we describe the evaluation protocol for each benchmark category.

4.2.1 Evaluation Prompt Templates for VQA and QA

For 2D and 3D medical VQA, as well as text-only QA problems in the benchmark, we adopt a set of carefully designed prompts and evaluation procedures. The complete prompt templates are shown in Extended Data Fig. 5a.

Multiple Choice Questions (MCQ). As shown in the left panel of Extended Data Fig. 5a, the MCQ prompt instructs the model to answer a multiple-choice question based on the provided image(s). Note that for text-only QA tasks, the phrase "based on the provided image(s)" is removed from the prompt. The prompt explicitly states that the model must select exactly one option and output only the single uppercase letter of the correct option enclosed in `<answer></answer>` tags. To reduce ambiguity, the prompt includes both a correct format example (*e.g.*, `<answer>A</answer>`) and several incorrect format examples that would fail automated grading, such as including extra text, missing tags, outputting option text instead of the letter, or selecting multiple letters. We designed this prompt to be strict for three reasons: **1)** we observed that some open-source models have poor instruction-following ability, and simpler prompts often fail to elicit the correct output format, which is essential for reliable rule-based evaluation via regular expression matching; **2)** suppressing unrelated output reduces decoding time and thus speeds up evaluation; and **3)** the prompt differs from the training prompts of most models, so it also serves as a test of generalization to unseen prompts, as recommended by [64]. The extracted letter is then directly compared against the GT answer for scoring.

Open-ended VQA. As shown in the upper-right panel of Extended Data Fig. 5a, the open-ended prompt asks the model to analyze the provided image(s) and give a concise, direct answer to the question. Similarly, for text-only QA tasks, the phrase "based on the provided image(s)" is removed from the prompt. The prompt is kept minimal to avoid biasing the response while still discouraging unnecessary verbosity. Since open-ended answers cannot be evaluated by exact string matching, we employ an LLM-as-a-Judge[2] approach for scoring. The judge prompt, shown in the lower-right panel of Extended Data Fig. 5a, provides the original question, the GT answer, and the model's answer to a language model judge. The judge is instructed to determine whether the model's answer is correct by checking for semantic equivalence with the GT—if the model's answer expresses a similar meaning or a valid alternative interpretation of the standard answer, it is considered correct. The judge outputs a structured verdict (`correct` or `incorrect`) under a `[JUDGEMENT]` tag, which is then parsed for scoring.

4.2.2 MedIF-Bench

We observe that medical domain fine-tuning often hurts a model's instruction-following ability. In our own training experiments, as well as in evaluations of several open-source medical MLLMs, we find that general-purpose MLLM models fine-tuned on medical SFT data frequently fail to comply with output format requirements—even when the underlying medical knowledge is correct. This is problematic because real clinical applications often require outputs in specific structured formats (*e.g.*, JSON for integration with electronic health records, reports with standardized section headers, or reasoning traces with confidence levels). A model that gives a correct diagnosis but ignores the requested format has limited practical utility. To track this issue, we developed MedIF-Bench, a benchmark that evaluates whether medical MLLMs can follow diverse, clinically relevant formatting instructions.

[2]Unless otherwise specified, all LLM-as-a-Judge evaluations in this work use OpenAI `gpt-4.1-2025-04-14` as the judge model.

Benchmark construction. MedIF-Bench contains 900 samples in total, built around a pool of manually curated formatting instructions organized by task category. For each sample, the model receives both a medical task and a specific formatting instruction that dictates how the output must be structured. Importantly, **this benchmark evaluates format compliance only, not medical accuracy**—the goal is to measure whether the model can produce outputs that conform to the requested structure, regardless of whether the medical content is correct. Each response is checked against a predefined regular expression pattern: a response scores 1.0 if and only if it fully matches the required format, and 0.0 otherwise.

The instruction pool covers seven task categories, each with its own pool of manually curated instructions and corresponding regex patterns. The first group of categories—**MCQ**, **MCQ with context**, and **Open**—are primarily constructed from the benchmark datasets listed in Extended Data Table 2. For these categories, we curate prompts that cover a range of common formatting requirements with special markers and structures, where the model must present its selected option or answer in a prescribed format. The second group of categories—**Prognosis**, **Report generation**, **SEER**, and **Treatment**—focus on extracting structured clinical variables from imaging or patient data for specific clinical tasks such as survival prognosis, report generation, and treatment recommendation. For these categories, some samples are drawn from the benchmark datasets in Extended Data Table 2, while others are constructed using real patient records from the SEER cancer registry [65], grounding the evaluation in realistic oncological data. Three examples from MedIF-Bench are shown in Extended Data Fig. 5b.

4.2.3 RoI-Grounded Report Generation

Existing evaluation pipelines for medical report generation suffer from two primary flaws. First, they typically prompt the model to generate a full report based solely on the image, entirely devoid of any clinical background information. Second, existing evaluation metrics, whether surface-level lexical or semantic matching (such as BLEU [49], ROUGE [30], BERTScore [33]), entity-level extraction that relies on small-scale models with limited accuracy (such as RadGraph-F1 [50]), or holistic LLM-based scoring that conflates accuracy and completeness (such as GREEN [36]), fail to provide a fine-grained, factuality-driven decomposition of clinical findings.

To address these limitations, we propose a novel evaluation paradigm: RoI-Grounded Report Generation, which is built upon two key innovations: **1)** context-aware report generation and **2)** factualness-driven LLM-as-a-Judge evaluation. The complete workflow (Extended Data Fig. 6) proceeds in three stages. First, we extract a clinical context from the GT report, consisting of potential background information and regions of interest. Next, we use this context to guide the model in generating a focused, RoI-grounded medical report. Finally, an LLM judge compares the claims in the generated report against the GT, categorizing the model claims to calculate exact clinical precision, recall, and F1 scores.

Context-aware report generation. Instead of naively asking the model to describe an entire image from scratch, we first extract a high-level clinical context for more clinically aligned report generation. Specifically, the clinical indication and the anatomical areas of focus—from the GT report using LLMs. This extracted context is then injected into the prompt for the model-under-test. A strict security constraint is applied during the extraction phase to ensure absolutely no specific findings or abnormalities are leaked. This design choice is driven by two fundamental justifications: **1) Mimicking real-world workflows:** Real-world radiologists do not read scans blindly; they consistently review patient background information and clinical indications before writing a report. Our pipeline faithfully simulates this standard clinical practice. **2) Ensuring verifiability:** A model's output can only be rigorously verified if its claims fall within the scope of the reference report. If a model is evaluated without context and describes a region not mentioned in the GT, it is impossible to objectively determine whether the model hallucinated or if the original human radiologist simply omitted a normal finding. By restricting the model to specific Regions of Interest (RoIs) defined by the GT, we ensure that the generated statements can be assessed against the reference report, substantially improving the verifiability of the evaluation.

Extended Data Fig. 6a presents the prompt design for context extraction alongside an actual example using the GPT 4.1 API. Extended Data Fig. 6b illustrates how this context is utilized as a condition for report generation on the same actual sample using our MLLM.

Factualness-driven LLM-as-a-Judge evaluation. To move beyond coarse semantic similarity, we leverage the

advanced text-processing capabilities of modern LLMs to evaluate exact clinical factualness. Instead of generating a single holistic score, the LLM judge is tasked with comparing the GT report against the model's generated report to categorize the diagnostic claims into three explicit lists of abnormalities: `[MATCHED_ABNORMALITY]` (True Positives, TP), meaning the GT mentions an abnormality and the model correctly identifies it; `[MISSED_ABNORMALITY]` (False Negatives, FN), meaning the GT mentions an abnormality but the model either omits it or falsely claims the area is normal; and `[HALLUCINATED_ABNORMALITY]` (False Positives, FP), meaning the model fabricates an abnormality that is either explicitly stated as normal or completely unmentioned in the GT.

Based on these exact counts, we compute rigorous clinical metrics: Precision $\frac{TP}{TP+FP}$, Recall $\frac{TP}{TP+FN}$, and the F1-score $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$, which provide transparent, granular indicators of diagnostic correctness. Extended Data Fig. 6c provides a worked example of this scoring process, showing how TP, FN, and FP counts are extracted and converted to Precision, Recall, and F1.

In the Results section (Section 2.4.1), we further validate this evaluation design by comparing against the conventional context-free + RadGraph-F1 pipeline [50, 8], demonstrating that **1)** context-free generation renders metrics unreliable by penalizing unverifiable claims as false positives, and **2)** RadGraph-F1 suffers from synonymy insensitivity and length bias that our LLM-as-a-Judge avoids.

4.3 Data Curation

Training a clinically useful medical MLLM requires data that simultaneously supports **1)** robust multimodal perception across heterogeneous medical imagery, **2)** medical reasoning and evidence-based answering, and **3)** realistic clinical workflows—especially 3D volumetric CT interpretation. To this end, we collect and construct a large-scale training corpus designed for our multi-stage training curriculum (Section 4.5), as shown in Extended Data Fig. 7. The overall corpus contains 22.2 million samples: 12.6M medical multimodal, 4.7M medical text, 3.4M general multimodal, and 1.5M general text (Extended Data Table 1).

Data sources. Our data originates from three complementary sources: **1)** open-source medical multimodal datasets spanning radiology and pathology; **2)** medical evidence corpora including clinical guidelines, journal articles, and case reports; and **3)** CT–report aligned data, combining open-source CT studies with high-quality reports and hospital-collected CT scans with synthesized reports.

Data synthesis pipelines. To address specific training bottlenecks, we develop four data synthesis pipelines. Detailed procedures and illustrations (Supplementary Fig. 1) for each pipeline are provided in Supplementary Methods.

- **Instruction Warping** (Supplementary Note S1.1). Medical adaptation tends to erode general instruction-following ability, which in turn constrains how the underlying medical knowledge can be expressed in realistic clinical settings (*e.g.*, structured outputs, scope restrictions, format compliance). To preserve this ability during fine-tuning, we manually author 120 seed instructions covering diverse clinical roles, scenarios, and formatting requirements, and use LLM-based expansion to produce 21k diverse instruction templates that are applied to existing samples via schema-based recomposition.
- **Dense Caption Generation** (Supplementary Note S1.2). Paper-figure captions provide only sparse supervision and composite figures introduce noise. We perform sub-figure segmentation and multi-stage filtering to isolate high-quality medical images, then generate dense visually-grounded descriptions using a dual-VLM consensus pipeline, yielding 610k dense-caption samples from 100k journal articles and 785k case reports.
- **Interactive Multimodal CoT Annotation** (Supplementary Note S1.3). To train long-horizon multimodal reasoning, we mine 100k reasoning-hard instances (high generation length yet low correctness) and solve them with a multi-agent pipeline [66] coupling a text-only orchestrator with a multimodal visual grounder through iterative evidence-seeking interaction, retaining 67k trajectories whose conclusions match reference answers.
- **Large-Scale CT Caption with Expert Toolset** (Supplementary Note S1.4). To enable native 3D understanding at scale, we curate 0.2M open-source CT studies with high-quality reports [67, 16, 68, 69] and develop a segmentation-assisted annotation pipeline based on lesion [70] and organ [71] segmentation models

to generate pseudo-report supervision for 1.1M hospital-collected CT scans, substantially expanding CT–report aligned data. The CT corpus supports both Stage ④ (rapid 3D encoder alignment) and Stage ⑤ (holistic 2D/3D instruction tuning).

4.4 Agentic Tool Use

Beyond the core architectural and evaluation contributions, we develop an agentic tool use extension to enhance ClinFusion's practical clinical deployment. Although MLLMs excel at language-based reasoning and flexible instruction following, their reliability in high-stakes settings can be further improved by addressing **1)** time-sensitive medical knowledge that cannot be fully memorized in parameters, **2)** strict traceability requirements for clinical accountability, and **3)** perception bottlenecks in certain imaging sub-tasks where specialist vision models remain stronger than generalist MLLMs.

To address these deployment considerations, we equip ClinFusion with a tool ecosystem and an explicit *plan–act* workflow, illustrated in Extended Data Fig. 4c. Given a multimodal clinical input (a user query together with one or more medical images), ClinFusion first enters a planning stage, in which it analyzes the query, optionally performs multi-query rewriting for text inputs [72], identifies the evidence required to answer it, and produces an ordered tool-invocation plan. In the subsequent stage of tool use and execution, ClinFusion sequentially calls the planned tools—which fall into two complementary categories, *knowledge tools* for evidence-grounded retrieval and *perception expert tools* for high-precision medical image analysis (described below)—and collects their structured outputs as auditable intermediate evidence. Finally, the model synthesizes the query, the original visual input, and the aggregated tool outputs into a single clinical response, ensuring that the final answer is supported by traceable evidence rather than implicit parametric recall alone. This design follows a principle of *capability compositionality*: ClinFusion serves as the central reasoning and orchestration engine, while tools provide verifiable evidence and specialized perception signals that can be audited and integrated into final outputs.

4.4.1 Knowledge Tools for Evidence-Grounded Retrieval

In clinical settings, answers must be supported by up-to-date, authoritative evidence and remain traceable. Pure parametric recall is inadequate because medical knowledge evolves quickly and unverifiable statements are unacceptable. We therefore equip ClinFusion with a RAG knowledge tool built on a hybrid retrieval engine that combines vector-based passage retrieval with knowledge-graph (KG)–based concept recall. To handle the pervasive ambiguity in medical text (synonyms, abbreviations, brand–generic drug names), the retrieval engine performs concept normalization using UMLS [73] before querying, mapping medical entities in the query to standardized concepts. The vector channel then retrieves semantically matched passages from large-scale text corpora, while the KG channel leverages UMLS and PrimeKG [74] to expand normalized concepts via clinically meaningful relations (*e.g.*, drug–disease, symptom–disease, contraindication), reducing misses caused by surface-form mismatch. Retrieved passages from both channels are then aggregated by clustering content that states the same clinical claim, and ClinFusion summarizes each cluster with attached citations to produce a compact, evidence-grounded context.

We organize our evidence resources into three complementary blocks: **1)** large-scale unstructured text corpora for broad coverage, including PubMed [75], Wikipedia, StatPearls [76], and medical textbooks [77], following MedRAG [78]; **2)** open-source medical knowledge graphs (UMLS and PrimeKG, described above); and **3)** curated in-house repositories comprising a guideline knowledge base of 18K clinical guidelines and expert consensuses released since 2006 by official authorities in China and the United States, and a journal knowledge base of 100K research articles from leading medical journals published since 2015. Detailed retrieval and aggregation procedures are provided in Supplementary Methods.

4.4.2 Perception Tools for Specialized Visual Analysis

MLLMs offer flexible multimodal reasoning, but can be unreliable on perception-critical imaging tasks that demand voxel-level precision and stable quantification (*e.g.*, exact lesion boundaries, small-structure detection, and subtle CT signs). In clinical settings, these perception outputs often define the downstream conclusion—measurements, staging, response assessment, and follow-up comparability. We therefore equip ClinFusion

with perception expert tools and treat them as external senses: when a question hinges on fine-grained visual evidence, ClinFusion actively delegates perception to specialized models and then reasons over their structured results. We currently integrate three CT-focused experts and two chest X-ray-focused experts:

- **Multi-organ lesion segmentation expert** [70] segments 29 common lesion types across eight organs (breast, colon, esophagus, stomach, kidney, lung, liver, pancreas) and produces structured lesion inventories with counts, locations, and sizes.
- **Fatty liver assessment expert** [79] estimates fatty liver severity from CT and localizes the liver parenchyma region used for assessment, producing a severity grade with the supporting parenchyma region.
- **Abdominal lymph node segmentation expert** [80] segments abdominal lymph nodes and outputs node enumeration with size measurements, supporting the detection of lymphadenopathy.
- **Chest X-ray finding classification expert** [81, 82] predicts structured probabilities for thoracic findings from chest radiographs, integrating TorchXRayVision (18 common findings) and CheXFound (40 findings).
- **Chest X-ray finding generation expert** [83] generates clinically coherent chest X-ray finding descriptions, used for sentence-level imaging evidence when classification scores alone are insufficient.

Each tool produces structured outputs that are normalized into compact evidence objects and verbalized into natural-language evidence statements before being fed back to ClinFusion. Full tool descriptions and routing details are provided in Supplementary Methods.

4.5 Progressive Staged Training

Training a complex, multi-component architecture like ClinFusion from scratch is challenging and computationally prohibitive. To ensure stability and effective knowledge transfer from pre-trained components, we devise a progressive training strategy. This curriculum-like approach systematically builds the model's capabilities by gradually increasing task complexity and data modality, moving from 2D alignment to holistic 3D instruction following. The entire process is summarized in Table 3 and detailed below.

- **Stage 1: Shallow multi-encoder alignment (2D)**. The initial stage focuses on rapidly aligning the feature spaces of the newly introduced specialist 2D encoders with the foundational Qwen ViT. Using a large corpus of 2D medical image-text pairs, we exclusively train the projection layers, the CaSL Fusion modules and all the encoders. The LLM and all base vision encoders remain frozen, allowing for an efficient and targeted alignment of the perception components, specifically in the medical domain.
- **Stage 2: Deep vision-language alignment (2D)**. Once the visual features are harmonized, this stage aims to establish a deep semantic alignment between the fused 2D visual representation and the entire LLM. We unfreeze the LLM and continue training on the same 2D image-text alignment data. This step fine-tunes the full model to understand and describe the rich, multi-encoder visual information.
- **Stage 3: 2D instruction tuning**. With the model now capable of grounded 2D perception, we teach it to follow complex instructions and perform diverse medical tasks. We fine-tune the entire model on a high-quality, multi-task 2D visual instruction dataset. This stage imbues the model with advanced reasoning and task-execution capabilities for 2D imagery.
- **Stage 4: Rapid 3D encoder alignment**. This stage introduces the 3D modality. Leveraging our 2D-anchored strategy, the goal is to efficiently align the new 3D encoder with the rest of the model and to retain the task-execution capabilities for 2D imagery. Using 3D volume-text alignment data, we freeze the LLM and all 2D perception components (encoders and fusion modules) and train only the 3D encoder and the 2D-anchored depth-aware fusion modules. This isolates the training to the new 3D components, enabling rapid and stable convergence.
- **Stage 5: Holistic 2D/3D instruction tuning**. In the final stage, we unlock the model's full potential by fine-tuning all components on a comprehensive visual instruction tuning dataset combining 2D and 3D medical imagery. This step unifies the model's capabilities, enabling it to handle complex instructions that require a deep understanding of both 2D and 3D medical data.

This staged training ensures that each new capability is built upon a stable foundation, culminating in a robust and powerful multimodal medical model.

4.6 Implementation Details

Architectural details. The final ClinFusion builds on Qwen3-VL, whose native ViT processes images at dynamic resolution (patch size 16 with 2×2 token merging). We augment it with two specialist 2D encoders (DINOv2-Large and ConvNeXt-Large-d-320) and a native 3D encoder (PE-3D). Each encoder operates at the input resolution of its pre-trained checkpoint so as to preserve its original representational quality; these resolutions therefore differ across encoders: 224×224 (center crop) for DINOv2-Large and 320×320 for ConvNeXt-Large-d-320, while PE-3D processes volumes at $32 \times 256 \times 256$ (depth×height×width) with a patch size of 8^3. MedSigLIP, assessed only as an ablation candidate at 448×448, is not part of the final configuration. For 3D inputs, we sample $S = 4$ anchor slices, and the resulting 3D feature depth is $D_{\text{vol}} = 4$ (a 32-frame input downsampled by a factor of 8 along the depth dimension). During training, the stochastic residual regularization uses a drop probability of $p_{\text{drop}} = 0.1$, which is disabled at inference by setting $\lambda = 1$. The local cross-attention window size k in CaSL Fusion (equation (3)) is set per encoder rather than shared: $k = 3$ for DINOv2 and the native 3D encoder, and $k = 5$ for ConvNeXt. The larger window for ConvNeXt is tied to a dedicated resizing scheme: the ConvNeXt feature map is directly upsampled to $5H \times 5W$ to retain finer spatial detail (a larger upsampling factor was not adopted due to memory constraints during training). Consequently, each Qwen ViT token corresponds to a 5×5 high-frequency sub-region of the ConvNeXt feature map, which preserves the high-frequency textural detail that ConvNeXt is specialized to capture.

Training details. All stages share the same optimization setup. We use the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) with weight decay 0 and gradient clipping at a maximum norm of 1.0. The learning rate is set to 1×10^{-5} for all trainable components (the LLM, the projection layers, and the vision encoders), following a cosine schedule with 100 warmup steps, and all training is performed in BF16 precision. Training is implemented with DeepSpeed ZeRO-3, FlashAttention-2, and gradient checkpointing, using a sequence cutoff length of 4096, and is conducted on NVIDIA H20 GPUs. The per-stage configurations differ mainly in the number of GPUs, effective batch size, and number of epochs. Stages 1–3 use an effective batch size of 128 on 16 GPUs, trained for 1, 1, and 3 epochs, respectively. Stage 4 uses an effective batch size of 256 on 16 GPUs for 1 epoch. Stage 5 uses an effective batch size of 256 on 32 GPUs for 3 epochs.

References

[1] 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.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

[4] 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.

