

From Compound Figures to Composite Understanding: Developing a Multi-Modal LLM from Biomedical Literature with Medical Multiple-Image Benchmarking and Validation

Zhen Chen^{1†}, Yihang Fu^{1†}, Gabriel Madera^{1,2}, Mauro Giuffre¹, Serina Applebaum¹, Hyunjae Kim¹, Hua Xu¹,
Qingyu Chen^{1*}

¹*Department of Biomedical Informatics and Data Science, Yale School of Medicine, Yale University, New Haven, CT 06510, USA*

²*School of Medicine, University of Puerto Rico, San Juan, PR 00921, USA*

[†]*Contributed Equally*

***Corresponding author**

1 Abstract

Multi-modal large language models (MLLMs) have shown tremendous promise in advancing healthcare. However, most existing models remain confined to single-image understanding, which greatly limits their applicability in real-world clinical workflows. In practice, medical diagnosis and disease progression assessment often require synthesizing information across multiple images from different modalities or time points. The development of medical MLLMs capable of such multi-image understanding has been hindered by the lack of large-scale, high-quality annotated training data. To address this limitation, we propose a novel framework that leverages license-permissive compound images, widely available in biomedical literature, as a rich yet under-utilized data source for training medical MLLMs in multi-image analysis. Specifically, we design a five-stage, context-aware instruction generation paradigm underpinned by a *divide-and-conquer* strategy that systematically transforms compound figures and their accompanying expert text into high-quality training instructions. By decomposing the complex task of multi-image analysis into manageable sub-tasks, this paradigm empowers MLLMs to move beyond single-panel analysis and provide a composite understanding by learning the complex spatial, temporal, and cross-modal relationships inherent in these compound figures. By parsing over 237,000 compound figures and their contextual text for instruction generation, we develop M³LLM, a medical multi-image multi-modal large language model. For comprehensive benchmarking, we construct PMC-MI-Bench for composite understanding, manually validated by medical experts. Extensive experiments show that M³LLM significantly outperforms both general-purpose and specialized medical MLLMs across multi-image, single-image, text-only, and multi-choice scenarios. Notably, M³LLM exhibits strong generalization to real-world clinical settings, achieving superior performance on longitudinal chest X-ray analysis using the MIMIC dataset. This work establishes a scalable and efficient paradigm for developing next-generation medical MLLMs, capable of composite reasoning across complex multi-image scenarios, bridging the gap between biomedical literature and real-world clinical applications.

2 Introduction

Multi-modal large language models (MLLMs)^{1–3} combine natural language processing with multi-modal perception capabilities, and are capable of processing and reasoning across textual and visual data. In the general domain, MLLMs have demonstrated remarkable capability in understanding and integrating information across modalities, paving the way for their adaptation to specialized fields⁴. Preliminary results in healthcare applications have revealed promising potentials, particularly in processing clinical text, answering medical questions, and analyzing visual medical data^{5–12}. These advancements indicate the prospect of MLLMs to enhance diagnostic processes^{13,14}, streamline clinical decision-making¹⁵, and support medical education¹⁶. Despite these advances, a critical limitation persists: most existing MLLMs are primarily designed for single-image understanding, which significantly constrains their applicability in real-world medical scenarios involving complex multi-image, multi-modal data.

Compared to single-image tasks, multi-image tasks hold greater practical significance in real-world clinical workflows^{17–19}. For example, longitudinal monitoring requires comparing multiple images collected across different time points to track disease progression, while clinical diagnosis often integrates medical images from different modalities to provide a comprehensive understanding of a medical case^{20,21}. For instance, oncologists routinely analyze Magnetic Resonance Imaging (MRI) scans for tumor morphology, Positron Emission Tomography (PET) scans for metabolic activity, and histopathology slides collectively to formulate a comprehensive diagnostic picture²⁰, while cardiologists and neurologists similarly combine modalities like echocardiography, Computed Tomography (CT), and functional MRI to evaluate heart disease and brain disorders^{22,23}. These multiple-image scenarios, which constitute a substantial portion of clinical workflows, demand the composite understanding capabilities that synthesize information across multiple medical images. However, existing MLLMs^{5–12} fail to adequately address these, severely limiting the applicability and adoption. The scarcity of multiple-image MLLMs stems largely from a fundamental data challenge. Medical imaging data is inherently difficult to collect due to privacy and ethical constraints^{13,24}, and the complexity increases substantially for multiple-image datasets that require curated collections of related images across modalities and time points.