[5] 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*, pages 26296–26306, 2024.

[6] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

[7] Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

[8] 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.

[9] 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.

[10] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer, 2025.

[11] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

[12] 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*, 2024.

[13] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024.

[14] 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 (CVPR)*, pages 22170–22183, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Hu_OmniMedVQA_A_New_Large-Scale_Comprehensive_Evaluation_Benchmark_for_Medical_LVLM_CVPR_2024_paper.pdf.

[15] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654, 2021. URL https://ieeexplore.ieee.org/document/9434010.

[16] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pages 1–19, 2026.

[17] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[18] Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *International Conference on Machine Learning*, pages 80961–80990. PMLR, 2025.

[19] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.

[20] Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, et al. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. *Nature Communications*, 16(1):2258, 2025. doi: 10.1038/s41467-025-57426-0.

[21] Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[22] 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.

[23] Suraj Pai, Dennis Bontempi, Ibrahim Hadzic, Vasco Prudente, Mateo Sokač, Tafadzwa L. Chaunzwa, Simon Bernatz, Ahmed Hosny, Raymond H. Mak, Nicolai J. Birkbak, and Hugo J. W. L. Aerts. Foundation model for cancer imaging biomarkers. *Nature Machine Intelligence*, 6(3):354–367, 2024. doi: 10.1038/s42256-024-00807-9.

[24] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *International Conference on Learning Representations*, volume 2025, pages 92628–92646, 2025.

[25] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle-2: Faster inference of language models with dynamic draft trees. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7421–7432, 2024.

[26] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[27] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[28] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024.

[29] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. In *International Conference on Learning Representations*, volume 2025, pages 84491–84506, 2025.

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[31] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[32] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[33] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[34] Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.319.

[35] Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, 2024.

[36] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 374–390, 2024.

[37] Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J. Topol. Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420, 2024.

[38] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of llm-based agents in medicine: How far are we from baymax? *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10345–10359, 2025.

[39] Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025.

[40] Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760, 2024.

[41] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[42] Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, et al. Bimedix2: Bio-medical expert lmm for diverse medical modalities. In *Findings of the association for computational linguistics: EMNLP 2025*, pages 14051–14071, 2025.

[43] Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Song Xiaohui, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *International Conference on Machine Learning*, pages 37975–37995. PMLR, 2025.

[44] Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Luyang Luo, Huiling Xiang, et al. Gsco: Towards generalizable ai in medicine via generalist-specialist collaboration. *arXiv preprint arXiv:2404.15127*, 2024.

[45] Ryutaro Tanno, David G. T. Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, et al. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608, 2025. doi: 10.1038/s41591-024-03302-1.

[46] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *ArXiv*, abs/2405.19538, 2024. URL https://arxiv.org/abs/2405.19538.

[47] Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *arXiv preprint arXiv:2506.11147*, 2025.

[48] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

[49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[50] Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 12902–12915, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.765.

[51] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[52] Yilun Du and Leslie Pack Kaelbling. Position: Compositional generative modeling: A single model is not all you need. In *International Conference on Machine Learning*, pages 11721–11732. PMLR, 2024.

[53] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[54] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

[55] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, 2025.

[56] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm Ai*, 2(1):AIoa2400640, 2025.

[57] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19792–19802, 2025.

[58] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *International Conference on Machine Learning*, pages 74840–74857. PMLR, 2025.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[60] Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

[61] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

[62] Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40, 2022. doi: 10.1038/s42256-021-00425-9.

[63] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 10859–10885, 2023.

[64] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12: 933–949, 2024.

[65] National Cancer Institute, Surveillance Research Program. Surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov), 2025. URL https://seer.cancer.gov/data/. SEER database; data analyzed using SEER*Stat. Visit https://seer.cancer.gov/data/.

[66] Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9262–9281, 2025.

[67] Fan Bai, Yuxin Du, Tiejun Huang, Max Q.-H. Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

[68] Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, pages 1–11, 2026.

[69] Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Curtis Langlotz, Matthew Lungren, Serena Yeung, Nigam Shah, and Jason Fries. Inspect: A multimodal dataset for patient outcome prediction of pulmonary embolisms. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, pages 17742–17772, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/39736af1b9d87a1fddad9f84a6bcf64c-Abstract-Datasets_and_Benchmarks.html.

[70] Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, et al. Vista3d: A unified segmentation foundation model for 3d medical imaging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20863–20873, 2025.

[71] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

[72] Zhicong Li, Jiahao Wang, Zhishu Jiang, Hangyu Mao, Zhongxia Chen, Jiazhen Du, Yuanxing Zhang, Fuzheng Zhang, Di Zhang, and Yong Liu. Dmqr-rag: Diverse multi-query rewriting for rag. *arXiv preprint arXiv:2411.13154*, 2024.

[73] National Library of Medicine (US). UMLS knowledge sources, release 2024aa, 2024. http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html (accessed 15 July 2024).

[74] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023. doi: 10.1038/s41597-023-01960-3.

[75] Eric W. Sayers, Evan E. Bolton, J. Rodney Brister, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 52(D1):D33–D43, 2024. doi: 10.1093/nar/gkad1044.

[76] StatPearls Publishing. *StatPearls [Internet]*. StatPearls Publishing, Treasure Island, FL, 2026. Available from: https://www.ncbi.nlm.nih.gov/books/NBK430685/ (accessed 27 February 2026).

[77] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11 (14):6421, 2021. doi: 10.3390/app11146421.

[78] G. Xiong, Q. Jin, Z. Lu, et al. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, 2024.

[79] Yuan Gao, Chunli Li, Wanxing Chang, Bai Du, Xianghua Ye, Yee Hui Yeo, Yingda Xia, Heng Guo, Xiaoming Zhang, Wei Liu, et al. Multi-modal ai for opportunistic screening, staging and progression risk stratification of steatotic liver disease. *Nature Communications*, 17(1):1562, 2026.

[80] Qinji Yu, Yirui Wang, Ke Yan, Haoshen Li, Dazhou Guo, Li Zhang, Na Shen, Qifeng Wang, Xiaowei Ding, Le Lu, et al. Effective lymph nodes detection in ct scans using location debiased query selection and contrastive query representation in transformer. In *European Conference on Computer Vision*, pages 180–198. Springer, 2024.

[81] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning (MIDL)*, volume 121 of *Proceedings of Machine Learning Research*, pages 136–155. PMLR, 2020. URL https://proceedings.mlr.press/v121/cohen20a.html.

[82] Zefan Yang, Xuanang Xu, Jiajin Zhang, Ge Wang, Mannudeep K. Kalra, and Pingkun Yan. Chest x-ray foundation model with global and local representations integration. *IEEE Transactions on Medical Imaging*, 44 (12):4787–4799, 2025. doi: 10.1109/TMI.2025.3581907.

[83] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024. URL https://arxiv.org/abs/2401.12208.

[84] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

[85] Obioma Pelka, Svenja Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS) 2018, held in conjunction with MICCAI 2018*, volume 11043 of *Lecture Notes in Computer Science*, pages 180–189. Springer, 2018. doi: 10.1007/978-3-030-01364-6_20.

[86] 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.

[87] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 7346–7370. Association for Computational Linguistics, November 2024. doi: 10.18653/v1/2024.emnlp-main.418.

[88] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[89] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. doi: 10.1093/jamia/ocv080.

[90] Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine Van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medicat: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120, 2020.

[91] Irene Siragusa, Salvatore Contino, Massimo La Ciura, Rosario Alicata, and Roberto Pirrone. Medpix 2.0: a comprehensive multimodal biomedical data set for advanced ai applications with retrieval augmented generation and knowledge graphs. *Data Science and Engineering*, pages 1–17, 2025.

[92] Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J. Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Austin Wolfgang Katzer, Collin Chiu, Anita Rau, Xiaohan Wang, Yuhui Zhang, Alfred Seunghoon Song, Robert Tibshirani, and Serena Yeung-Levy. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19724–19735, 2025. doi: 10.1109/CVPR52734.2025.01837.

[93] 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 (CVPR)*, pages 26296–26306, June 2024.

[94] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025.

[95] 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.

[96] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192, 2024.

[97] Sushant Gautam, Andrea Storås, Cise Midoglu, Steven A. Hicks, Vajira Thambawita, Pål Halvorsen, and Michael A. Riegler. Kvasir-vqa: A text-image pair gi tract dataset. In *Proceedings of the 2nd International Workshop on Vision-Language Models for Biomedical Applications (VLM4Bio '24)*, pages 15–24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 979-8-4007-0435-2. doi: 10.1145/3689096.3689458.

[98] Seongsu Bae, Daeun Kyung, Jaehee Ryu, et al. MIMIC-Ext-MIMIC-CXR-VQA: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images, 2024. PhysioNet, https://physionet.org/content/mimic-ext-mimic-cxr-vqa/.

[99] 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 (Volume 2: Short Papers)*, pages 708–718. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-short.90. URL https://aclanthology.org/2021.acl-short.90/.

[100] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):23, 2024. doi: 10.1038/s43856-024-00709-2.

[101] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. In *CLEF 2019 Working Notes, CEUR Workshop Proceedings*, volume 2380, 2019. URL https://ceur-ws.org/Vol-2380/paper_272.pdf.

[102] 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. URL https://doi.org/10.1038/sdata.2018.251.

[103] 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.

[104] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024.

[105] Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511, 2019. doi: 10.1186/s12859-019-3119-4.

[106] 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*, 2024.

[107] Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, et al. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*, 2024.

[108] Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*, 2023.

[109] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus*, 15(6): e40895, 2023. doi: 10.7759/cureus.40895.

[110] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024.

[111] Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, et al. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*, 2023.

[112] 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.

[113] Hao Wei. Medthoughts-8k: A medical reasoning dataset distilled from deepseek-r1. https://huggingface.co/datasets/hw-hwei/MedThoughts-8K, 2025.

[114] Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. CMB: A comprehensive medical benchmark in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.343. URL https://aclanthology.org/2024.naacl-long.343/.

[115] Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071, 2018. doi: 10.1109/ACCESS.2018.2883637.

[116] Guoxin Wang, Minyu Gao, Shuai Yang, et al. Citrus: Leveraging expert cognitive pathways in a medical language model for advanced medical decision support. *arXiv preprint arXiv:2502.18274*, 2025.

[117] Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, et al. Small language models learn enhanced reasoning skills from medical textbooks. *npj Digital Medicine*, 8(1):240, 2025. doi: 10.1038/s41746-025-01653-8.

[118] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. Hugging Face Datasets, https://huggingface.co/datasets/teknium/OpenHermes-2.5.

[119] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*, 2024.

[120] AMOS-MM: Abdominal multimodal analysis challenge dataset, 2024. Zenodo, https://zenodo.org/doi/10.5281/zenodo.10992154 (accessed 23 February 2026).

[121] Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[122] Suhao Yu, Haojin Wang, Juncheng Wu, Cihang Xie, and Yuyin Zhou. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*, 2025.

[123] 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.

[124] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

[125] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.naacl-long.182/.

[126] Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shuyue Guo, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, Dehua Ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tianshun Xing, Ming Xu, Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. In *Advances in Neural Information Processing Systems 38 (NeurIPS), Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=6WgflzYQpf.

[127] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. URL https://doi.org/10.1093/jamia/ocv080.

[128] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128(2):336–359, 2020.

Acknowledgements

We thank all the board-certified radiologists who participated in the blinded expert evaluation study.

Author Contributions

H.Y., Y.Q., Z.T., X.X., P.W. and W.C. conceived and designed the study. J.W. (Jitao Wang), J.D., W.C., F.X., F.W., Y.Y. and J.T. supervised the project. H.Y., Y.Q., Z.T., X.X., Y.X., P.W. and Y.H. designed the algorithm. H.Y., X.X., Z.T., Q.X., W.Y., P.W. and Y.X. contributed to data interpretation and visualization. H.Y., Y.Q., Z.T., X.X., J.W. (Jinwang Wang), S.Y., Z.L. and S.W. contributed to the initial drafting and revision of the manuscript. X.X., H.Y., Z.T., L.W., Z.Z. and S.L. conducted the data collection. H.Y., Y.Q., S.Y., J.W. (Jinwang Wang) and T.F. participated in data analysis. All authors discussed the results and approved the paper.

Competing Interests

H.Y., Y.Q., Z.T., X.X., L.W., J.W. (Jinwang Wang), P.W., Y.H., Y.X., S.L., J.T., W.C., and F.W. are employees of DAMO Academy, Alibaba Group. The remaining authors declare no competing interests.

Data Availability

The public datasets used for model training and evaluation are listed in Extended Data Table 1 and Extended Data Table 2. Their download links and access terms are documented in Extended Data Table 4, and all datasets are available from their respective original sources. Data from the SEER cancer registry, used to construct part of the MedIF-Bench evaluation, are available at https://seer.cancer.gov/data/. The knowledge tool draws on publicly available medical knowledge sources, including UMLS (https://www.nlm.nih.gov/research/umls/), PrimeKG (https://github.com/mims-harvard/PrimeKG), PubMed (https://pubmed.ncbi.nlm.nih.gov/), StatPearls (https://www.statpearls.com/), Wikipedia (https://www.wikipedia.org/), and standard medical textbooks. The in-house knowledge base of clinical guidelines and journal articles, together with the hospital-collected CT data used in this study, are not publicly available because of patient privacy, ethical restrictions, and institutional data-use agreements. De-identified research data may be made available for academic and non-commercial purposes upon reasonable request to the corresponding authors, subject to institutional approval and a data-use agreement.

Code Availability

The source code for ClinFusion, including all evaluation scripts, benchmark implementations, and the proposed MedIF-Bench, is available at https://github.com/alibaba-damo-academy/ClinFusion. Model weights for both ClinFusion-8B and ClinFusion-32B are available at https://huggingface.co/collections/Alibaba-DAMO-Academy/clinfusion.

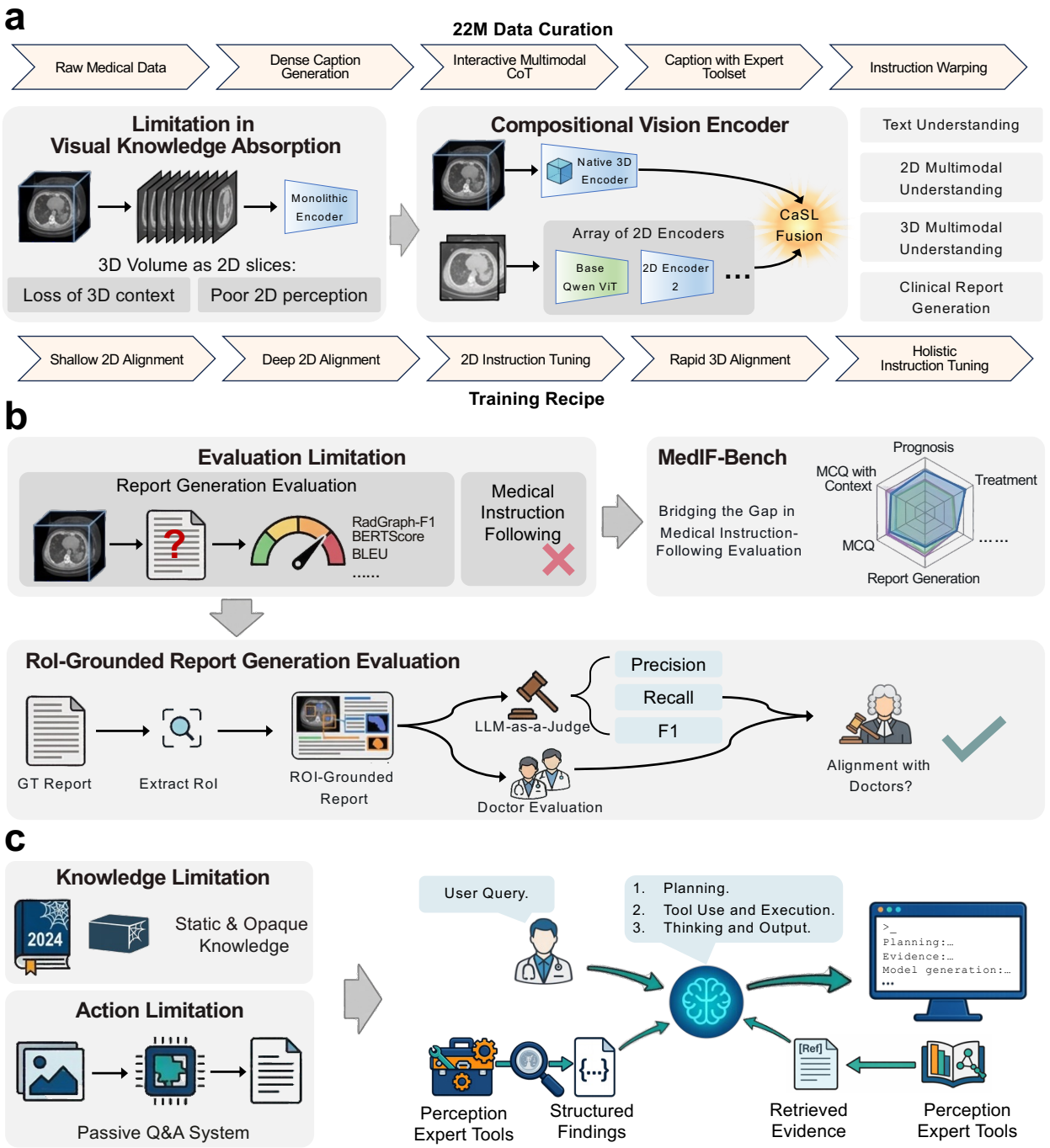


Figure 1 Overview of the ClinFusion framework. **a**, Compositional vision encoder. We address the limitation of monolithic encoders by proposing a compositional architecture with a native 3D encoder and an array of 2D encoders unified via CaSL Fusion, supported by a 22M-sample data curation pipeline and a progressive 5-stage training recipe. **b**, Vision-grounded evaluation. We identify key evaluation limitations—lack of instruction-following assessment and reliance on global-level text-matching metrics—and introduce MedIF-Bench for medical instruction following and an RoI-Grounded Report Generation Evaluation scheme that assesses diagnostic accuracy at the region level using an LLM-as-a-Judge. **c**, Agentic tool use extension. To enhance deployment, we equip ClinFusion with perception expert tools and a retrieval-augmented generation pipeline within a plan–act workflow, demonstrating consistent improvements in both text-only and multimodal clinical scenarios.

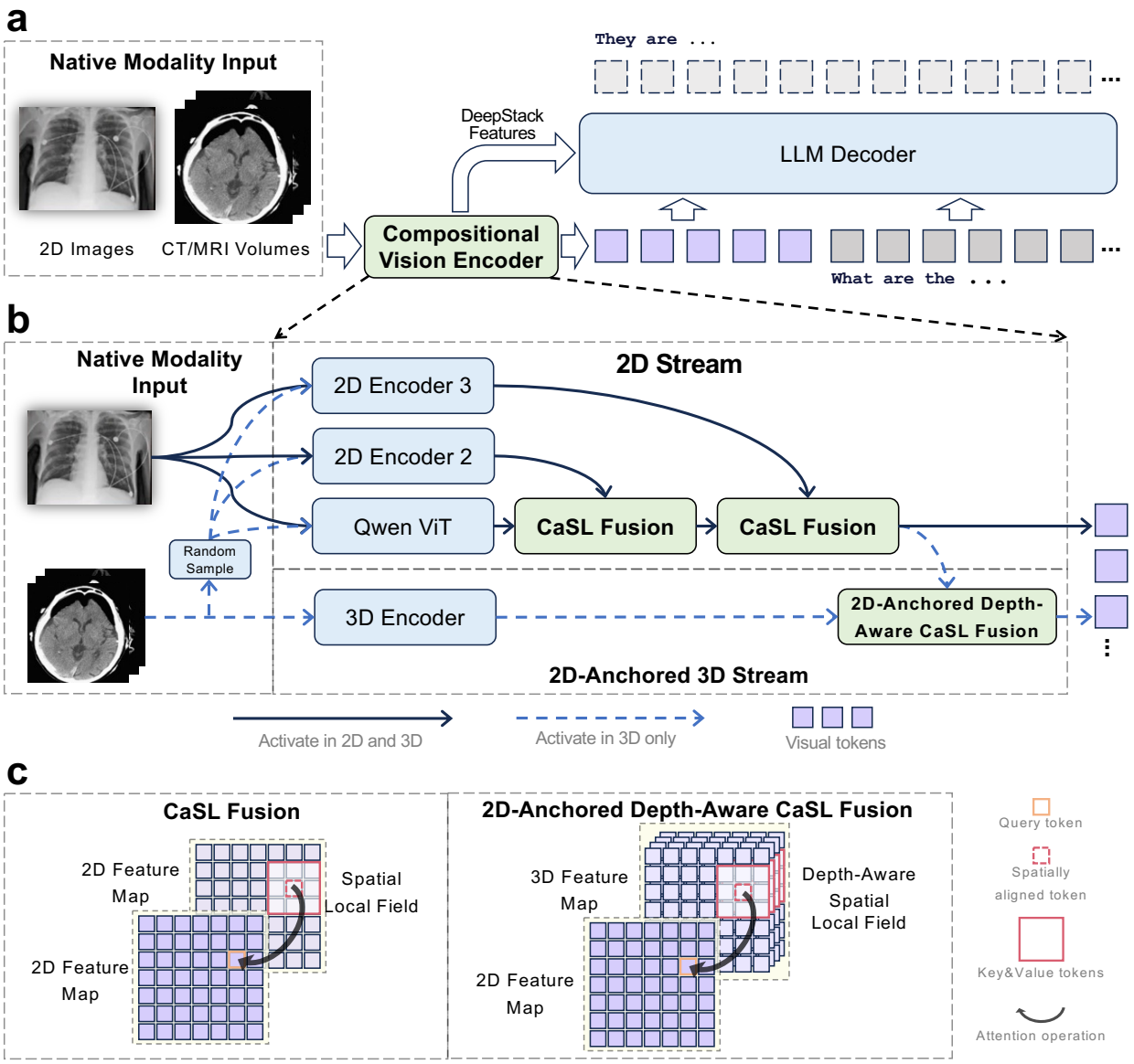
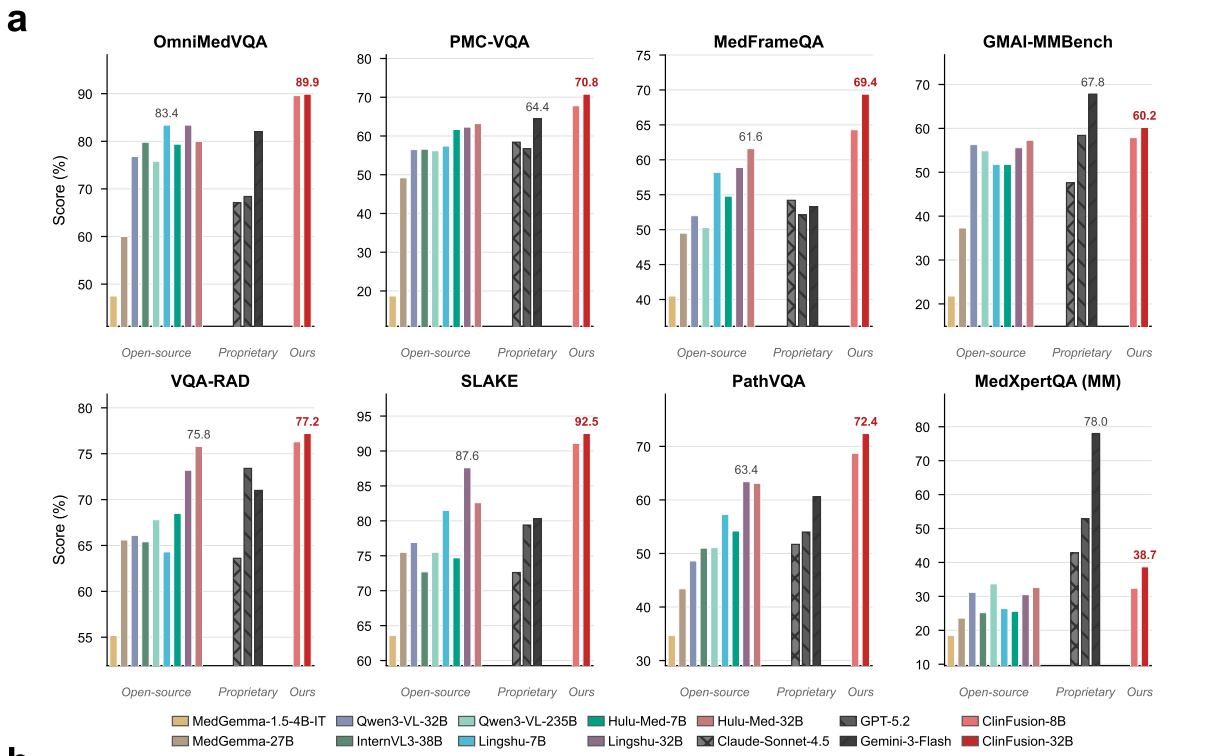


Figure 2 The architecture details of ClinFusion, a multimodal large language model for holistic medical understanding. **a**, The overall framework of ClinFusion. The model accepts native 2D and 3D medical images as input. A Compositional Vision Encoder module processes these inputs to generate a sequence of information-dense visual tokens. These tokens are then fed into an LLM decoder to enable complex reasoning and generation. The architecture is compatible with deep fusion mechanisms like DeepStack, where visual features are injected at multiple layers of the LLM. **b**, A detailed view of the Compositional Vision Encoder module, which comprises two primary processing streams. The 2D Stream handles 2D images using a compositional ensemble of encoders: a foundational Qwen ViT and several specialist encoders (*e.g.*, Encoder 2, Encoder 3). Features are progressively refined through a series of CaSL Fusion blocks in a cascaded manner. The 2D-Anchored 3D Stream is activated for 3D volumetric input. It processes the full 3D volume with a native 3D encoder while simultaneously processing randomly sampled 2D slices (anchors) through the 2D stream. A final CaSL Fusion block then integrates the 3D features, guided by the well-aligned 2D anchor representations. **c**, A schematic of the core CaSL Fusion operator. Left: The standard CaSL Fusion operator used in the 2D stream. A query token from one feature map attends to a small, spatially local field of key/value tokens from another feature map, enforcing a spatial inductive bias. Right: The 2D-Anchored Depth-Aware CaSL Fusion operator. A 2D query token attends to a 3D local field that spans both a spatial neighborhood and the depth dimension of the 3D feature map, enabling the 2D representation to be enriched with volumetric context.



b

	CheXpert Plus				IU-XRAY		
	Precision	Recall	F1		Precision	Recall	F1
ClinFusion-32B	**56.1**	**46.6**	**48.0**	*Ours*	60.2	54.7	55.6
ClinFusion-8B	46.4	35.3	37.8		**63.3**	**55.8**	**57.3**
Gemini-3-Flash	38.4	33.5	33.0	*Proprie-tary*	56.8	55.2	54.4
GPT-5.2	49.1	38.1	40.2		55.3	50.3	50.6
Claude-Sonnet-4.5	33.8	28.3	28.9		39.2	39.2	37.7
Hulu-Med-32B	51.0	35.5	39.6	*Open-source*	53.5	46.8	48.2
Lingshu-32B	28.0	21.5	22.8		44.5	40.5	41.3
MedGemma-27B	26.9	21.9	22.5		41.3	40.7	40.1
Hulu-Med-7B	40.5	29.1	31.9		51.1	45.3	46.5
Lingshu-7B	26.2	17.8	19.9		43.7	39.3	40.4
MedGemma-1.5-4B-IT	34.1	23.1	25.4		43.1	38.9	39.8
Qwen3-VL-235B	29.5	25.2	25.6		43.2	41.4	41.2
Qwen3-VL-32B	30.6	27.9	27.3		40.8	41.0	40.0

Figure 3 Performance of ClinFusion on 2D multimodal medical benchmarks. **a**, Results on 2D multimodal medical VQA benchmarks, grouped into multi-modality VQA (OmniMedVQA, PMC-VQA, MedFrameQA, GMAI-MMBench), specific-modality VQA (VQA-RAD, SLAKE, PathVQA), and a reasoning-oriented benchmark (MedXpertQA-MM). We compare ClinFusion against leading proprietary models, general-purpose MLLMs, and medical MLLMs at various parameter scales. Results are all re-benchmarked by us in a fair protocol. **b**, Results on 2D report generation benchmarks (CheXpert-Plus and IU-Xray), evaluated using the proposed RoI-Grounded Report Generation protocol and reported as Precision, Recall and F1.

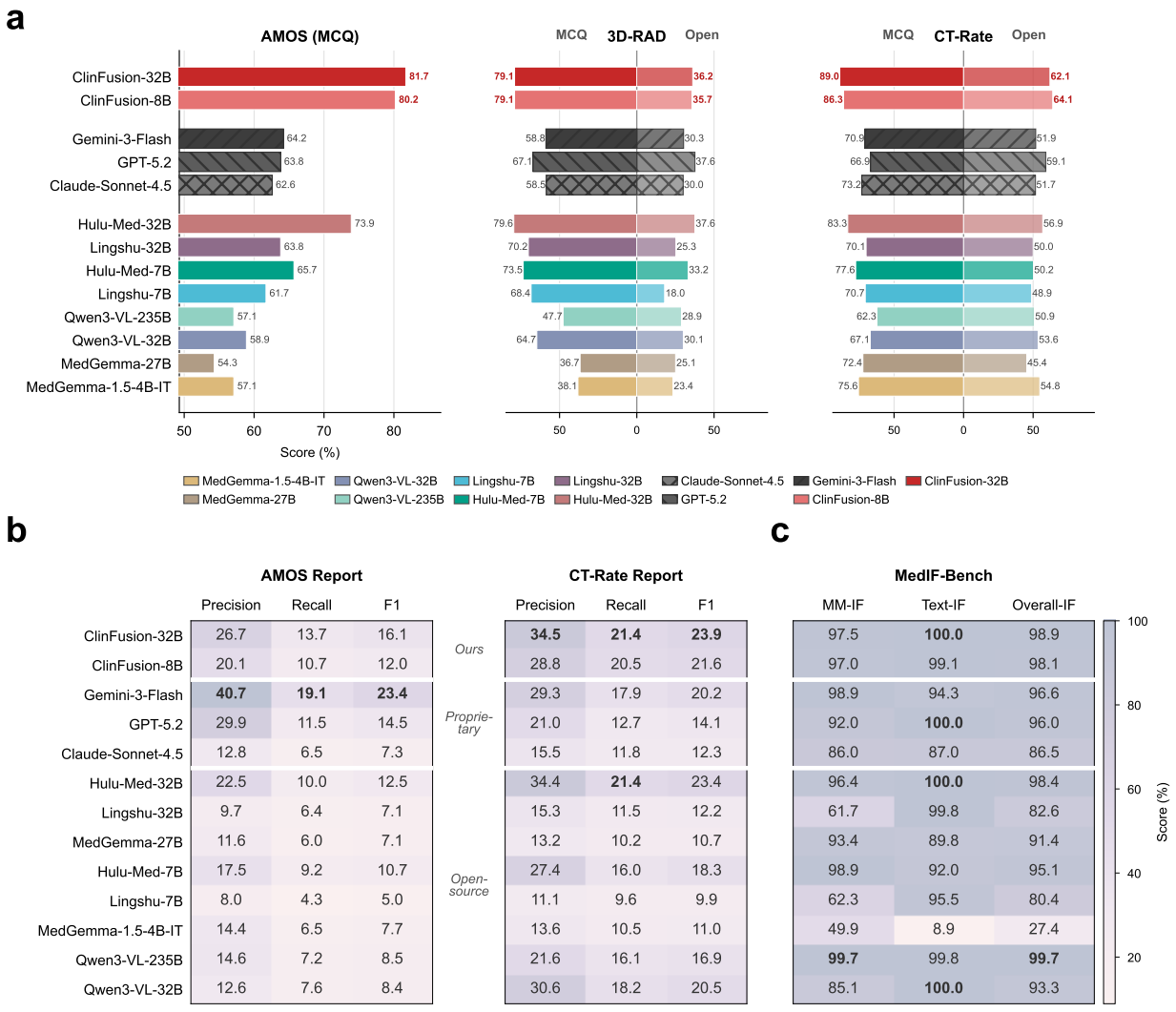


Figure 4 Performance of ClinFusion on 3D multimodal medical benchmarks and MedIF-Bench. **a**, 3D multimodal medical benchmark results on AMOS-MM (MCQ), 3D-RAD (MCQ and Open-ended), and CT-Rate-VQA (MCQ and Open-ended). We compare our model with proprietary and open-source models, re-benchmarked by us in a fair protocol. **b**, 3D report generation results on AMOS-MM Report and CT-Rate Report, evaluated using the RoI-Grounded protocol and reported as Precision, Recall and F1. **c**, Results on our proposed MedIF-Bench, reporting composite instruction-following scores for multimodal inputs (MM-IF), text-only inputs (Text-IF), and the overall average (Overall-IF).

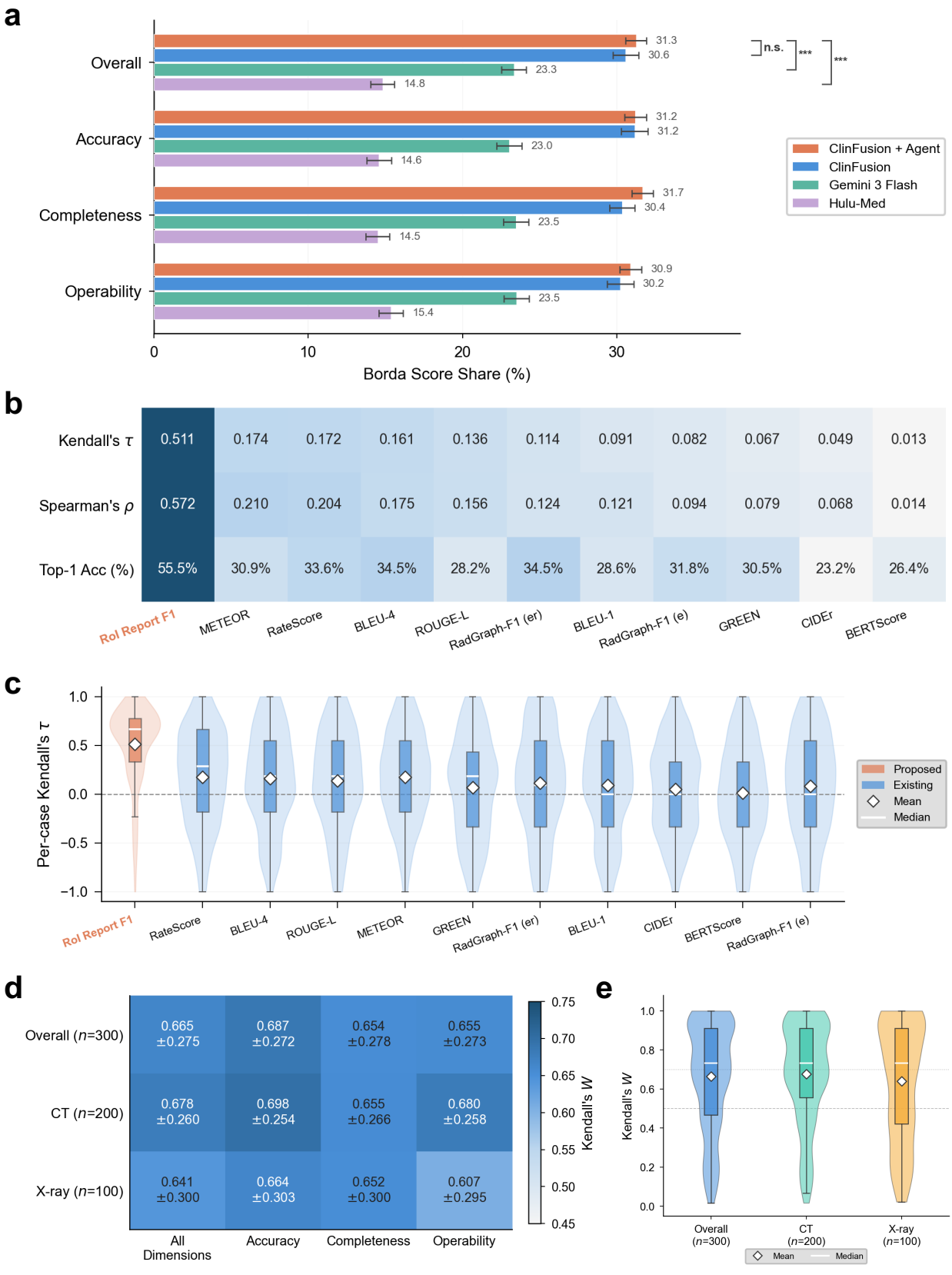
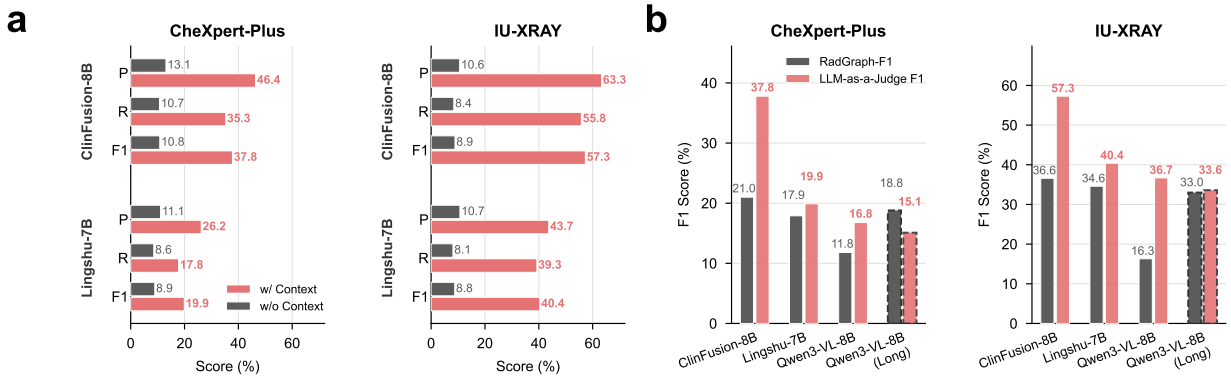

a
Overall
Accuracy
Completeness
Operability
31.3
30.6
23.3
14.8
31.2
31.2
23.0
14.6
31.7
30.4
23.5
14.5
30.9
30.2
23.5
15.4
n.s.
***
***
ClinFusion + Agent
ClinFusion
Gemini 3 Flash
Hulu-Med
0
10
20
30
Borda Score Share (%)
b
Kendall's τ
Spearman's ρ
Top-1 Acc (%)
0.511
0.174
0.172
0.161
0.136
0.114
0.091
0.082
0.067
0.049
0.013
0.572
0.210
0.204
0.175
0.156
0.124
0.121
0.094
0.079
0.068
0.014
55.5%
30.9%
33.6%
34.5%
28.2%
34.5%
28.6%
31.8%
30.5%
23.2%
26.4%
RoI Report F1
METEOR
RateScore
BLEU-4
ROUGE-L
RadGraph-F1 (er)
BLEU-1
RadGraph-F1 (e)
GREEN
CIDEr
BERTScore
c
Per-case Kendall's τ
1.0
0.5
0.0
−0.5
−1.0
Proposed
Existing
Mean
Median
RoI Report F1
RateScore
BLEU-4
ROUGE-L
METEOR
GREEN
RadGraph-F1 (er)
BLEU-1
CIDEr
BERTScore
RadGraph-F1 (e)
d
Overall (n=300)
CT (n=200)
X-ray (n=100)
0.665 ±0.275
0.687 ±0.272
0.654 ±0.278
0.655 ±0.273
0.678 ±0.260
0.698 ±0.254
0.655 ±0.266
0.680 ±0.258
0.641 ±0.300
0.664 ±0.303
0.652 ±0.300
0.607 ±0.295
All Dimensions
Accuracy
Completeness
Operability
Kendall's W
0.75
0.70
0.65
0.60
0.55
0.50
0.45
e
Kendall's W
1.0
0.8
0.6
0.4
0.2
0.0
Overall (n=300)
CT (n=200)
X-ray (n=100)
Mean
Median

Figure 5 Expert radiologist evaluation. **a**, Expert radiologist model ranking. Aggregated Borda score shares from radiologists across three original evaluation dimensions (Accuracy, Completeness, Operability) and one aggregated dimension (Overall). Scores were aggregated using the Borda count method (score = 4 - rank). Error bars denote standard error of the mean over per-case Borda shares. Brackets indicate Wilcoxon signed-rank test (two-sided, paired by case) comparing ClinFusion + Agent with each competing model on the Overall dimension (n.s.: not significant; ***: $p < 0.001$). **b**, Correlation between automated metrics and expert annotations. Kendall's τ, Spearman's ρ, and Top-1 Accuracy (the fraction of cases where the metric correctly identifies the doctor-preferred model) for each automatic metric computed against the radiologist consensus ranking. Metrics are sorted by Kendall's τ in descending order. **c**, Per-case distribution of ranking agreement between automated metrics and doctor annotations. Violin plots showing the distribution of per-case Kendall's τ between each automated metric and expert-annotated rankings. White diamonds and horizontal white lines indicate the mean and median, respectively. **d**, Inter-annotator agreement: heatmap of Kendall's coefficient of concordance (W) stratified by imaging modality (CT, X-ray) and evaluation dimension. Values represent mean $W \pm s.d.$. **e**, Inter-annotator agreement: violin plots showing the distribution of W across individual cases. Box plots indicate the median (white horizontal line), interquartile range (box) and 1.5 x IQR (whiskers). White diamonds indicate the mean. Dashed lines indicate thresholds for moderate (W = 0.5) and strong (W = 0.7) agreement.