To overcome the critical bottleneck of data scarcity, we turn to compound figures from license-permissive biomedical literature, *i.e.*, multi-panel figures that integrate multiple sub-images within a single structured layout, where each panel typically represents a distinct but related aspect of the same medical case. Their significance lies not just in their public availability, but in their nature as a rich proxy for real-world clinical scenarios. As exemplified in Fig. 1, the compound figure exhibits the diverse inter-image complexities, including spatial arrangements highlighting anatomical correspondence, cross-modal combinations integrating complementary diagnostic information from CT and histopathology images, and temporal sequences showing disease evolution with postoperative examination. These complex relationships demand fundamentally more advanced reasoning capabilities compared to single-image understanding. As such, traditional instruction generation methods^{5,9}, primarily designed for single-image scenarios with straightforward image-text pairing, fail to capture the multi-dimensional dependencies inherent in compound figures, thus presenting a significant methodological challenge for composite understanding.

To address these challenges, we present the first systematic MLLM framework specifically designed for medical multiple-image understanding, by leveraging the compound figure data derived from biomedical literature. Our primary contribution is a novel five-stage, context-aware instruction generation paradigm underpinned by a *divide-and-conquer* strategy. This paradigm decomposes the complex challenge of compound figure understanding into a sequence of manageable, specialized tasks, ranging from medical knowledge complementation to visual perception enhancement, to transform raw compound figures and their associated textual content into

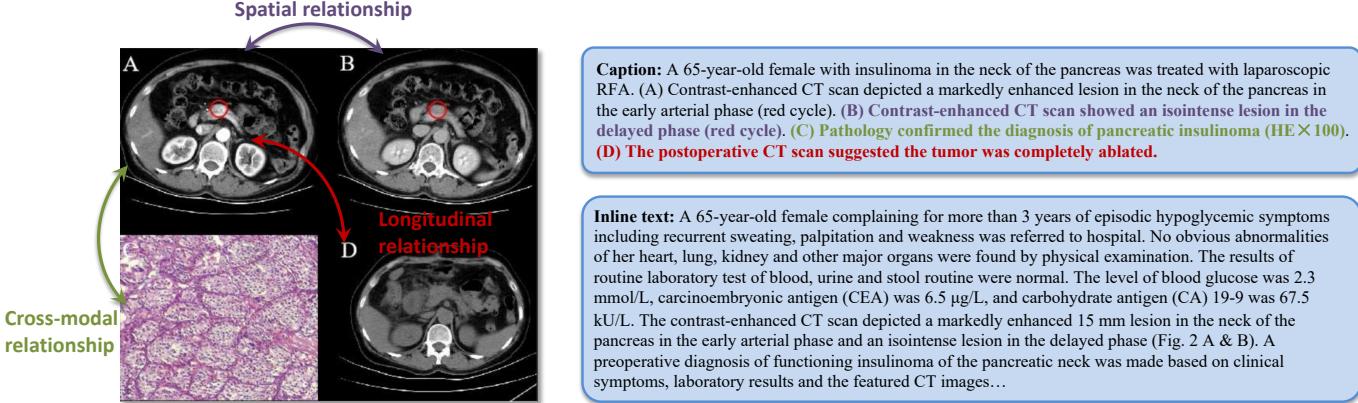


Figure 1: Illustration of a compound figure example in PMC literature. This example, derived from PMC7029651, features a compound figure composed of multiple sub-images. The example highlights longitudinal patient records with radiology and histopathology images for a case of insulinoma located in the neck of the pancreas. It integrates the accompanying image caption, which describes the visual contents, along with inline text from the manuscript that references the compound figure. To fully understand this medical compound figure, it is essential to comprehend the rich visual content and associated textual information. This includes analyzing the spatial, cross-modal, and longitudinal relationships of the sub-images, particularly concerning the first CT scan.