c Fusion Mechanism

	VQA-RAD	PMC-VQA	MedXQA	SLAKE	PathVQA	OMVQA	Avg
No Fusion	69.4	57.4	24.6	81.0	58.0	83.5	62.3
(Ours) Local Cross-Attn	72.3	57.5	25.3	84.4	65.5	82.3	**64.5**
Global Cross-Attn	69.5	55.0	24.2	82.6	61.0	84.1	62.7
Mixture of Vision Experts	70.0	55.2	24.2	81.1	63.3	83.2	62.8
Channel Concat	70.5	54.4	24.7	83.4	63.2	82.3	63.1

d Fusion of One Additional Encoder

	VQA-RAD	PMC-VQA	MedXQA	SLAKE	PathVQA	OMVQA	CheXpert-Plus	IU-XRAY	Report Avg	VQA Avg
DINOv2	72.3	57.5	25.3	84.2	65.5	82.3	20.2	35.5	27.8	64.5
ConvNext	72.1	57.1	24.8	83.5	64.7	82.7	20.0	35.9	**27.9**	64.1
Medsiglip	70.7	57.8	25.9	83.5	66.9	84.2	20.9	32.3	26.6	**64.8**

e Fusion of Two Additional Encoders

		VQA-RAD	PMC-VQA	MedXQA	SLAKE	PathVQA	OMVQA	CheXpert-Plus	IU-XRAY	Report Avg	VQA Avg
Cascade	DINOv2 → ConvNext	70.1	57.1	25.0	84.7	66.2	83.4	21.0	36.6	**28.8**	64.4
Cascade	ConvNext → DINOv2	70.3	54.8	26.2	84.2	66.5	84.2	17.9	37.4	27.6	64.4
Cascade	DINOv2 → Medsiglip	72.7	58.0	25.2	84.7	66.5	84.4	20.7	34.1	27.4	**65.2**
Cascade	Medsiglip → DINOv2	69.0	53.6	25.5	84.6	68.0	85.9	20.0	32.5	26.3	64.4
Parallel	DINOv2 + ConvNext	71.4	58.0	24.8	83.9	65.3	83.9	20.0	34.6	27.3	64.6

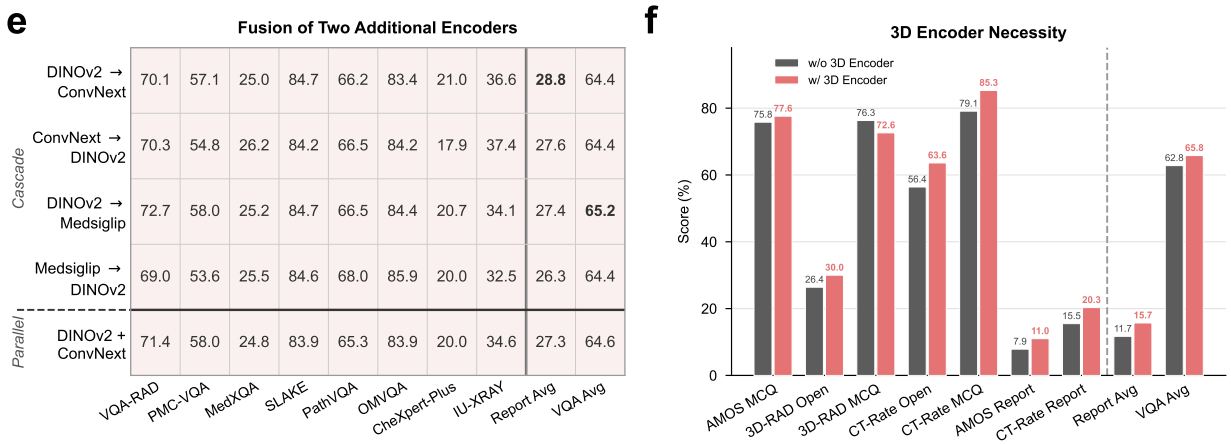


Figure 6 Ablation studies on the RoI-Grounded report generation evaluation and on the architectural design of ClinFusion. **a**, Ablation on the necessity of clinical context for report generation. ClinFusion-8B and Lingshu-7B [6] are evaluated under two settings: with the extracted clinical context ("w/ Context") and without it ("w/o Context"), where the Clinical Indication and Area of Focus are removed from the generation prompt. Precision, Recall, and F1 are reported on CheXpert-Plus and IU-XRAY. **b**, Comparison between RadGraph-F1 and our LLM-as-a-Judge metric on CheXpert-Plus and IU-XRAY across four model variants, including a Qwen3-VL-8B (Long Output) variant prompted to encourage verbose generation. RadGraph-F1 reports entity-relation F1, while the LLM-as-a-Judge reports F1 for clarity. **c**, Ablation on different fusion mechanisms for the compositional vision encoder. A single-encoder baseline (Qwen ViT) is compared against a dual-encoder setup (Qwen ViT + DINOv2) using channel concatenation, global cross-attention, mixture of vision experts, and our local cross-attention (CaSL Fusion). Accuracy (%) is reported on six 2D medical VQA benchmarks. **d**, Ablation on the choice of the one additional specialist 2D encoder. The base Qwen ViT is augmented with one additional encoder (DINOv2, ConvNext, or Medsiglip) and evaluated on 2D VQA (accuracy, %) and report generation. **e**, Ablation on the choice of multiple additional 2D encoders, the multi-encoder fusion strategy, and the fusion order. The base Qwen ViT is augmented with two additional encoders combined either through Cascade fusion (with different orderings) or Parallel fusion, and evaluated on 2D VQA and report generation. **f**, Ablation on the necessity of the native 3D encoder during inference. The full model is compared against an ablated version where the 3D encoder is deactivated, forcing the model to rely solely on sampled 2D slices. Accuracy (%) is reported on 3D VQA benchmarks (AMOS, 3D-RAD, CT-Rate) and Precision, Recall, and F1 are reported on 3D report generation (AMOS Report, CT-Rate Report).

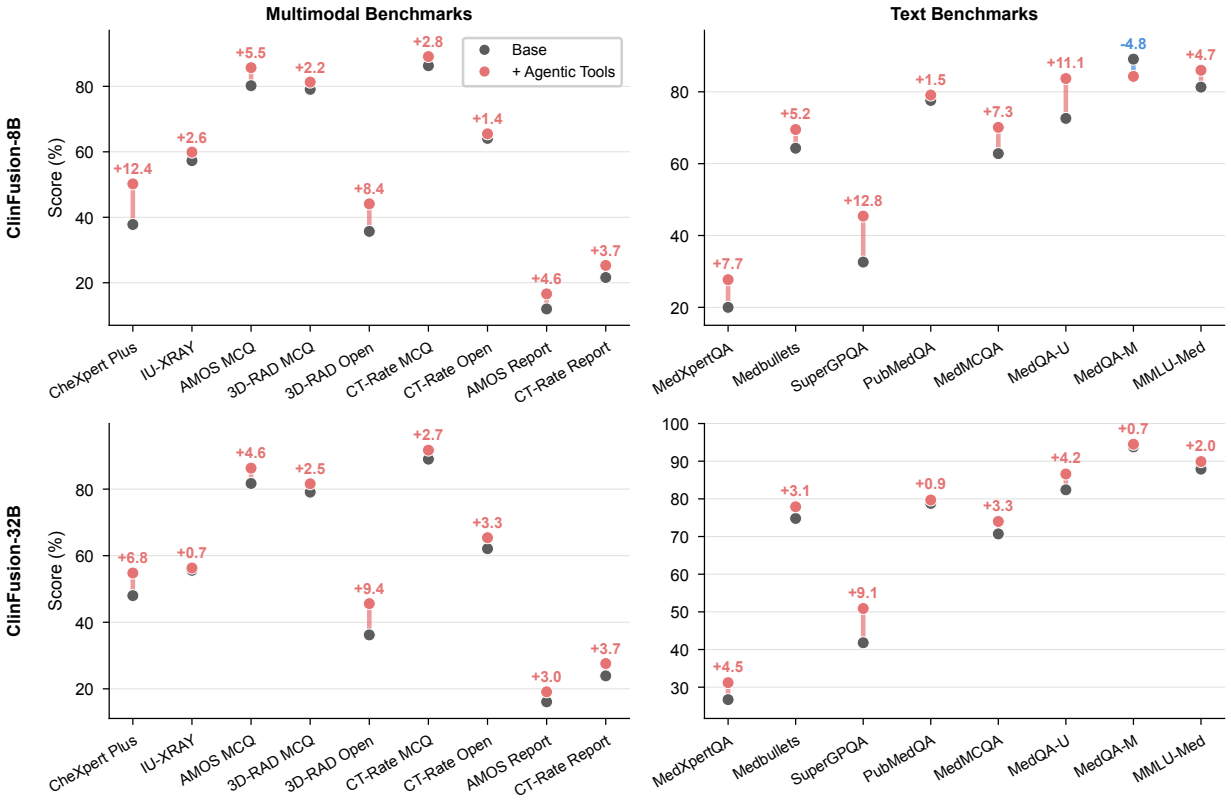
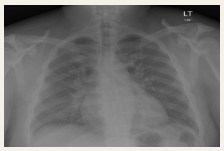


Figure 7 Performance gains of ClinFusion when augmented with agentic tool use. For each benchmark, grey and red dots denote the Base and +Agentic Tools scores respectively, connected by a vertical segment, with the absolute difference annotated above (red for positive, blue for negative). Rows correspond to ClinFusion-8B (top) and ClinFusion-32B (bottom); columns correspond to multimodal benchmarks (left) and text-only medical benchmarks (right). For VQA and exam benchmarks scores are accuracy (%); for report generation scores are F1 (%) under the RoI-Grounded protocol.

Extended Data

a

Input



You are an expert radiologist. Your task is to analyze a medical image and write a report based on the provided clinical context.

Clinical Indication: Assessment of cardiac size and pulmonary vascular status.
Based on this indication, you are asked to evaluate the following key areas.

Your report must focus on these topics.
Areas of Focus:
- Heart and great vessels
- Pulmonary vasculature
- Lungs and pleura

……

Model Prediction

GT answer:
Findings: the heart is large, and the pulmonary are engorged. no infiltrates.
Impression: cardiomegaly and pulmonary venous hypertension.

Hulu-Med:
Findings: Normal heart size and mediastinal contours. No pleural effusion or pneumothorax. Pulmonary vasculature appears normal with no evidence of congestion or edema.
Impression: Normal chest radiograph with no acute cardiopulmonary abnormalities.

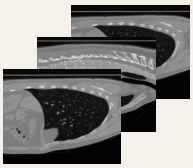
Lingshu:
Findings: The heart and great vessels are within normal limits. The pulmonary vasculature is normal. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax.
Impression: Normal heart and great vessels. Normal pulmonary vasculature. Normal lungs.

Gemini-3-Flash:
Findings: The cardiac silhouette is normal in size. The mediastinal and hilar contours are unremarkable. The pulmonary vasculature is normal in caliber and distribution, without evidence of pulmonary venous congestion. The lungs are clear bilaterally, with no focal consolidation, masses, or nodules. There are no pleural effusions or pneumothoraces.
Impression: Normal cardiac size and pulmonary vascularity. No acute cardiopulmonary abnormality.

Ours:
Findings: Heart and great vessels: Cardiac silhouette is mildly enlarged. Mediastinal contours are unremarkable. Pulmonary vasculature: There is mild prominence of the pulmonary vasculature. Lungs and pleura: No focal consolidation. No pleural effusion. No pneumothorax.
Impression: Mild cardiomegaly with mild pulmonary vascular prominence and no acute pulmonary abnormality.

b

Input



(Native CT Volume)

You are an expert radiologist. Your task is to analyze a medical image and write a report based on the provided clinical context.

Clinical Indication: Evaluation of thoracic and upper abdominal structures, likely for assessment of liver and pulmonary status. Based on this indication, you are asked to evaluate the following key areas. Your report must focus on these topics.

Areas of Focus:
- Lungs and pleura
- Mediastinum and heart
- Liver and biliary system
- Adrenal glands
- Bone structures

……

Model Prediction

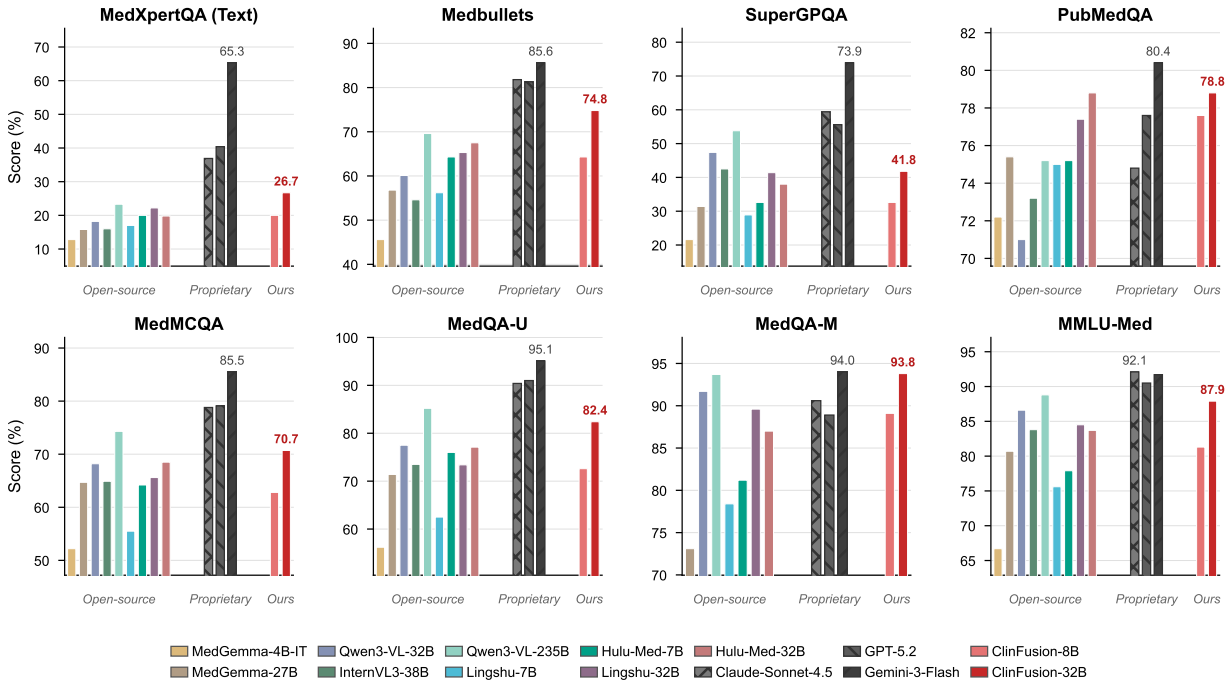
GT answer:
Findings: Trachea, both main bronchi are open. Mediastinal main vascular structures, heart contour, size are normal. Thoracic aorta diameter is normal. Pericardial effusion-thickening was not observed. Thoracic esophagus calibration was normal and no significant tumoral wall thickening was detected. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. When examined in the lung parenchyma window; Aeration of both lung parenchyma is normal and no nodular or infiltrative lesion is detected in the lung parenchyma. Pleural effusion-thickening was not detected. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Liver parenchyma shows diffuse changes in favor of steatosis. Bilateral adrenal glands were normal and no space-occupying lesion was detected. Bone structures in the study area are natural. Vertebral corpus heights are preserved.
Impression: Hepatosteatosis.

Gemini-3-Flash:
Findings: Lungs and airways: The lungs are clear without focal consolidation, suspicious nodules, or masses. No pleural effusion or pneumothorax is present. The trachea and major bronchi are patent. Heart and great vessels: The heart is normal in size. No pericardial effusion is identified. The thoracic aorta and pulmonary arteries are normal in caliber. Mediastinum: No significant mediastinal or hilar lymphadenopathy is seen. Upper abdominal organs: The liver is normal in size and attenuation without focal lesions. The adrenal glands are unremarkable. The spleen, pancreas, and visualized portions of the kidneys are normal. Bone structures: No suspicious lytic or blastic bone lesions are identified. The visualized vertebral bodies and ribs are intact.
Impression: Clear lungs and patent airways. Normal appearance of the heart, great vessels, and mediastinum. Unremarkable upper abdominal organs, including the liver and adrenal glands. No suspicious osseous findings.

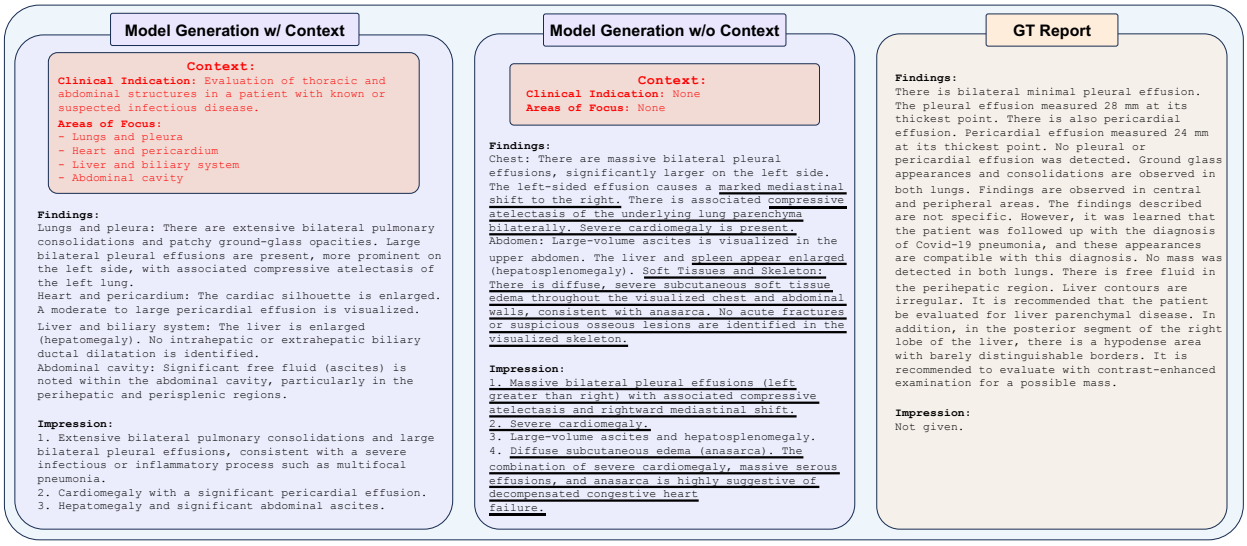
Ours:
Findings:
- Lungs and pleura: The trachea and main bronchi are unobstructed. The mediastinum and trachea are centered. No significant abnormalities are observed in the morphology, size, and proportional distribution of the bilateral lungs. No obvious active lesions are seen in the lung window. No pleural effusion or thickening is observed.
- Mediastinum and heart: The morphology, size, and proportion of the heart and mediastinal major blood vessels are normal. No enlarged lymph nodes are observed in the mediastinum and bilateral hilum. No signs of fluid accumulation are seen in the mediastinum.
- Liver and biliary system: The liver is of normal size, with a regular contour. The liver shows diffusely decreased density. The intrahepatic and extrahepatic bile ducts are not dilated. No abnormalities are observed in the gallbladder.
- Adrenal glands: The bilateral adrenal glands are normal in morphology, size, and density.
- Bone structures: The bone structures are normal in morphology and density. No obvious abnormalities are observed in the vertebral bodies and intervertebral disc spaces. The vertebral alignment is normal.

Impression: Mild fatty liver.