clinically rich and relevant training instructions. Unlike traditional methods^{5,6,9} that rely on simple image-text pairing, this paradigm constructs comprehensive learning scenarios that emulate real-world clinical reasoning processes, enabling MLLMs to effectively process and analyze the complex interrelationships inherent in medical compound figures. Then, using this paradigm on a large-scale dataset of over 237,000 compound figures, we develop and train M³LLM, a medical multi-image multi-modal large language model, to understand and reason over complex visual and textual information in clinical contexts. Furthermore, to facilitate rigorous benchmarking for this domain, we curate and release the PMC-MI-Bench, an expert-validated benchmark with comprehensive multi-image understanding tasks. Systematic evaluations demonstrate that M³LLM significantly outperforms state-of-the-art general-purpose and specialized medical MLLMs in multi-image, single-image, text-only, and multi-choice scenarios. Notably, the capabilities of M³LLM successfully generalize to clinical practice, as shown by its substantial improvements in a longitudinal patient analysis task using chest X-ray images from the MIMIC database^{18,19}, such as disease diagnosis and progression monitoring. To promote transparency and further advancements, we release the weights of our M³LLM, the training dataset, and our benchmark to the research community.

3 Results

We conduct extensive benchmarking and validation to assess the performance of our proposed M³LLM against state-of-the-art general-purpose MLLMs (*e.g.*, LLaVA-7B¹, LLaVA-NeXT-7B²⁵, QWen2.5-VL-7B², and InternVL3-8B³) and medical-specific ones (*e.g.*, LLaVA-Med-7B⁵, HuatuoGPT-Vision-7B⁹, Lingshu-7B²⁶, HealthGPT-14B¹², and MedGemma-27B²⁷). To ensure a comprehensive and diverse evaluation, our experiments assess performance across several key dimensions. We utilize a wide range of datasets, including our newly curated PMC-MI-Bench, public OmniMedVQA²⁸ and MMMU-Med²⁹ benchmarks, and a real-world clinical validation task using MIMIC longitudinal X-rays^{18,19}. The evaluation spans multiple task types, including multi-image VQA, single-image VQA, text-only QA, and multi-choice VQA. We employ a robust suite of evaluation metrics, ranging from accuracy for classification tasks to semantic metrics, including BLEU@4, ROUGE-L, BERTScore, and Semantic Textual Similarity (STS), and LLM-as-a-judge using GPT-4o for open-