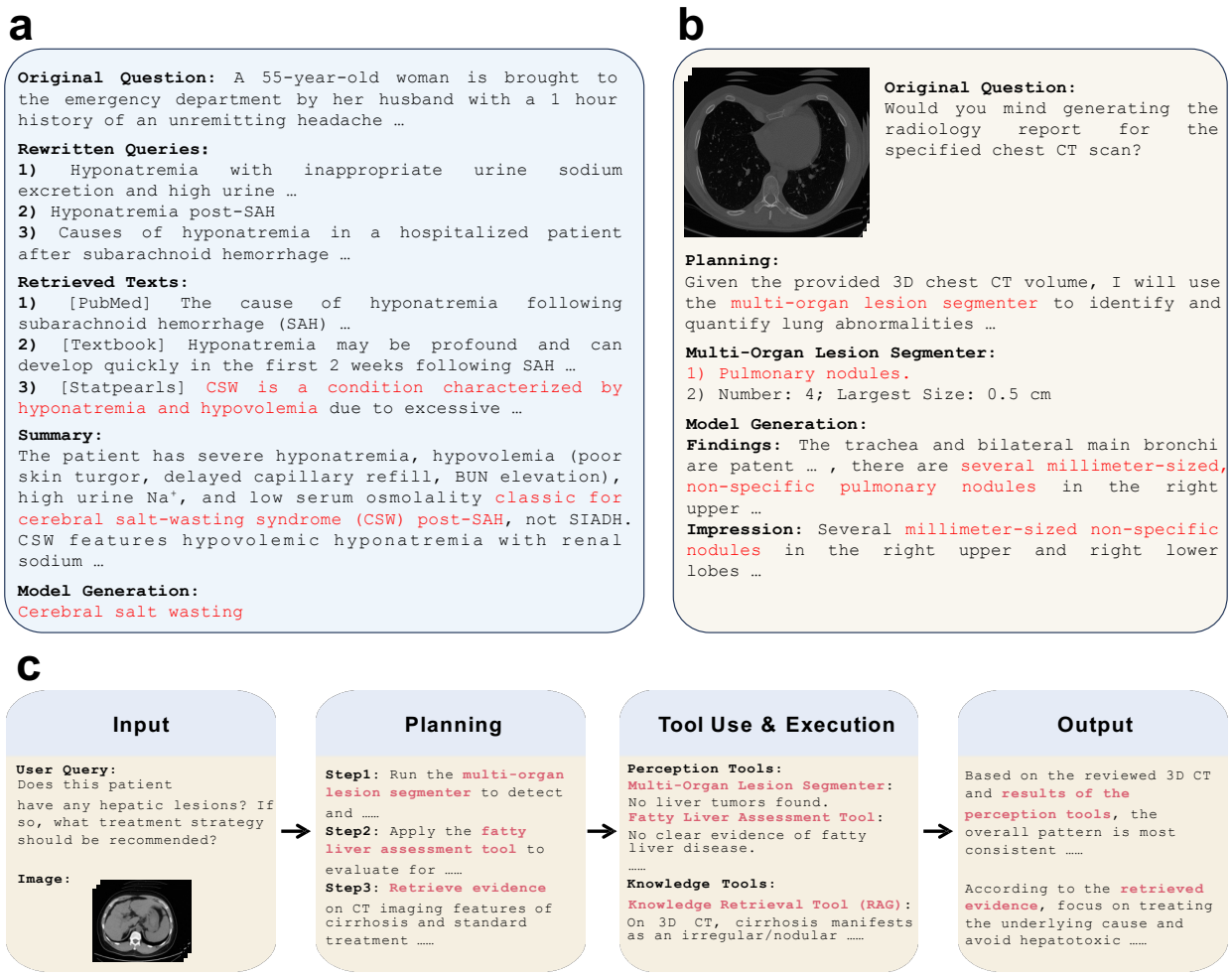
Extended Data Fig. 1 Case visualization of 2D and 3D report generation comparing ClinFusion with other methods. **a**, 2D report generation case. **b**, 3D report generation case. The input prompt directs the MLLMs to focus on specific ROIs.



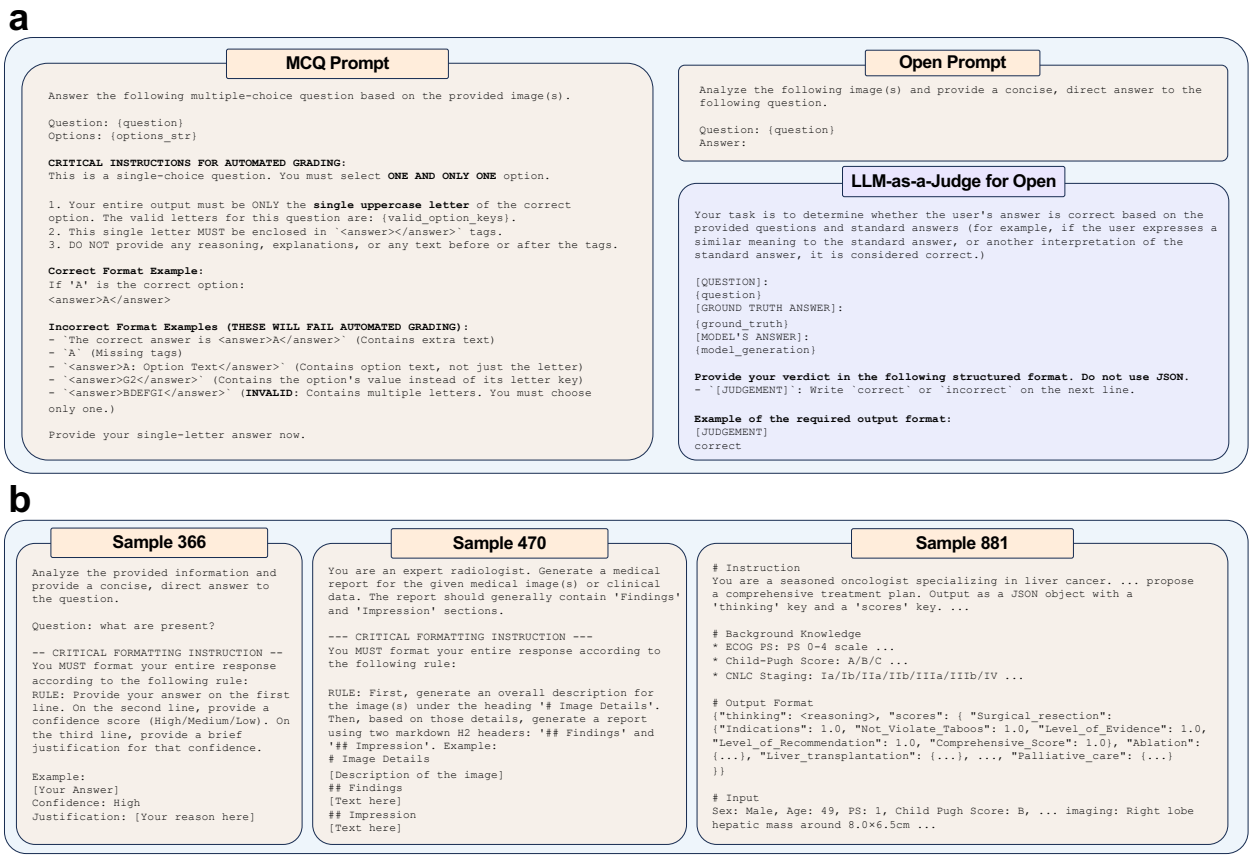
Extended Data Fig. 2 Performance of ClinFusion on textual medical benchmarks. Results on eight text-only benchmarks spanning complex reasoning (MedXpertQA-Text, Medbullets, SuperGPQA), text understanding (PubMedQA), and medical examinations (MedMCQA, MedQA-USMLE, MedQA-MCMLE, MMLU-Medical). Models are compared across open-source (general-purpose and medical MLLMs), proprietary, and our models in a fair protocol re-benchmarked by us.



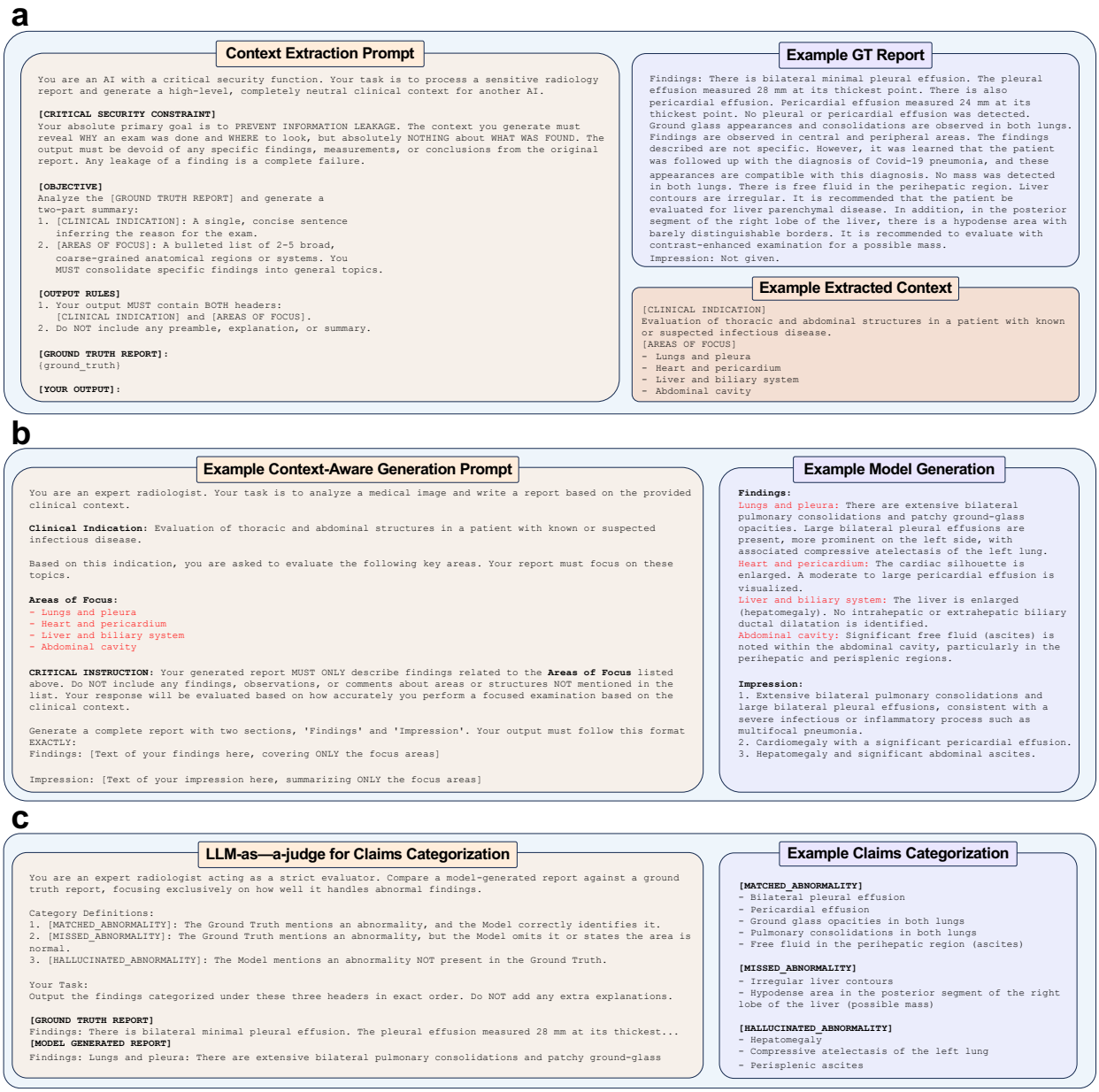
Extended Data Fig. 3 Comparison of report generation with and without clinical context on the same CT scan. We reuse the example from Section 4.2.3. Right: The GT report. Left: The output of ClinFusion when provided with the extracted clinical context (Clinical Indication and Area of Focus). Middle: the output of the same model without any clinical context. Underlined text indicates claims that are *not verifiable* against the GT.



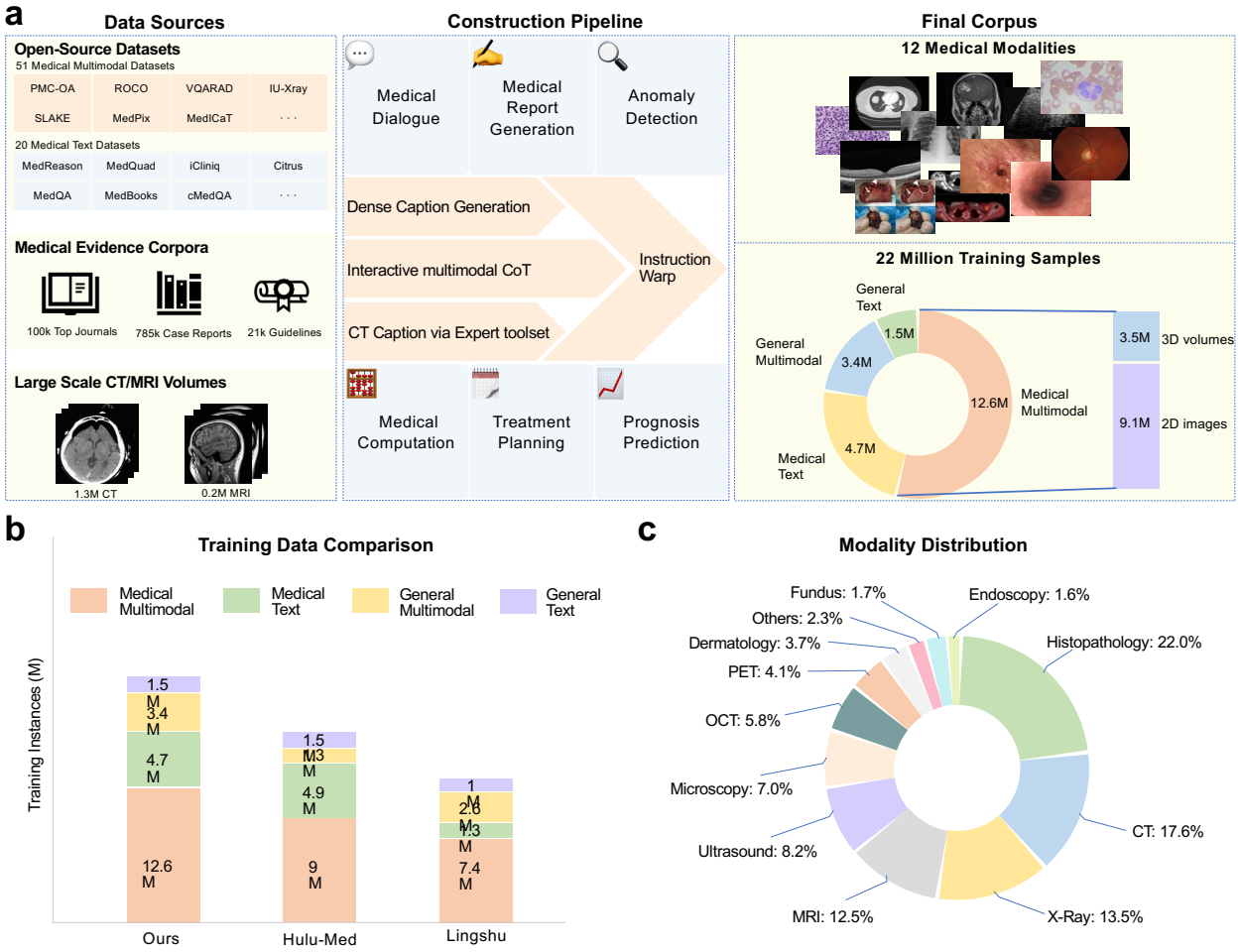
Extended Data Fig. 4 Agentic tool use: case studies and pipeline overview. **a**, Text-only task: The agent rewrites the original clinical question into focused retrieval queries, invokes the hybrid retrieval tool to collect provenance-aware evidence (*e.g.*, PubMed/textbooks/StatPearls), and then synthesizes an evidence-grounded diagnosis in the final generation. **b**, Multimodal task: Radiology report generation from a 3D chest CT volume. The agent performs modality-aware routing, selects a Multi-Organ Lesion Segmenter to produce structured, quantitative findings (*e.g.*, pulmonary nodule count and size), and then integrates the verified tool outputs into a radiology-style report. **c**, Pipeline: Given multimodal clinical inputs (text and medical images), ClinFusion performs task-aware planning to select and invoke appropriate tools (including knowledge tools and perception expert tools), aggregates the returned tool outputs, and generates the final clinical response. In the illustrated liver CT case, ClinFusion executes a three-step workflow by calling a Multi-Organ Lesion Segmenter, a Fatty Liver Assessment Tool, and a Knowledge Retrieval Tool (RAG), and then integrates their results to produce a coherent case analysis and evidence-grounded management recommendations.



Extended Data Fig. 5 Evaluation prompt templates and MedIF-Bench examples. **a**, Prompt templates used for VQA evaluation. Left: The MCQ prompt enforces a strict single-letter output format enclosed in `<answer></answer>` tags, with explicit correct and incorrect format examples to ensure reliable rule-based extraction. Upper right: The open-ended prompt requests a concise, direct answer. Lower right: The LLM-as-a-Judge prompt used to evaluate open-ended responses by comparing the model's answer against the GT answer for semantic correctness. **b**, Three examples from MedIF-Bench (prompts are abbreviated for display; full prompts are substantially longer). Left (Sample 366): An open-ended VQA task where the model must structure its answer on the first line, followed by a confidence score and justification on subsequent lines, representing clinical workflows that require both a concise answer and an explicit confidence assessment. Middle (Sample 470): A report generation task where the model must produce a radiology report with a `# Image Details` section followed by markdown `##` headers for "Findings" and "Impression," mirroring standardized institutional reporting formats. Right (Sample 881): A SEER-based treatment planning task where the model must output a complex nested JSON object containing a reasoning trace and structured scores for multiple treatment modalities, representing integration with clinical decision-support systems.



Extended Data Fig. 6 ROI-Grounded report generation evaluation pipeline. **a**, Context extraction procedure via GPT 4.1 API. The left panel displays the prompt for context extraction, while the right panels present an actual example of an input GT report (top right) and its corresponding extracted context (bottom right). This procedure is designed to minimize leakage of actual diagnostic answers and reveals only the regions of interest. **b**, Context-conditioned report generation. The context successfully extracted in panel (a) is utilized here to condition the generation process. This extracted context is combined with the instruction prompt and presented to our MLLM, strictly constraining the model to generate findings exclusively for the provided Regions of Interest. **c**, LLM-as-a-Judge evaluation. The left panel shows the evaluation prompt presented to the LLM judge (abbreviated for presentation), which strictly defines the categorization rules. The right panel displays the actual categorized output, evaluating the MLLM-generated report from panel (b) against the original GT report from panel (a). In this specific case, the True Positives (TP, matched) are 5, False Negatives (FN, missed) are 2, and False Positives (FP, hallucinated) are 3. Consequently, the metrics are computed as: Precision $= 5/(5+3) = 0.6250$, Recall $= 5/(5+2) = 0.7143$, and F1-score $= 2 \times (0.6250 \times 0.7143)/(0.6250 + 0.7143) = 0.6667$.



Extended Data Fig. 7 Overview of data curation for ClinFusion. **a**, Data sources and construction pipeline. We collect open-source datasets, medical evidence corpora, and large-scale CT/MRI volumes. These data are then processed through the Dense Caption Generation, Medical Multimodal CoT, and Large-Scale CT Caption pipelines, yielding 22.2M training samples spanning 12 medical imaging modalities and supporting diverse clinical tasks, including medical dialogue, medical report generation, anomaly detection, medical computation, treatment planning and prognosis prediction. **b**, Training data scale comparison among ClinFusion, Hulu-Med [7], and Lingshu [6], decomposed into medical multimodal, medical text, general multimodal, and general text instances. **c**, Modality distribution of medical images in the final corpus.

Extended Data Table 1 Data sources used for ClinFusion training. We report an approximate number of training items for each data type in **Scale**. * denotes datasets constructed by our data synthetic pipeline.

Stage	Modality	Datasets	Scale
S1	**Medical multimodal**	PMC-OA [84]; ROCO [85]; ROCOv2 [86]; PubMedVision [87]; MIMIC-CXR [88]; IU-Xray [89]; open-i [89]; MedICaT [90]; MedPix-2.0 [91]; BIOMEDICA Clinical Subset [92]; BIOMEDICA Dermatology Subset [92]; BIOMEDICA Histopathology Subset [92]; BIOMEDICA Microscopy Subset [92]; BIOMEDICA Surgery Subset [92]; Synthetic Case Report Caption*; Synthetic Journal Caption*.	∼3.9M
	General multimodal	Llava-v1.5-Caption [93]; PixMo [94].	∼2.3M
S2	**Medical multimodal**	*Same as S1 Medical Multimodal datasets.*	∼3.8M
	General multimodal	*Same as S1 General multimodal datasets.*	∼2.3M
S3	**Medical multimodal**	LLaVA-Med [95]; Quilt-LLaVA [96]; FairVLMed; CheXpert Plus [46]; Kvasir-VQA [97]; LLaVA-1.5-Instruct [93]; MIMIC-Ext-MIMICCXR-VQA [98]; PathVQA [99]; PMC-VQA [100]; SLAKE [15]; VQA-Med-2019 [101]; VQA-RAD [102]; PMC-VQA-2-recaptions; SLAKE-recaptions; VQA-Med-2019-recaptions; VQA-RAD-recaptions; Path-VQA-recaptions; GMAI-Reasoning10K [103]; Synthetic Medical OCR*; Private CT*; Synthetic Case Report VQA*; Synthetic Journal VQA*.	∼5.3M
	General multimodal	Llava-v1.5-MM [93]; ALLaVA [104].	∼1.1M
	Medical text	MedQuAD [105]; medical-o1-verifiable-problem [106]; medical-o1-reasoning-SFT; ApolloCorpus [107]; Medical-R1-Distill-Data; AlpaCare-MedInstruct-52k [108]; HealthCareMagic-100k [109]; PMC-LLaMA [110]; HuatuoGPT2-GPT4-SFT-140K [111]; MedReason [112]; MedThoughts-8K [113]; MedQA [77]; open-book-CMB-exam [114]; cMedQA [115]; HealthCareMagic [109]; iCliniq-10K [109]; Citrus_S3 [116]; MedBooks-18-CoT [117]; Synthetic Case Report Long CoT*.	∼4.7M
	General text	Llava-v1.5-text-subset [93]; OpenHermes-2.5 [118].	∼1.5M
S4	**Medical multimodal**	CT-Rate-report [16]; Merlin-report [68]; RadGenome-ChestCT [119]; M3D-CAP [67]; OLD-INSPECT [69]; AMOS-MM-report [120]; AbdomenCT*; ChronicCT*; CardiovascularCT*; PulmonaryNodulesCT*.	∼1.3M
S5	**Medical multimodal**	*All S3 medical multimodal datasets* +AMOS-MM-VQA [120]; CT-Rate-VQA [16]; 3D-RAD [121]; Synthetic-M3d-VQA*; Merlin-MCQ [68]; Merlin-VQA [68].	∼6.4M
	General multimodal	*Same as S3 General multimodal datasets.*	∼1.1M
	Medical text	*Same as S3 Medical text datasets.*	∼4.7M
	General text	*Same as S3 General text datasets.*	∼1.5M

Extended Data Table 2 Summary of evaluation benchmarks. We categorize our evaluation framework into five key dimensions: 2D multimodal medical VQA, text-only medical knowledge, 3D volumetric medical VQA, clinical report generation, and instruction-following (# indicates sample counts). Under the Modality column, *Single* and *Multi* distinguish between single-image and multi-image inputs; notably, for 3D, report generation, and instruction-following tasks, the model typically processes a single multimodal file (*e.g.*, one 3D volume or one 2D image) per instance. Task types are classified as Multiple Choice (MCQ), Open-ended (Open), Clinical Report Generation (Report), or Instruction-Following (IF). † denotes a manually curated subset of the original dataset, filtered to ensure data quality and evaluation fidelity.