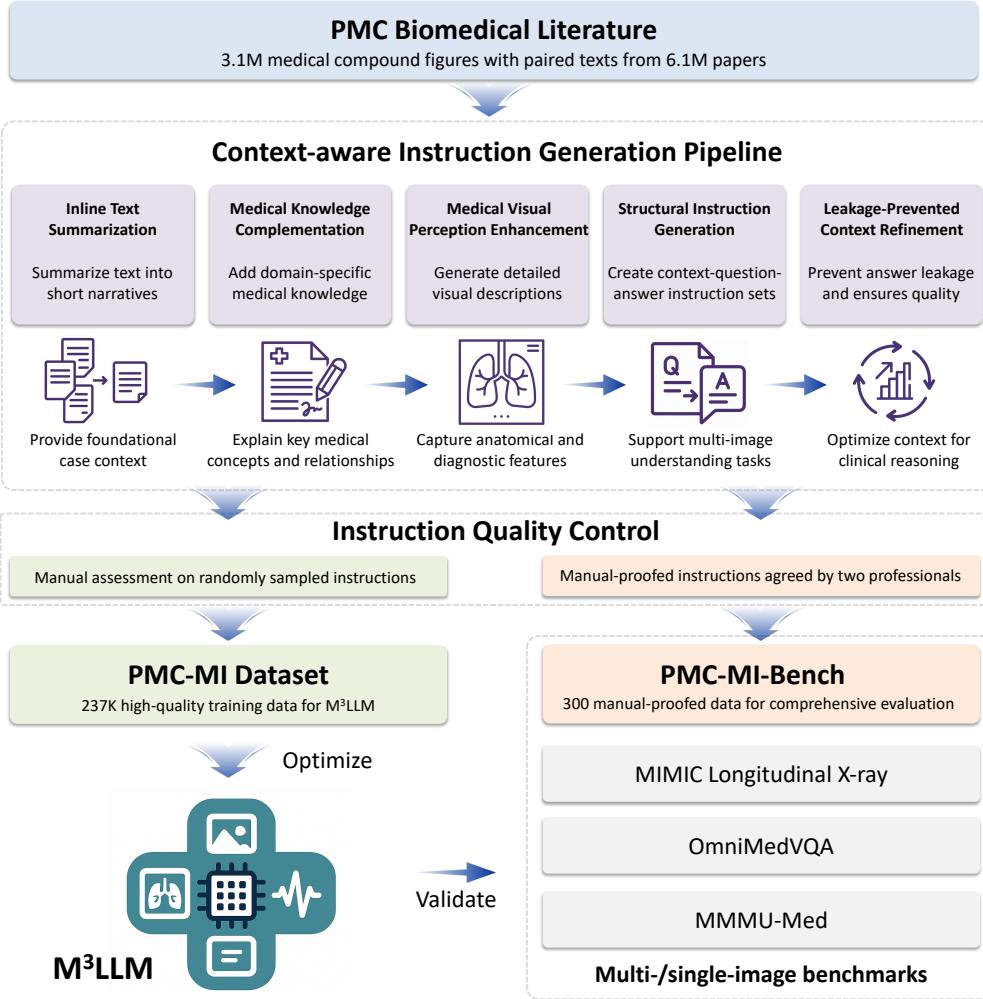


Figure 2: Overview of the study for medical compound figure understanding and clinical validation. The framework integrates PMC-derived compound figure data. Through a five-stage, context-aware instruction generation paradigm, the proposed M³LLM processes medical compound figures and paired texts. The core architecture of M³LLM includes a Vision Transformer (ViT), a connector module for visual-to-text alignment, and a large language model (LLM) for clinical reasoning. On this basis, the context-aware instruction tuning enables efficient and accurate multi-image comprehension. Extensive evaluation is conducted on the curated PMC-MI-Bench, public benchmarks, and MIMIC clinical cases.

ended generation. A holistic visualization of these comparisons in Fig. 3 concisely demonstrates that M³LLM achieves superior and well-rounded capabilities across this diverse suite of tasks. In this section, we detail these findings, followed by comprehensive ablation studies and a manual quality assessment of our generated PMC-MI dataset for training.

3.1 Performance Comparison on PMC-MI-Bench

We conduct comprehensive comparisons with state-of-the-art MLLMs across four instruction types of the curated PMC-MI-Bench, including the multi-image VQA, single-image VQA, text-only QA and multi-choice VQA. As elaborated in Table 1, 2, 3, and 4, our M³LLM achieves significant improvements across all QA settings, substantially outperforming existing MLLMs^{1–3, 5, 9, 12, 25–27}. These results demonstrate the effectiveness of our five-stage, context-aware instruction generation paradigm in creating clinically relevant training data that enables sophisticated medical reasoning with multi-image, single-image and text-only settings, as shown

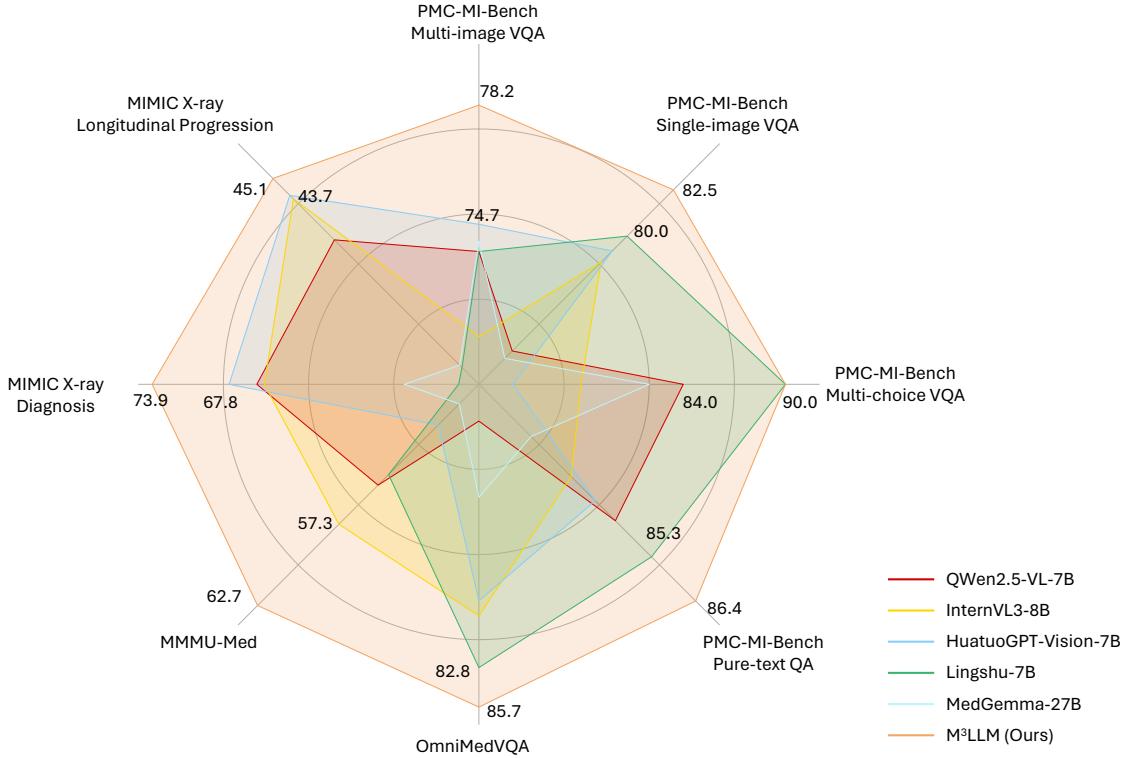


Figure 3: Comprehensive performance comparison of M³LLM against state-of-the-art MLLMs across eight tasks. This radar chart visualizes model capabilities across diverse medical tasks, utilizing Accuracy for classification and multi-choice tasks (*e.g.*, PMC-MI-Bench Multi-choice VQA, MIMIC X-ray Diagnosis, MIMIC X-ray Longitudinal Progression, OmniMedVQA, and MMMU-Med) and Semantic Textual Similarity (STS) for open-ended generation tasks (*e.g.*, PMC-MI-Bench Multi-image VQA, Single-image VQA, and Text-only QA). The best and second-best performances are marked on each axis, respectively. The expansive area covered by M³LLM (orange) visually demonstrates its superior and well-rounded capabilities across this diverse suite of tasks and metrics.

in Fig. 4.