Benchmark	Modality/Genre	Scale (#)	Type	Lang.	Source
2D Multimodal Medical VQA (Total: 141,587)					
VQA-RAD	Radiology (Single)	451	MCQ & Open	EN	[102]
SLAKE	Radiology (Single)	2,094	MCQ & Open	EN & CN	[15]
PathVQA	Pathology (Single)	6,719	MCQ & Open	EN	[99]
PMC-VQA	Diverse (Single)	33,430	MCQ	EN	[100]
OmniMedVQA	Diverse (Single)	88,996	MCQ	EN	[14]
MedFrameQA	Diverse (Multi)	2,851	MCQ	EN	[122]
GMAI-MMBench	Diverse (Single)	4,550	MCQ	EN	[123]
MedXpertQA (MM)	Diverse (Multi)	2,000	MCQ	EN	[18]
Text-only Medical Knowledge (Total: 17,074)					
MedQA (USMLE)	Clinical Exam	1,273	MCQ	EN	[77]
MedQA (MCMLE)	Clinical Exam	3,426	MCQ	CN	[77]
MedMCQA	Entrance Exam	4,183	MCQ	EN	[124]
MMLU (Medical)	Multi-discipline	1,871	MCQ	EN	[17]
PubMedQA	Research QA	500	MCQ	EN	[19]
MedBullets	Clinical Knowledge	616	MCQ	EN	[125]
SuperGPQA	General Practice	2,755	MCQ	EN	[126]
MedXpertQA (Text)	Diverse	2,450	MCQ	EN	[18]
3D Volumetric Medical VQA (Total: 48,907)					
CT-Rate-VQA†	Non-Contrast Chest CT	12,210	MCQ & Open	EN	[16]
3D-RAD	Non-Contrast Chest CT	33,910	MCQ & Open	EN	[47]
AMOS-MM†	Abdominal CT	2,787	MCQ	EN	[48]
2D/3D Clinical Report Generation (Total: 3,935)					
IU-Xray	Chest X-ray	296	Report	EN	[127]
CheXpert-Plus	Chest X-ray	200	Report	EN	[46]
CT-Rate	Non-Contrast Chest CT	3,039	Report	EN	[16]
AMOS-MM	Abdominal CT	400	Report	EN	[48]
Medical Instruction Following					
MedIF-Bench	Mixed	900	IF	EN	Ours

Extended Data Table 3 The progressive training strategy for ClinFusion. "Trainable Components" lists the parts of the model being updated in each stage.

Stage	Objective	Training Data	Trainable Components
①	**Shallow multi-encoder alignment (2D)** *Align specialist 2D encoders.*	2D Image-Text Pairs	All Encoders, 2D Fusion Modules
②	**Deep vision-language alignment (2D)** *Align fused 2D features with LLM.*	2D Image-Text Pairs	Full Model (All Encoders, 2D Fusion Modules, LLM)
③	**2D Instruction Tuning** *Teach 2D task execution.*	2D Instruction Datasets	Full Model (All Encoders, 2D Fusion Modules, LLM)
④	**Rapid 3D encoder alignment** *Efficiently align 3D encoder.*	3D Volume-Text Pairs	3D Encoder, 3D Fusion Modules
⑤	**Holistic 2D/3D instruction tuning** *Unlock full 2D/3D capabilities.*	2D/3D Instruction Datasets	Full Model (All Encoders, All Fusion Modules, LLM)

Extended Data Table 4 Dataset availability, download links, and access terms for all data sources used in ClinFusion training. Datasets are grouped by training stages described in Section 4.5. The Access column reports the verified license, access category, or source-license inheritance where applicable.

Dataset Name	Link	Access
Application Stage: S1 – Medical multimodal		
PMC-OA	https://huggingface.co/datasets/axiong/pmc_oa	Under CC
ROCO	https://github.com/razorx89/roco-dataset	Under CC
ROCOv2	https://huggingface.co/datasets/eltorio/ROCOv2-radiology	CC BY-NC-SA 4.0
PubMedVision	https://huggingface.co/datasets/FreedomIntelligence/PubMedVision	Apache 2.0
MIMIC-CXR	https://physionet.org/content/mimic-cxr/	PhysioNet License
IU-Xray	https://openi.nlm.nih.gov/	Open Access
open-i	https://openi.nlm.nih.gov/	Open Access
MedICaT	https://github.com/allenai/medicat	Under CC
MedPix-2.0	https://github.com/CHILab1/MedPix-2.0	CC BY-NC-SA 4.0
BIOMEDICA Clinical Subset	https://huggingface.co/datasets/BIOMEDICA/biomedica_webdataset_24M	Under CC
BIOMEDICA Dermatology Subset	https://huggingface.co/datasets/BIOMEDICA/biomedica_webdataset_24M	Under CC
BIOMEDICA Histopathology Subset	https://huggingface.co/datasets/BIOMEDICA/biomedica_webdataset_24M	Under CC
BIOMEDICA Microscopy Subset	https://huggingface.co/datasets/BIOMEDICA/biomedica_webdataset_24M	Under CC
BIOMEDICA Surgery Subset	https://huggingface.co/datasets/BIOMEDICA/biomedica_webdataset_24M	Under CC
Synthetic Case Report Caption	Synthetic Data	Internal Synthetic Data
Synthetic Journal Caption	Synthetic Data	Internal Synthetic Data
Application Stage: S1 – General multimodal		
LLaVA-v1.5-Caption	https://github.com/haotian-liu/LLaVA	CC BY-NC 4.0
PixMo	https://huggingface.co/datasets/allenai/pixmo-cap	ODC-BY v1.0
Application Stage: S2 – Medical multimodal		
Same as S1 Medical multimodal datasets		Same access terms as S1 Medical multimodal
Application Stage: S2 – General multimodal		
Same as S1 General multimodal datasets		Same access terms as S1 General multimodal
Application Stage: S3 – Medical multimodal		
LLaVA-Med	https://github.com/microsoft/LLaVA-Med	CC BY-NC 4.0
Quilt-LLaVA	https://huggingface.co/datasets/wisdomik/QUILT-LLaVA-Instruct-107K	CC BY-NC-ND 3.0
FairVLMed	https://zenodo.org/records/13178701	CC BY-NC-ND 4.0
CheXpert Plus	https://aimi.stanford.edu/datasets/chexpert-plus	Credentialed Access
Kvasir-VQA	https://datasets.simula.no/kvasir-vqa/	CC BY-NC 4.0
LLaVA-1.5-Instruct	https://github.com/haotian-liu/LLaVA	CC BY-NC 4.0
MIMIC-Ext-MIMICCXR-VQA	https://physionet.org/content/mimic-ext-mimic-cxr-vqa/1.0.0/	PhysioNet License
PathVQA	https://huggingface.co/datasets/flaviagiammarino/path-vqa	MIT
PMC-VQA	https://huggingface.co/datasets/RadGenome/PMC-VQA	MIT
SLAKE	https://huggingface.co/datasets/BoKelvin/SLAKE	CC BY 4.0
VQA-Med-2019	https://zenodo.org/records/10499039	CC BY 4.0
VQA-RAD	https://huggingface.co/datasets/flaviagiammarino/vqa-rad	CC0-1.0
PMC-VQA-2-recaptions	https://huggingface.co/datasets/RadGenome/PMC-VQA	Internal Synthetic Data
SLAKE-recaptions	https://huggingface.co/datasets/BoKelvin/SLAKE	Internal Synthetic Data
VQA-Med-2019-recaptions	https://zenodo.org/records/10499039	Internal Synthetic Data
VQA-RAD-recaptions	https://huggingface.co/datasets/flaviagiammarino/vqa-rad	Internal Synthetic Data
Path-VQA-recaptions	https://huggingface.co/datasets/flaviagiammarino/path-vqa	Internal Synthetic Data
GMAI-Reasoning10K	https://huggingface.co/datasets/General-Medical-AI/GMAI-Reasoning10K	CC BY 4.0
Synthetic Medical OCR	Synthetic Data	Internal Synthetic Data
Synthetic Case Report VQA	Synthetic Data	Internal Synthetic Data
Synthetic Journal VQA	Synthetic Data	Internal Synthetic Data
Application Stage: S3 – General multimodal		
LLaVA-v1.5-MM	https://github.com/haotian-liu/LLaVA	CC BY-NC 4.0
ALLaVA	https://huggingface.co/datasets/FreedomIntelligence/ALLaVA-4V	Apache 2.0
Application Stage: S3 – Medical text		
MedQuAD	https://huggingface.co/datasets/lavita/MedQuAD	Open Access
medical-o1-verifiable-problem	https://huggingface.co/datasets/FreedomIntelligence/medical-o1-verifiable-problem	Apache 2.0
medical-o1-reasoning-SFT	https://huggingface.co/datasets/FreedomIntelligence/medical-o1-reasoning-SFT	Apache 2.0
ApolloCorpus	https://huggingface.co/datasets/FreedomIntelligence/ApolloCorpus	Apache 2.0
Medical-R1-Distill-Data	https://huggingface.co/datasets/FreedomIntelligence/Medical-R1-Distill-Data	Apache 2.0
AlpaCare-MedInstruct-52k	https://huggingface.co/datasets/lavita/AlpaCare-MedInstruct-52k	CC BY 4.0
HealthCareMagic-100k	https://huggingface.co/datasets/lavita/ChatDoctor-HealthCareMagic-100k	Apache 2.0
icliniq10k	https://huggingface.co/datasets/zhengComing/iCliniq-10K	Apache 2.0
PMC-LLaMA	https://github.com/chaoyi-wu/PMC-LLaMA	Apache 2.0
HuatuoGPT2-GPT4-SFT-140K	https://huggingface.co/datasets/FreedomIntelligence/HuatuoGPT2-GPT4-SFT-140K	Apache 2.0
MedReason	https://huggingface.co/datasets/UCSC-VLAA/MedReason	CC BY 4.0
MedThoughts-8K	https://huggingface.co/datasets/hw-hwei/MedThoughts-8K	MIT
MedQA	https://github.com/jind11/MedQA	MIT
open-book-CMB-exam	https://huggingface.co/datasets/wangrongsheng/open-book-CMB-exam	Apache 2.0
cMedQA	https://github.com/zhangsheng93/cMedQA2	GPL-3.0
HealthCareMagic	https://huggingface.co/datasets/lavita/ChatDoctor-HealthCareMagic-100k	Apache 2.0
iCliniq-10K	https://huggingface.co/datasets/zhengComing/iCliniq-10K	Apache 2.0
Citrus_S3	https://huggingface.co/datasets/jdh-algo/Citrus_S3	MIT
MedBooks-18-CoT	https://huggingface.co/datasets/dmis-lab/meerkat-instructions	CC BY-NC 4.0
Synthetic Case Report Long CoT	Synthetic Data	Internal Synthetic Data
Application Stage: S3 – General text		
LLaVA-v1.5-text-subset	https://github.com/haotian-liu/LLaVA	CC BY-NC 4.0
OpenHermes-2.5	https://huggingface.co/datasets/teknium/OpenHermes-2.5	Apache 2.0
Application Stage: S4 – Medical multimodal		
CT-Rate-report	https://huggingface.co/datasets/ibrahimhamamci/CT-RATE	CC BY-NC-SA 4.0
Merlin-report	https://github.com/StanfordMIMI/Merlin	MIT
RadGenome-ChestCT	https://huggingface.co/datasets/RadGenome/RadGenome-ChestCT	CC BY 4.0
M3D-CAP	https://huggingface.co/datasets/GoodBaiBai88/M3D-Cap	Apache 2.0
OLD-INSPECT	https://aimi.stanford.edu/datasets/inspect-Multimodal-Dataset...	Credentialed Access
AMOS-MM-report	https://zenodo.org/doi/10.5281/zenodo.10992154	CC BY 4.0
AbdomenCT	Synthetic Data	Internal Synthetic Data
ChronicCT	Synthetic Data	Internal Synthetic Data
CardiovascularCT	Synthetic Data	Internal Synthetic Data
PulmonaryNodulesCT	Synthetic Data	Internal Synthetic Data
Application Stage: S5 – Medical multimodal		
All S3 Medical multimodal datasets		Same access terms as S3 Medical multimodal
AMOS-MM-VQA	https://zenodo.org/doi/10.5281/zenodo.10992154	CC BY 4.0
CT-Rate-VQA	https://huggingface.co/datasets/ibrahimhamamci/CT-RATE	CC BY-NC-SA 4.0
3D-RAD	https://github.com/Tang-xiaoxiao/3D-RAD	MIT
Synthetic-M3d-VQA	Synthetic Data	Internal Synthetic Data
Merlin-MCQ	https://github.com/StanfordMIMI/Merlin	MIT
Merlin-VQA	https://github.com/StanfordMIMI/Merlin	MIT
Application Stage: S5 – General multimodal		
Same as S3 General multimodal datasets		Same access terms as S3 General multimodal
Application Stage: S5 – Medical text		
Same as S3 Medical text datasets		Same access terms as S3 Medical text
Application Stage: S5 – General text		
Same as S3 General text datasets		Same access terms as S3 General text

Supplementary Information

S1 Supplementary Methods: Data Synthesis Pipelines

This section provides the detailed procedures for the four data synthesis pipelines summarized in the Methods (Section 4.3). These pipelines target distinct training bottlenecks: **1)** *instruction warping* diversifies medical instructions to mitigate the loss of instruction-following ability during medical adaptation; **2)** *dense caption generation* converts noisy paper-figure caption pairs into grounded, fine-grained vision–language supervision; **3)** *interactive multimodal CoT annotation* synthesizes long-horizon, visually grounded reasoning trajectories for hard multimodal QA; and **4)** *large-scale CT caption with expert toolset* produces 1.1M radiology-style pseudo-reports for hospital-collected CT volumes that lack reports. The complete workflows for Dense Caption Generation, CT Caption with Expert Toolset, and Interactive Multimodal CoT Annotation are illustrated in Supplementary Fig. 1.

S1.1 Instruction Warping

Medical adaptation tends to erode general instruction-following ability, which in turn constrains how the underlying medical knowledge can be expressed in realistic clinical settings, such as structured outputs (JSON), scope restriction ("only describe the RoI"), negative constraints ("ignore lungs"), or domain-specific formatting required by clinical workflows. To address this, we build an instruction augmentation pipeline via **instruction warping**, inspired by self-instruction, to diversify instructions along multiple axes while preserving medical semantics.

Seed templates. We manually author 120 seed instructions spanning diverse clinical application scenarios and task types, each with explicit formatting requirements (*e.g.*, strict headers, JSON schemas, table columns, and a fixed Final Answer anchor). The seeds are designed to cover four axes: **1)** dialogue roles (*e.g.*, radiologist note, patient-facing explanation, triage assistant); **2)** scenarios (*e.g.*, outpatient consult, ED triage, follow-up comparison, multidisciplinary discussion); **3)** format-following (JSON/tables/bullets/checklists with length and field constraints); and **4)** constraints (scope-only, negative constraints, uncertainty handling, citation requirements).

LLM-based expansion. Building on these seeds, we use GPT-5.1 to automatically expand them into 21k instruction templates. For each data sample, we parse it into structured schema fields and then recompose it using a sampled template, which substantially increases instruction diversity while introducing minimal semantic drift.

S1.2 Dense Caption Generation

A non-trivial portion of large-scale medical multimodal data is collected from paper figures paired with captions. However, these pairs often suffer from low alignment quality: captions typically state only the most critical conclusion and thus provide weak, non-dense supervision; meanwhile, paper figures are frequently composite, mixing medical images with plots, arrows, symbols, and annotations. Such visual clutter introduces noise and also deviates from real clinical imagery, reducing transferability to practical medical settings (Supplementary Fig. 1a).

Sub-figure segmentation and filtering. Following [92], we perform sub-figure segmentation on document screenshots to decompose complex figures into candidate sub-images. We then apply a multi-stage filtering pipeline to retain only high-quality medical imagery: **1)** remove non-image panels (plots/tables); **2)** remove heavily annotated or low-resolution regions; **3)** validate modality and medicalness; and **4)** deduplicate near-duplicates.

Dense caption regeneration. For retained sub-images, we generate dense descriptions by conditioning on the original caption and the visual content, moving from "headline" captions to grounded, fine-grained supervision that inventories visible findings, anatomical context, and modality-specific patterns, while avoiding non-visual speculation. We employ a two-pass generation scheme: **1) Visual inventory.** We use two complementary VLMs, GPT-5.1 and Gemini 3 Pro, to independently produce a visual inventory that enumerates directly observable elements (*e.g.*, findings, locations, and appearance cues). We then retain only the consistent, cross-model

overlapping items as a reliability filter. **2) Clinically coherent description.** The consensus inventory is then used to generate a clinically coherent description with appropriate medical phrasing and structure.

We curate our raw corpus from 100k journal articles and 785k case reports, yielding 749k high-quality medical images paired with captions. Applying the above dense-captioning pipeline produces 610k dense-caption samples, summarized as Synthetic Case Report Caption and Synthetic Journal Caption in Extended Data Table 1. These dense captions are primarily used in early alignment stages to strengthen vision–language grounding, and also serve as rich supervision for later instruction tuning.

S1.3 Interactive Multimodal CoT Annotation

Beyond perception, medical MLLMs must sustain multi-step reasoning over multimodal evidence. We find that many existing medical VQA instances are either too short or can be solved by shallow pattern matching, and thus do not strongly incentivize long-horizon multimodal reasoning. To explicitly train this capability, we construct a hard multimodal reasoning dataset via a multi-agent annotation pipeline, following the core idea of iterative look-again [66] reasoning (Supplementary Fig. 1c).

Difficulty mining. We assign each candidate multimodal QA instance a difficulty score using a fixed solver model. For each instance, we query the solver under a standardized prompt for n stochastic samples, and compute: **1)** the empirical correctness rate (pass@1); and **2)** the average generation length measured in output tokens. In practice, we use Qwen3-VL-8B [2] as the solver model.

Our intuition is that instances with low correctness yet high generation length often induce extensive but unproductive reasoning, which are desirable seeds for constructing reasoning-focused training data. Accordingly, we prioritize instances by:

$$S^{\text{hard}}(\text{sample}) = \big(1 - \text{pass@1}(\text{sample})\big) \cdot \log\big(1 + \overline{Len}(\text{sample})\big), \tag{11}$$

where pass@1(sample) denotes the solver's single-sample accuracy and $\overline{Len}(\text{sample})$ is the average response length. We label instances with $S^{\text{hard}} > 4$ as reasoning-hard samples and further apply diversity-aware sampling to avoid over-representing redundant patterns, yielding 100k diverse reasoning-hard instances.

Multi-agent reasoning. We design a multi-agent pipeline that couples a text-only orchestrator with a multimodal visual grounder through iterative, evidence-seeking interaction. For each instance, the orchestrator has access only to the textual components (question and candidate options) and cannot directly observe the image/volume. It first analyzes the question to identify the visual attributes, anatomical cues, or regions that must be verified, then converts them into targeted fact-checking sub-questions. The visual grounder receives the medical image/volume along with the orchestrator's queries, inspects the visual evidence, and returns grounded observations. We use GPT-5.1 as the orchestrator and Gemini 3 Pro as the visual grounder.

The orchestrator integrates the returned evidence, updates its hypotheses, and either requests additional, more specific visual checks or finalizes the answer once the accumulated evidence is sufficient. This closed-loop interaction proceeds for multiple rounds and enforces that each reasoning step is supported by newly retrieved visual evidence, thereby mitigating grounding drift as reasoning depth grows. We retain only trajectories whose final conclusion matches the reference answer, resulting in 67k high-confidence hard multimodal reasoning instances that are primarily used for instruction tuning to improve long-horizon, visually grounded reasoning fidelity.

S1.4 Large-Scale CT Caption with Expert Toolset

CT is a cornerstone modality for diagnosing and differentiating many diseases, and natively understanding 3D CT volumes is central to real clinical deployment. We therefore construct CT–report aligned supervision at scale through a two-pronged strategy: **1)** collecting open-source CT datasets that already include high-quality reports, and **2)** expanding coverage with hospital-collected CT scans that lack reports by generating report-style pseudo-labels (Supplementary Fig. 1b).