In the multi-image VQA, our M³LLM demonstrates exceptional capability to synthesize information across multiple sub-images for comprehensive medical queries in Table 1, achieving 15.0 BLEU@4, 37.8 ROUGE-L, 70.1 BERTScore, and 78.2 Semantic Textual Similarity (STS) compared to the second-best performance of 9.8 BLEU@4 (Lingshu-7B²⁶), and 31.4 ROUGE-L, 66.9 BERTScore, and 74.7 STS (HuatuoGPT-Vision-7B⁹). The LLM-as-a-judge evaluation in Fig. 5 (a) further confirms superior quality across semantic reasoning tasks, with our M³LLM achieving 58.0% win and 17.7% tie compared to MedGemma-27B²⁷. This substantial improvement directly demonstrates the effectiveness of our context-aware instruction generation paradigm, which systematically integrates diverse medical findings across multiple imaging perspectives.

For the single-image VQA and text-only QA, our M³LLM also achieves superior performance in metrics of BLEU@4, ROUGE-L, Semantic Textual Similarity (STS), and the LLM-as-a-judge, compared with state-of-the-art general-purpose and medical MLLMs in Table 2 and 3 and Fig. 5 (b) and (c). It is noteworthy that LLaVA-Med-7B⁵ shows relatively strong performance on automatic text generation metrics like BLEU@4 and ROUGE-L. This can be attributed to its pretraining strategy, which is specifically optimized for medical caption generation. While this focus enhances its ability to produce linguistically aligned outputs, it does not translate as effectively to tasks demanding deeper clinical reasoning. This is highlighted by its significantly

Table 1: Comparison on the PMC-MI-Bench regarding the multi-image VQA.

Method	BLEU@4	ROUGE-L	BERTScore	STS
LLaVA-7B ¹	3.9	27.1	58.3	68.2
LLaVA-NeXT-7B ²⁵	4.5	26.5	59.3	68.7
QWen2.5-VL-7B ²	8.5	29.5	64.7	73.9
InternVL3-8B ³	3.8	22.5	55.7	71.4
LLaVA-Med-7B ⁵	5.8	23.7	58.8	63.0
HuatuoGPT-Vision-7B ⁹	9.1	31.4	66.9	74.7
Lingshu-7B ²⁶	9.8	30.4	66.8	73.9
HealthGPT-14B ¹²	9.3	30.8	66.3	73.7
MedGemma-27B ²⁷	3.4	26.7	62.5	74.2
M³LLM-8B (Ours)	15.0	37.8	70.1	78.2

Table 2: Comparison on the PMC-MI-Bench regarding the single-image VQA.

Method	BLEU@4	ROUGE-L	BERTScore	STS
LLaVA-7B ¹	2.3	22.0	55.3	70.1
LLaVA-NeXT-7B ²⁵	2.7	20.6	66.0	67.4
QWen2.5-VL-7B ²	3.4	23.5	55.5	73.8
InternVL3-8B ³	6.8	29.5	58.7	78.6
LLaVA-Med-7B ⁵	11.6	34.5	67.2	79.4
HuatuoGPT-Vision-7B ⁹	9.1	31.7	65.0	79.2
Lingshu-7B ²⁶	10.0	33.8	66.3	80.0
HealthGPT-14B ¹²	9.8	34.0	67.1	79.8
MedGemma-27B ²⁷	2.3	19.0	51.9	73.4
M³LLM-8B (Ours)	15.4	38.4	65.8	82.5

lower accuracy of 46.0% on the multi-choice VQA task (Table 4), where semantic correctness is paramount. In contrast, our M³LLM achieves the highest accuracy of 90.0% in Table 4, outperforming the medical MLLM MedGemma-27B²⁷ with the accuracy of 82.0% and HealthGPT-14B¹² with the accuracy of 88.0%.

Furthermore, we present the qualitative comparison of our M³LLM and MedGemma-27B²⁷ in Fig. 9, 10, 11, 12, 13 and 14 in terms of diverse tasks. These consistent performance advantages fully demonstrate that our M³LLM not only has significant advantages in multi-image scenarios, but also can learn effective medical knowledge from context-aware instruction tuning in the single-image VQA and text-only QA tasks, as well as the multi-choice VQA that existing research focuses on, thereby achieving better prediction answers on multiple tasks and metrics.