Open-source CT with high-quality reports. We curate 0.2M CT studies from open-source datasets with high-quality radiology reports [67, 16, 68, 69]. These provide clean paired supervision for 3D alignment and report generation evaluation.

Hospital CT collection and pseudo-report generation. To further advance CT-centric modeling, we collaborate with multiple hospitals to curate a large-scale internal dataset comprising 1.5M CT scans. Due to data privacy agreements, we were only permitted to access the de-identified CT images, without paired radiology reports. To enable report-level supervision at scale, we developed an annotation-guided CT report annotation pipeline based on a visual-expert toolset, which primarily consists of a lesion segmentation model [70] and an organ segmentation model [71]. This toolset produces structured pseudo-annotations from the images, which are subsequently transformed into radiology-style captions.

The pipeline proceeds in four steps. **1)** The lesion segmentation model produces lesion masks with anatomical labels, which are aggregated into a structured CT representation capturing clinically relevant attributes (*e.g.*, organ presence, lesion location, and lesion patterns). **2)** For each lesion entry, we retrieve representative CT slices that cover the lesion and overlay lesion-region markers on the selected slices. **3)** The marked slices, together with lesion metadata, are provided to a VLM to generate lesion-focused local captions. **4)** We then synthesize the structured attributes and lesion captions into standardized radiology-style Findings/Impression text via medically constrained generation.

Rigorous quality control is applied throughout, including mask sanity checks (size bounds and topology constraints), cross-slice consistency verification, and text-level validation to reduce spurious outputs and avoid unsupported claims. Using this pipeline, we generate pseudo-report supervision for 1.1M CT cases, substantially expanding the scale and diversity of CT–report aligned data.

The full CT–report aligned corpus (0.2M open-source + 1.1M pseudo-labeled) supports both Stage ④ (rapid 3D encoder alignment) and Stage ⑤ (holistic 2D/3D instruction tuning). The large pseudo-labeled portion improves volumetric diversity and robustness, while the smaller open-source portion provides high-fidelity alignment anchors.

S2 Supplementary Methods: Agentic Tool Use

This section provides the detailed retrieval algorithms for the knowledge tool (Section 4.4.1) and the tool routing and interface design for the perception expert tools (Section 4.4.2) summarized in the Methods.

S2.1 Knowledge Tool: Hybrid Retrieval and Cross-Source Aggregation

The knowledge tool combines two complementary retrieval channels followed by cross-source evidence aggregation.

Hybrid retrieval with concept normalization. Medical queries and documents contain pervasive ambiguity (synonyms, abbreviations, brand–generic drug names, variant spellings). To reduce such ambiguity at the retrieval stage, we perform concept normalization using UMLS before querying: medical entities mentioned in the query are mapped to standardized medical concepts. Retrieval then proceeds through two complementary channels. First, a vector search engine retrieves semantically matched evidence passages from large-scale corpora, improving recall under linguistic variation. Second, KG-based recall leverages two open-source medical knowledge graphs: UMLS [73], which provides standardized medical concept identifiers for robust normalization across synonyms and abbreviations, and PrimeKG [74], which offers rich biomedical relations (*e.g.*, drug–disease, symptom–disease, contraindication) that support link-based expansion to clinically relevant upstream/downstream concepts. This structured recall enables concept-to-evidence routing and reduces misses caused by surface-form mismatch. The two channels are integrated by normalized concepts and semantic similarity, producing a candidate evidence set grounded in explicit medical entities and relations.

Cross-source evidence aggregation. Medical QA often benefits from combining complementary evidence views: vector retrieval typically returns highly query-aligned passages, while KG-based recall provides richer upstream/downstream context anchored to normalized entities. Since these two channels differ in granularity and surface form, we aggregate evidence by clustering retrieved chunks across both channels using semantic similarity and overlap of normalized medical entities, grouping content that states the same clinical claim or offers mutual support from different sources. ClinFusion then summarizes each cluster to preserve the shared clinical conclusion together with key qualifiers (*e.g.*, indication, contraindication, dose constraints) and attaches citations to the supporting chunks. This cluster-then-summarize aggregation jointly leverages

the precision of embedding-based retrieval and the relational breadth of KG recall, producing a compact, evidence-grounded context for faithful, traceable clinical answers.

S2.2 Perception Expert Tools: Tool Descriptions and Routing

Detailed tool descriptions. We provide expanded descriptions of the five perception experts integrated into ClinFusion, each producing structured findings that the agent can directly use for evidence-based reasoning.

- **Multi-organ lesion segmentation expert** [70] segments 29 common lesion types across eight organs: breast, colon, esophagus, stomach, kidney, lung, liver, and pancreas. It enables ClinFusion to ground answers in explicit lesion inventories (how many, where, how large), support tumor burden quantification, and maintain consistency across multi-lesion or multi-organ cases where free-form visual description is error-prone.
- **Fatty liver assessment expert** [79] estimates fatty liver severity from CT and localizes the liver parenchyma region used for assessment. It enables ClinFusion to provide standardized steatosis assessment with traceable imaging evidence (parenchyma region), improving reproducibility for screening, risk stratification, and longitudinal monitoring.
- **Abdominal lymph node segmentation expert** [80] segments abdominal lymph nodes and is sensitive to node size, supporting the detection of lymphadenopathy. It enables ClinFusion to perform structured nodal evaluation (node enumeration + size-based criteria), which is central to oncologic staging and response tracking and is difficult to do reliably via generic vision-language perception.
- **Chest X-ray finding classification expert** [81, 82] predicts structured probabilities for a broad set of thoracic findings from chest radiographs. In our system, this expert combines TorchXRayVision, which provides standardized preprocessing and strong pre-trained chest X-ray models for 18 common findings, with CheXFound, a chest X-ray foundation model pretrained on up to one million images and supporting classification over 40 findings. This expert enables ClinFusion to supplement free-form image understanding with explicit pathology-level evidence, improving robustness and consistency in report generation and medical VQA.
- **Chest X-ray finding generation expert** [83] generates clinically coherent chest X-ray findings descriptions from radiographs. Built on CheXagent, a vision-language foundation model trained on the large-scale CheXinstruct dataset and evaluated across diverse chest X-ray interpretation tasks, this expert is mainly used to draft findings-style descriptions that summarize the visible thoracic abnormalities and their imaging context. It enables ClinFusion to obtain richer sentence-level imaging evidence when classification scores alone are insufficient, improving the completeness and fluency of chest X-ray report drafting and finding-oriented question answering.

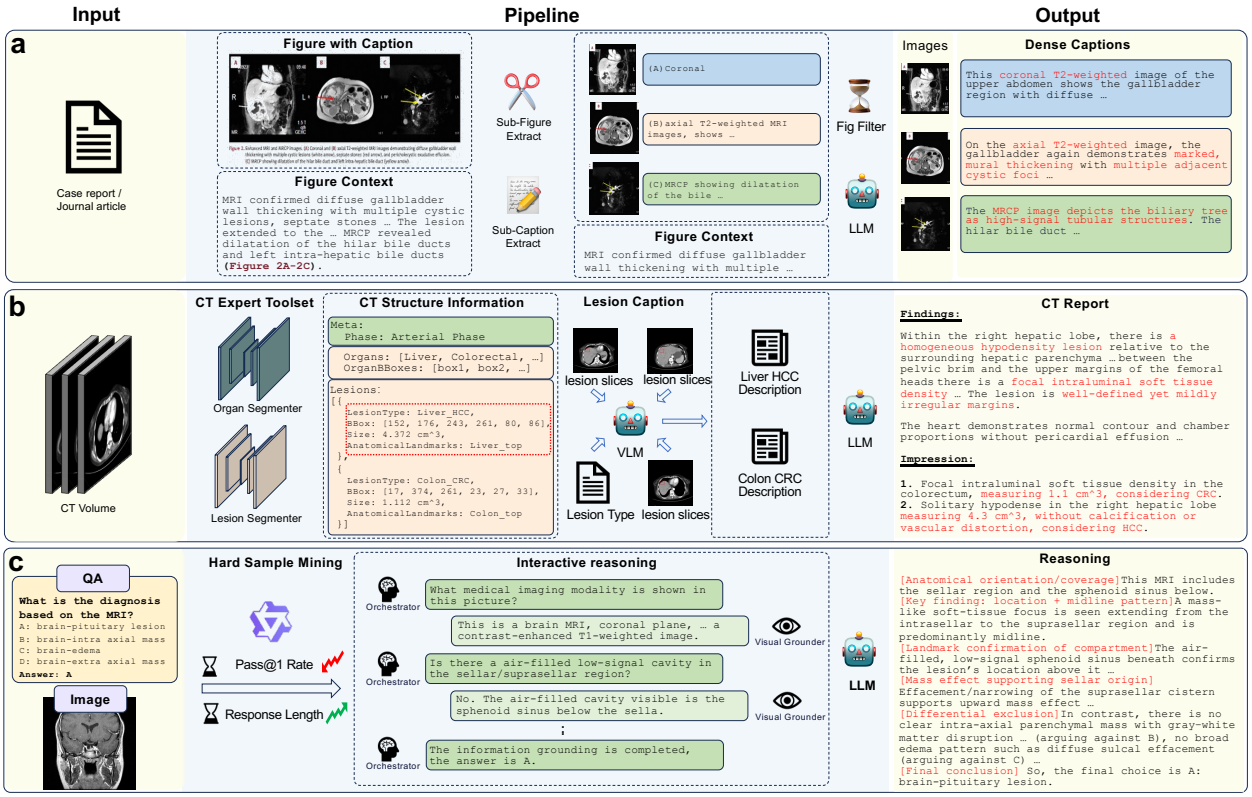
Tool routing and schema-first interface. ClinFusion does not follow a rigid workflow; instead it seeks the specific visual evidence required by the user's question and calls tools accordingly. When the query implies that the answer depends on precise localization or measurement——such as requesting the largest lesion size, asking whether lymph nodes are enlarged, asking for fatty liver grade, or asking for lesion counts in a target organ—ClinFusion treats this as an evidence gap and initiates tool calls rather than relying on approximate visual judgment. To make tool use stable, controllable, and consumable by ClinFusion, we adopt a schema-first interface. Each perception expert is registered with a capability card that specifies the supported anatomies and label set, the measurable endpoints such as diameters, volumes, and severity grades, and the types of clinical questions the tool can resolve. This explicit specification enables intentional routing. Tool outputs are normalized into compact evidence objects and then verbalized into natural-language evidence statements before being fed back to ClinFusion. Overly granular numeric fields such as voxel-level coordinates are mapped to coarse standardized location descriptors, *e.g.*, `liver_top`, which reduces sensitivity to coordinate noise while preserving clinically actionable localization.

S3 Supplementary Ablation Studies

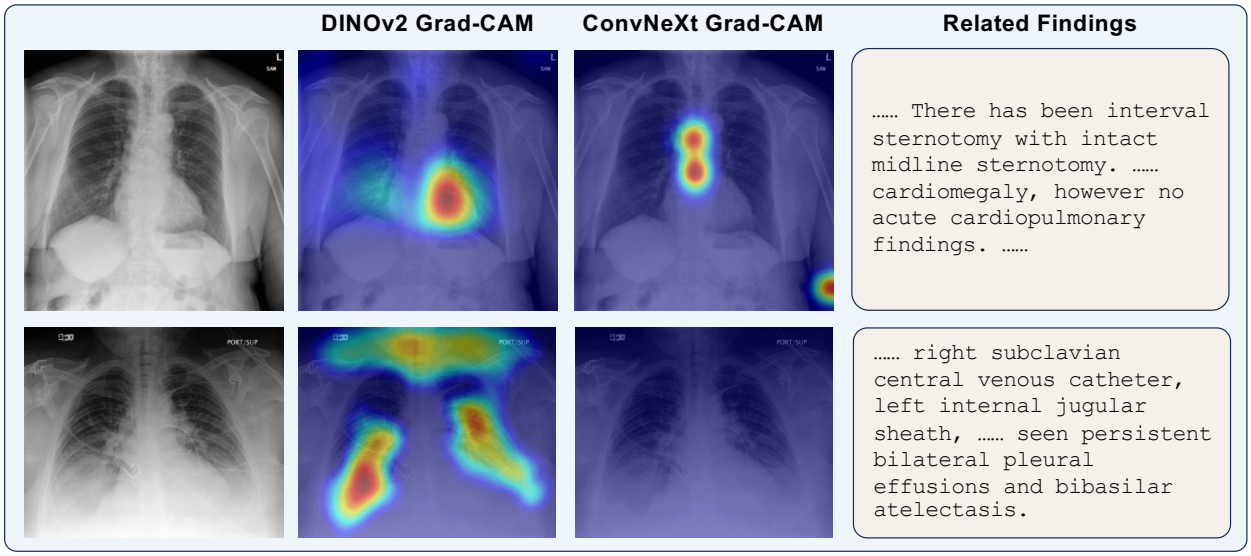
Stochastic residual regularization shows benefits to 3D volumetric understanding. Based on the above experiments, we add the 3D ViT as the fourth vision encoder to enable native volumetric understanding. We train the model with all five stages with partial 2D and 3D data to ensure fast feedback. Our prior experiments indicated an effect of diminishing returns when adding a third vision encoder, suggesting that later-stage components may struggle to integrate effectively into an already powerful model. We hypothesized that a stronger regularization scheme could mitigate this side effect by encouraging more robust feature learning. To test this, we evaluated the impact of applying stochastic residual regularization during training. The results, presented in Supplementary Table 12, provide evidence for our hypothesis. Employing the stochastic approach leads to a clear improvement in both the average 3D VQA score (from 64.8 to 65.8) and the average report generation metrics (from 23.4 / 13.3 / 15.1 to 23.7 / 14.0 / 15.7). However, we also observed that these benefits are contingent on sufficient training duration. In shorter training runs, the performance gap between the stochastic and deterministic models was negligible or even inverted. This suggests that while stochastic regularization provides a more robust learning signal, it requires adequate optimization steps to fully manifest its advantages.

Interpreting the functionality of additional encoders. To qualitatively interpret how the compositional vision architecture contributes to the final prediction, we employ Grad-CAM [128] to visualize the regions of focus for our specialist encoders, DINOv2 and ConvNext. We analyze two chest X-ray report generation examples to illustrate how these encoders exhibit a complementary division of labor, which is then unified by our CaSL Fusion mechanism to produce a comprehensive and accurate report. As shown in Supplementary Fig. 2, the specialist encoders demonstrate distinct yet synergistic attentional patterns. In the first example (top row), which describes a post-operative chest, the Grad-CAM visualizations reveal a clear separation of concerns. DINOv2, known for its strong semantic feature extraction, directs its focus squarely on the cardiac silhouette. This global, semantic focus is critical for assessing overall organ size and shape, leading to the identification of cardiomegaly. In comparison, ConvNext, with its powerful inductive bias for local patterns and textures, precisely highlights the midline sternotomy wires. This demonstrates its specialization in identifying fine-grained, man-made objects or textural abnormalities. The second example (bottom row) further reinforces this observation. The case involves multiple findings, including bilateral pleural effusions and atelectasis. Here, DINOv2 casts a wide attentional net, focusing on the bilateral pleural effusions and the lung bases, which are key for identifying the reported bibasilar atelectasis. Intriguingly, in this specific case, ConvNext's visualization does not show significant activation, suggesting that its role may be more nuanced or context-dependent. This might indicate that for cases dominated by large-scale fluid and density changes (like effusions), DINOv2's global semantic understanding dominates, while ConvNext's contribution becomes more critical when fine-grained textural details are diagnostically relevant. This dynamic interplay, managed by our cascaded fusion, allows ClinFusion to flexibly leverage the most relevant visual evidence for any useful case, leading to a more robust and holistic perception.

S4 Supplementary Figures



Supplementary Fig. 1 Data synthesis pipeline of ClinFusion. **a**, Dense caption generation. We parse case reports and journal articles to obtain figure images together with their original captions and surrounding context, segment composite figures into sub-panels, and extract panel-wise subcaptions aligned to each sub-figure. After multi-stage filtering to remove non-clinical, heavily annotated, or low-quality regions, we regenerate dense, visually grounded descriptions by conditioning on both the sub-image and the original (sub)caption/context. **b**, CT annotation with visual expert toolset. For CT volumes without paired reports, we employ a CT Visual Expert to produce a structured CT representation that includes organ context and lesion-level attributes (*e.g.*, type, coarse 3D location, and extent proxies). Using the lesion entries in this representation, we retrieve representative CT slices that cover each lesion and overlay lesion-region markers on the selected slices. The marked slices, together with the corresponding lesion metadata, are then provided to a VLM to generate lesion-focused captions. Finally, these lesion captions and global structured attributes are consolidated by an LLM into standardized radiology-style Findings/Impression pseudo-reports. **c**, Interactive multimodal CoT annotation for reasoning-hard samples. We first score candidate multimodal QA instances with Qwen3-VL and prioritize samples that elicit long analytical responses yet exhibit low pass@1, indicating failures in long-horizon grounded reasoning rather than superficial knowledge gaps. The selected hard instances are then solved with a carefully designed interactive reasoning protocol that separates a text-only orchestrator from a multimodal visual grounder. Given the question, the orchestrator decomposes it into a sequence of visual fact-check queries, iteratively interrogates the visual grounder to obtain image-grounded evidence, and terminates once the accumulated evidence is sufficient to support a final answer. After completion, an additional LLM pass rewrites the full dialogue trajectory into a coherent reasoning trace for instruction tuning.



Supplementary Fig. 2 Grad-CAM visualizations revealing the complementary attentional focus of specialist encoders. For each example, we show the original image, the Grad-CAM heatmaps for DINOv2 and ConvNext, and the corresponding related findings in the report.

S5 Supplementary Tables

Supplementary Table 1 Results on 2D Multimodal Medical Benchmarks. We list results of leading proprietary models, leading generalist MLLMs, and other medical MLLMs. **Bold** results denote the best scores in medical MLLMs and underlined results denote the second-best scores. ♡ denotes that the model does not support multi-image input. † indicates that results were re-benchmarked using open-source model weights or model APIs within our own evaluation framework.

	Multi-modality Benchmarks				Specific-modality Benchmarks			Reasoning Benchmark
Models	OM.VQA	PMCVQA	MedF	GMAI	VQA-RAD	SLAKE	PathVQA	MedXQA
Proprietary Models								
GPT-5.2†	68.4	56.7	52.1	58.4	73.4	79.4	54.0	52.9
Gemini-3-Flash†	82.0	64.4	53.3	67.8	71.0	80.3	60.6	78.0
Claude-Sonnet-4.5†	67.1	58.4	54.2	47.6	63.6	72.6	51.7	42.8
General-purpose MLLMs								
Qwen2.5-VL-7B†	63.3	51.7	46.9	44.0	59.6	67.4	43.4	22.5
Qwen3-VL-8B†	76.4	54.0	42.1	54.0	61.9	70.4	42.8	24.9
InternVL3-8B	79.1	53.8	–	–	65.4	72.8	48.6	22.4
InternVL3-14B	78.9	54.1	–	–	66.3	72.8	48.0	23.1
Qwen2.5-VL-32B†	67.2	53.2	51.5	47.4	72.7	77.2	46.3	25.4
Qwen3-VL-32B†	76.8	56.5	52.0	56.3	66.1	76.9	48.6	31.2
InternVL3-38B	79.8	56.6	–	–	65.4	72.7	51.0	25.2
Qwen3-VL-235B†	75.8	56.2	50.3	54.9	67.8	75.5	51.1	33.7
Medical MLLMs								
— Models < 10B —								
BiomedGPT♡	27.9	27.6	–	–	16.6	13.6	11.3	–
Med-R1-2B	–	47.4	–	–	39.0	54.5	15.3	21.1
MedVLM-R1-2B	77.6	48.8	–	–	49.2	56.3	36.0	21.4
HealthGPT-M3	71.5	55.4	–	–	56.8	70.8	55.4	22.4
BioMediX2-8B	66.0	41.8	–	–	55.7	54.1	34.6	21.9
LLaVA-Med-7B	34.8	22.7	–	–	46.6	51.9	35.2	20.8
MedGemma-4B-IT	70.7	49.2	–	–	72.3	78.2	48.1	25.4
HuatuoGPT-V-7B	74.3	53.1	–	–	67.6	68.1	44.8	23.2
Lingshu-7B	82.9	56.3	–	–	67.9	83.1	61.9	26.7
Hulu-Med-7B	84.2	66.8	–	–	**78.0**	86.8	65.6	29.0
MedGemma-1.5-4B-IT†	47.5	18.7	40.5	21.8	55.2	63.6	34.7	18.5
Lingshu-7B†	83.4	57.4	58.2	51.8	64.3	81.5	57.3	26.4
Hulu-Med-7B†	79.4	61.7	54.8	51.8	68.5	74.7	54.2	25.6
ClinFusion-8B	**89.6**	**67.8**	**64.3**	**57.9**	76.3	**91.1**	**68.7**	**32.4**
— Models > 10B —								
HealthGPT-14B	75.2	56.4	–	–	65.0	66.1	56.7	24.7
HuatuoGPT-V-34B	74.0	56.6	–	–	61.4	69.5	44.4	22.1
Lingshu-32B	83.4	57.9	–	–	76.7	86.7	65.5	30.9
MedDr-40B♡	64.3	13.9	–	–	65.2	66.4	53.5	–
Hulu-Med-14B	85.1	68.9	–	–	76.1	86.5	64.4	30.0
Hulu-Med-32B	84.6	69.4	–	–	**81.4**	85.7	67.3	34.0
MedGemma-27B-IT†	60.0	49.2	49.5	37.3	65.6	75.5	43.4	23.6
Lingshu-32B†	83.4	62.3	58.9	55.6	73.2	87.6	63.4	30.5
Hulu-Med-32B†	80.0	63.2	61.6	57.3	75.8	82.6	63.1	32.6
ClinFusion-32B	**89.9**	**70.8**	**69.4**	**60.2**	77.2	**92.5**	**72.4**	**38.7**

Supplementary Table 2 Results on 2D Report Generation Benchmarks. **Bold** results denote the best scores among medical MLLMs, and underlined results denote the second-best scores. † indicates that results were re-benchmarked using open-source model weights or model APIs within our own evaluation framework.

Models	CheXpert-Plus			IU-XRAY		
	Precision	Recall	F1	Precision	Recall	F1
Proprietary Models						
GPT-5.2†	49.1	38.1	40.2	55.3	50.3	50.6
Gemini-3-Flash†	38.4	33.5	33.0	56.8	55.2	54.4
Claude-Sonnet-4.5†	33.8	28.3	28.9	39.2	39.2	37.7
General-purpose MLLMs						
Qwen2.5-VL-7B†	13.7	11.5	12.0	33.0	32.9	32.4
Qwen3-VL-8B†	20.0	16.3	16.8	37.0	37.3	36.7
Qwen2.5-VL-32B†	22.2	16.2	17.3	33.5	31.9	31.9
Qwen3-VL-32B†	30.6	27.9	27.3	40.8	41.0	40.0
Qwen3-VL-235B†	29.5	25.2	25.6	43.2	41.4	41.2
Medical MLLMs						
— *Models < 10B* —						
MedGemma-1.5-4B-IT†	34.1	23.1	25.4	43.1	38.9	39.8
Lingshu-7B†	26.2	17.8	19.9	43.7	39.3	40.4
Hulu-Med-7B†	40.5	29.1	31.9	51.1	45.3	46.5
ClinFusion-8B	<u>46.4</u>	<u>35.3</u>	<u>37.8</u>	<u>63.3</u>	<u>55.8</u>	<u>57.3</u>
ClinFusion-8B + Agentic tools	**57.3**	**49.8**	**50.2**	**64.5**	**59.4**	**59.9**
— *Models > 10B* —						
MedGemma-27B-IT†	26.9	21.9	22.5	41.3	40.7	40.1
Lingshu-32B†	28.0	21.5	22.8	44.5	40.5	41.3
Hulu-Med-32B†	51.0	35.5	39.6	53.5	46.8	48.2
ClinFusion-32B	<u>56.1</u>	<u>46.6</u>	<u>48.0</u>	<u>60.2</u>	<u>54.7</u>	<u>55.6</u>
ClinFusion-32B + Agentic tools	**60.2**	**55.4**	**54.8**	**60.8**	**56.1**	**56.3**

Supplementary Table 3 Results on Textual Medical Benchmarks. We list results of leading proprietary models, leading generalist MLLMs, and other medical MLLMs. **Bold** results denote the best scores in medical MLLMs and underlined results denote the second-best scores. † indicates that results were re-benchmarked using open-source model weights or model APIs within our own evaluation framework.

	Complex Reasoning Benchmarks			Text Understanding Benchmark	Medical Exam Benchmarks			
Models	MedXQA	Medbullets	SGPQA	PubMedQA	MedMCQA	MedQA-U	MedQA-M	MMLU-Med
Proprietary Models								
GPT-5.2	40.4	81.3	55.6	77.6	79.1	91.0	88.9	90.5
Gemini-3-Flash	65.3	85.6	73.9	80.4	85.5	95.1	94.0	91.7
Claude-Sonnet-4.5	36.9	81.8	59.4	74.8	78.8	90.4	90.6	92.1
General-purpose MLLMs								
Qwen2.5-VL-7B†	12.0	43.0	25.8	74.6	53.2	56.4	76.9	74.1
Qwen3-VL-8B†	14.6	51.9	36.9	72.2	60.7	66.2	87.2	80.0
InternVL3-8B	13.1	48.5	31.2	75.4	57.7	62.1	–	77.5
InternVL3-14B	14.1	49.5	37.9	77.2	62.0	70.1	–	81.7
Qwen2.5-VL-32B†	15.8	54.2	39.3	71.6	62.9	70.8	90.0	83.3
Qwen3-VL-32B†	18.2	60.1	47.4	71.0	68.2	77.5	91.7	86.6
InternVL3-38B	16.0	54.6	42.5	73.2	64.9	73.5	–	83.8
Qwen3-VL-235B†	23.3	69.6	53.8	75.2	74.3	85.2	93.7	88.8
Medical MLLMs								
— Models < 10B —								
MedVLM-R1-2B	11.8	33.8	19.1	66.4	39.7	42.3	–	51.8
BioMediX2-8B	13.4	45.9	25.2	75.2	52.9	58.9	–	68.6
MedGemma-4B-IT	12.8	45.6	21.6	72.2	52.2	56.2	–	66.7
HealthGPT-M3	11.5	41.4	18.9	57.8	54.2	55.0	–	72.5
LLaVA-Med-7B	9.90	34.4	16.1	26.4	39.4	42.0	–	50.6
HuatuoGPT-V-7B	10.1	40.9	21.9	72.8	51.2	52.9	–	69.3
Lingshu-7B	16.5	56.2	26.3	76.6	55.9	63.3	–	74.5
Hulu-Med-7B	19.6	61.5	31.1	77.4	67.6	73.5	–	79.5
MedGemma-1.5-4B-IT†	9.18	33.4	2.25	6.26	6.29	35.7	4.23	10.3
Lingshu-7B†	17.0	56.2	28.9	75.0	55.5	62.5	78.4	75.6
Hulu-Med-7B†	20.0	64.3	32.6	75.2	64.2	76.0	81.2	77.9
ClinFusion-8B	20.0	64.3	32.6	77.6	62.8	72.6	**89.1**	81.3
ClinFusion-8B + Agentic tools	**27.7**	**69.5**	**45.4**	**79.1**	**70.1**	**83.7**	84.3	**86.0**
— Models > 10B —								
HealthGPT-14B	11.3	39.8	25.7	68.0	63.4	66.2	–	80.2
Lingshu-32B	22.7	65.4	41.1	77.8	66.1	74.7	–	84.7
HuatuoGPT-V-34B	11.4	42.7	26.5	72.2	54.7	58.8	–	74.7
MedDr-40B	12.0	44.3	24.0	77.4	38.4	59.2	–	65.2
Hulu-Med-14B	23.2	68.5	37.7	79.8	70.4	78.1	–	83.3
Hulu-Med-32B	24.2	68.8	41.8	**80.8**	72.8	80.4	–	85.6
MedGemma-27B-IT†	15.8	56.8	31.4	75.4	64.7	71.4	73.1	80.7
Lingshu-32B†	22.2	65.3	41.4	77.4	65.6	73.4	89.6	84.5
Hulu-Med-32B†	19.8	67.5	38.0	78.8	68.5	77.1	87.0	83.7
ClinFusion-32B	26.7	74.8	41.8	78.8	70.7	82.4	93.8	87.9
ClinFusion-32B + Agentic tools	**31.2**	**77.9**	**50.9**	79.7	**74.0**	**86.6**	**94.5**	**89.9**

Supplementary Table 4 Results on 3D Multimodal Medical Benchmarks. **Bold** results denote the best scores among medical MLLMs, and underlined results denote the second-best scores. † indicates that results were re-benchmarked using open-source model weights or model APIs within our own evaluation framework.

	AMOS	3D-RAD		CT-Rate		AMOS Report			CT-Rate Report		
Models	MCQ	MCQ	Open	MCQ	Open	Precision	Recall	F1	Precision	Recall	F1
Proprietary Models											
GPT-5.2	63.8	67.1	37.6	66.9	59.1	29.9	11.5	14.5	21.0	12.7	14.1
Gemini-3-Flash	64.2	58.8	30.3	70.9	51.9	40.7	19.1	23.4	29.3	17.9	20.2
Claude-Sonnet-4.5	62.6	58.5	30.0	73.2	51.7	12.8	6.49	7.35	15.5	11.8	12.3
General-purpose MLLMs											
Qwen2.5-VL-7B†	53.4	40.8	28.6	67.3	44.5	6.31	4.18	4.69	10.6	9.42	9.61
Qwen3-VL-8B†	50.2	52.7	28.5	59.6	44.7	11.6	6.51	7.30	22.6	16.3	17.3
Qwen2.5-VL-32B†	55.8	46.4	24.0	64.0	53.1	6.55	4.09	4.57	16.9	13.7	14.2
Qwen3-VL-32B†	58.9	64.7	30.1	67.1	53.6	12.6	7.61	8.37	30.6	18.2	20.5
Qwen3-VL-235B†	57.1	47.7	28.9	62.3	50.9	14.6	7.25	8.53	21.6	16.1	16.9
Medical MLLMs											
— Models < 10B —											
MedGemma-1.5-4B-IT†	57.1	38.1	23.4	75.6	54.8	14.4	6.53	7.74	13.6	10.5	11.0
Lingshu-7B†	61.7	68.4	18.0	70.7	48.9	7.96	4.35	5.04	11.1	9.61	9.86
Hulu-Med-7B†	65.7	73.5	33.2	77.6	50.2	17.5	9.18	10.7	27.4	16.0	18.3
ClinFusion-8B	80.2	79.1	35.7	86.3	64.1	20.1	10.7	12.0	28.8	20.5	21.6
ClinFusion-8B + Agentic tools	**85.7**	**81.3**	**44.1**	**89.1**	**65.5**	**23.2**	**17.1**	**16.6**	**30.0**	**23.5**	**25.3**
— Models > 10B —											
MedGemma-27B-IT†	54.3	36.7	25.1	72.4	45.4	11.6	6.00	7.07	13.2	10.2	10.7
Lingshu-32B†	63.8	70.2	25.3	70.1	50.0	9.65	6.41	7.10	15.3	11.5	12.2
Hulu-Med-32B†	73.9	79.6	37.6	83.3	56.9	22.5	9.97	12.5	34.4	21.4	23.4
ClinFusion-32B	81.7	79.1	36.2	89.0	62.1	26.7	13.7	16.1	34.5	21.4	23.9
ClinFusion-32B + Agentic tools	**86.3**	**81.6**	**45.6**	**91.7**	**65.4**	**28.4**	**17.8**	**19.1**	**34.9**	**25.9**	**27.6**

Supplementary Table 5 Results on our proposed MedIF-Bench. We report scores on individual tasks and composite instruction-following scores. **Bold** results denote the best scores among medical MLLMs, and underlined results denote the second-best scores. † indicates that results were re-benchmarked using open-source model weights or model APIs within our own evaluation framework.

	Individual Metrics							Composite IF Scores		
Models	MCQ	MCQ with context	Open	Prognosis	Report Generation	Seer	Treatment	MM-IF	Text-IF	Overall-IF
Proprietary Models										
GPT-5.2	100.0	100.0	87.0	100.0	84.0	100.0	100.0	92.0	100.0	96.0
Gemini-3-Flash	98.0	100.0	100.0	98.9	97.0	83.0	97.0	98.9	94.3	96.6
Claude-Sonnet-4.5	73.5	76.0	90.0	96.2	80.0	100.0	100.0	86.0	87.0	86.5
General-purpose MLLMs										
Qwen2.5-VL-7B†	100.0	100.0	99.0	100.0	100.0	100.0	99.0	99.7	99.8	99.8
Qwen3-VL-8B†	99.0	100.0	99.0	100.0	99.0	95.0	100.0	99.4	98.4	98.9
Qwen2.5-VL-32B†	77.0	83.0	88.0	100.0	89.0	20.0	100.0	87.3	72.7	79.3
Qwen3-VL-32B†	100.0	100.0	91.0	100.0	55.0	100.0	100.0	85.1	100.0	93.3
Qwen3-VL-235B†	100.0	100.0	99.0	100.0	100.0	99.0	100.0	99.7	99.8	99.7
Medical MLLMs										
— Models < 10B —										
MedGemma-1.5-4B-IT†	50.5	21.0	97.0	3.8	0.0	0.0	0.0	49.9	8.9	27.4
Lingshu-7B†	80.5	92.0	67.0	100.0	35.0	88.0	100.0	62.3	95.5	80.4
Hulu-Med-7B†	99.0	100.0	100.0	98.9	96.0	69.0	100.0	98.9	92.0	95.1
ClinFusion-8B	95.0	100.0	99.0	100.0	98.0	99.0	99.0	97.0	99.1	**98.1**
— Models > 10B —										
MedGemma-27B-IT†	79.0	85.0	100.0	100.0	100.0	92.0	96.0	93.4	89.8	91.4
Lingshu-32B†	82.5	100.0	73.0	100.0	31.0	100.0	91.0	61.7	99.8	82.6
Hulu-Med-32B†	100.0	100.0	98.0	100.0	89.0	100.0	100.0	96.4	100.0	98.4
ClinFusion-32B	100.0	100.0	91.0	100.0	100.0	100.0	100.0	97.5	100.0	**98.9**

Supplementary Table 6 Ablation study on the necessity of clinical context for report generation. We compare two models—ClinFusion-8B and Lingshu-7B [6]—under two settings: with the extracted clinical context ("w/ Context") and without it ("w/o Context"), where the Clinical Indication and Area of Focus are removed from the generation prompt.

Model	Setting	AMOS Report			CT Rate Report			CheXpert-Plus			IU-XRAY		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
ClinFusion-8B	w/ Context	20.1	10.7	12.0	28.8	20.5	21.6	46.4	35.3	37.8	63.3	55.8	57.3
	w/o Context	0.89	0.39	0.46	7.99	3.68	4.71	13.1	10.7	10.8	10.6	8.43	8.87
Lingshu-7B	w/ Context	7.96	4.35	5.04	11.1	9.61	9.86	26.2	17.8	19.9	43.7	39.3	40.4
	w/o Context	0.17	0.12	0.13	5.83	4.92	4.91	11.1	8.58	8.94	10.7	8.10	8.76

Supplementary Table 7 Comparison of RadGraph-F1 and our LLM-as-a-Judge metric on CheXpert-Plus and IU-XRAY. Four model variants are evaluated, including a Qwen3-VL-8B (Long Output) variant prompted to encourage verbose generation in order to expose length bias. RadGraph-F1 reports entity F1 and entity-relation F1, and our LLM-as-a-Judge reports Precision, Recall, and F1.

Model	CheXpert-Plus					IU-XRAY				
	RadGraph-F1		LLM-as-a-Judge (Ours)			RadGraph-F1		LLM-as-a-Judge (Ours)		
	Ent. F1	Ent.-Rel. F1	Prec.	Rec.	F1	Ent. F1	Ent.-Rel. F1	Prec.	Rec.	F1
ClinFusion-8B	23.2	21.0	46.4	35.3	37.8	38.7	36.6	63.3	55.8	57.3
Lingshu-7B	20.0	17.9	26.2	17.8	19.9	35.1	34.6	43.7	39.3	40.4
Qwen3-VL-8B	13.3	11.8	20.0	16.3	16.8	17.8	16.3	37.0	37.3	36.7
Qwen3-VL-8B (Long Output)	20.6	18.8	13.8	18.9	15.1	34.7	33.0	29.4	42.1	33.6

Supplementary Table 8 Ablation study on different fusion mechanisms for the compositional vision encoder. We compare the performance of a single encoder (Qwen ViT baseline) against a dual-encoder setup (Qwen ViT + DINOv2) using various fusion strategies. All scores are reported as accuracy (%).

Encoder Types	Fusion Mechanism	VQA-RAD	PMC-VQA	MedXQA	SLAKE	PathVQA	OMVQA	Average
Qwen ViT	-	69.4	57.4	24.6	81.0	58.0	83.5	62.3
Qwen ViT + DINOv2	Local Cross-Attn (Ours)	72.3	57.5	25.3	84.4	65.5	82.3	**64.5**
	Global Cross-Attn	69.5	55.0	24.2	82.6	61.0	84.1	62.7
	Mixture of Vision Experts	70.0	55.2	24.2	81.1	63.3	83.2	62.8
	Channel Concat	70.5	54.4	24.7	83.4	63.2	82.3	63.1

Supplementary Table 9 Ablation study on the choice of the specialist 2D encoder. We augment the base Qwen ViT with one additional encoder (DINOv2, ConvNext, or Medsiglip) and evaluate performance on two task categories: VQA and report generation. For VQA, we report accuracy (%). For report generation, we report RadGraph-F1 [8] for entities and entity-relations. The best average result is highlighted in **bold**.

Additional Encoder	VQA Benchmarks							Report Generation		
	VQA-RAD	PMC-VQA	MedXQA	SLAKE	PathVQA	OMVQA	Average	CheXpert-Plus	IU-XRAY	Average
DINOv2	72.3	57.5	25.3	84.2	65.5	82.3	64.5	22.6 / 20.2	37.7 / 35.5	**30.2 / 27.8**
ConvNext	72.1	57.1	24.8	83.5	64.7	82.7	64.1	21.8 / 20.0	37.9 / 35.9	29.9 / 27.9
Medsiglip	70.7	57.8	25.9	83.5	66.9	84.2	**64.8**	23.1 / 20.9	34.6 / 32.3	28.9 / 26.6

Supplementary Table 10 Ablation study on the choice of multiple additional encoders, the multi-encoder fusion strategy and fusion order. We augment the base Qwen ViT with two additional encoders and evaluate performance on two task categories: VQA and report generation. For VQA, we report accuracy (%). For report generation, we report RadGraph-F1 [8] for entities and entity-relations. The best average result is highlighted in **bold**.

Fusion Mode	Additional Encoder 1	Additional Encoder 2	VQA Benchmarks							Report Generation		
			VQA-RAD	PMC-VQA	MedXQA	SLAKE	PathVQA	OMVQA	Average	CheXpert-Plus	IU-XRAY	Average Report
Cascade	DINOv2	ConvNext	70.1	57.1	25.0	84.7	66.2	83.4	64.4	23.2 / 21.0	38.7 / 36.6	**31.0 / 28.8**
	ConvNext	DINOv2	70.3	54.8	26.2	84.2	66.5	84.2	64.4	20.2 / 17.9	39.2 / 37.4	29.7 / 27.6
	DINOv2	Medsiglip	72.7	58.0	25.2	84.7	66.5	84.4	**65.2**	23.1 / 20.7	36.6 / 34.1	29.8 / 27.4
	Medsiglip	DINOv2	69.0	53.6	25.5	84.6	68.0	85.9	64.4	22.4 / 20.0	35.1 / 32.5	28.8 / 26.3
Parallel	DINOv2	ConvNext	71.4	58.0	24.8	83.9	65.3	83.9	64.6	22.3 / 20.0	36.4 / 34.6	29.3 / 27.3

Supplementary Table 11 Ablation study on the necessity of the native 3D encoder during inference. We compare the performance of our full model against an ablated version where the 3D encoder is deactivated during inference, forcing the model to rely solely on sampled 2D slices for 3D understanding. For VQA tasks, we report accuracy (%). For report generation, we present Precision/Recall/F1 scores. Best results are highlighted in **bold**.

Regularization	AMOS	3D RAD		CT Rate		AMOS Report			CT Rate Report			Average	Average
	MCQ	Open	MCQ	Open	MCQ	Precision	Recall	F1	Precision	Recall	F1	VQA	Report
w/ 3D Encoder	77.6	72.6	30.0	85.3	63.6	18.1	9.67	11.0	29.3	18.4	20.3	**65.8**	**23.7 / 14.0 / 15.7**
w/o 3D Encoder	75.8	76.3	26.4	79.1	56.4	12.1	6.82	7.86	27.0	13.4	15.5	62.8	19.6 / 10.3 / 11.7

Supplementary Table 12 Ablation study on stochastic residual regularization for 3D tasks. The "Deterministic" setting uses standard residual connections, while the "Stochastic" setting applies stochastic residual regularization. For VQA tasks (AMOS, 3D RAD, CT Rate), we report accuracy (%). For report generation tasks, we present average Precision/Recall/F1 scores. Best average results are highlighted in **bold**.

Regularization	AMOS	3D RAD		CT Rate		AMOS Report			CT Rate Report			Average	Average
	MCQ	Open	MCQ	Open	MCQ	Precision	Recall	F1	Precision	Recall	F1	VQA	Report
Stochastic	77.6	72.6	30.0	85.3	63.6	18.1	9.67	11.0	29.3	18.4	20.3	**65.8**	**23.7 / 14.0 / 15.7**
Deterministic	78.8	67.9	30.4	85.9	61.2	20.3	8.40	10.6	26.5	18.2	19.6	64.8	23.4 / 13.3 / 15.1