3.2 Performance Comparison on Public Medical Benchmarks

We further compare our M³LLM with state-of-the-art MLLMs on public single-image medical benchmarks, including OmniMedVQA²⁸ and MMMU-Med²⁹. Extensive evaluation validates that our comprehensive instruction generation paradigm yields substantial improvements beyond multi-image scenarios, confirming the positive transfer effects of systematic medical knowledge integration achieved through our five-stage, context-aware instruction generation paradigm.

Table 3: Comparison on the PMC-MI-Bench regarding the text-only QA.

Method	BLEU@4	ROUGE-L	BERTScore	STS
LLaVA-7B ¹	9.3	32.4	66.1	81.9
LLaVA-NeXT-7B ²⁵	9.8	35.2	68.3	83.2
QWen2.5-VL-7B ²	10.9	37.7	70.3	84.4
InternVL3-8B ³	9.1	35.4	68.3	83.3
LLaVA-Med-7B ⁵	11.5	36.3	69.6	83.5
HuatuoGPT-Vision-7B ⁹	11.3	37.9	69.1	83.9
Lingshu-7B ²⁶	11.0	38.8	68.3	85.3
HealthGPT-14B ¹²	11.8	39.0	69.0	84.2
MedGemma-27B ²⁷	7.1	31.6	65.3	82.3
M³LLM-8B (Ours)	13.0	38.5	73.4	86.4

Table 4: Comparison on the PMC-MI-Bench regarding the multi-choice VQA.

Method	Accuracy	F1	Recall	Precision
LLaVA-7B ¹	66.0	68.0	67.6	68.8
LLaVA-NeXT-7B ²⁵	70.0	70.0	74.4	72.3
QWen2.5-VL-7B ²	84.0	84.0	86.3	85.0
InternVL3-8B ³	78.0	78.0	80.9	79.2
LLaVA-Med-7B ⁵	46.0	47.9	45.0	51.1
HuatuoGPT-Vision-7B ⁹	74.0	69.7	71.9	75.6
Lingshu-7B ²⁶	90.0	90.8	90.8	91.8
HealthGPT-14B ¹²	88.0	87.5	89.7	87.4
MedGemma-27B ²⁷	82.0	81.0	80.5	83.1
M³LLM-8B (Ours)	90.0	89.9	91.2	89.8

On the OmniMedVQA benchmark (in Table 5), our M³LLM achieves 85.7% average accuracy across all imaging modalities, substantially outperforming both specialized medical MLLMs (*e.g.*, HuatuoGPT-Vision-7B⁹: 77.9%) and general-purpose models (InternVL3-8B³: 79.0%). The improvement over the existing general-purpose and specialized medical MLLMs demonstrates that the medical knowledge complementation (Stage 2) and medical visual perception enhancement (Stage 3) provide richer medical knowledge representation compared to conventional single-image focused training approaches. Modality-specific improvements are particularly notable in Computed Tomography (CT) (85.1% vs. 78.9% of the best baseline InternVL3-8B³) and Magnetic Resonance Imaging (MRI) (89.3% vs. 83.6% of HealthGPT-14B¹²), where our systematic instruction generation captures complex visual-clinical relationships essential for radiological diagnosis. X-Ray analysis shows consistent improvement (88.7% vs. 87.3% of InternVL3-8B³), while microscopy imaging demonstrates substantial gains (83.6% vs. 82.7% of Lingshu-7B²⁶), confirming that our multi-image instruction generation paradigm enhances understanding across diverse medical imaging modalities. It is noteworthy that the performance of our M³LLM is not the best in all modalities, particularly Ultrasound (US), Fundus Photography (FP), and Dermoscopy (Der). This directly correlates with the modality distribution of training data (Fig. 7), where these modalities are significantly underrepresented (*e.g.*, Ultrasound samples account for 2.3% and Fundus Photography samples account for 0.4%). The modest performance on these specific tasks highlights the impact of training data diversity and suggests a clear path for future improvement. In general, the exceptional performance of M³LLM in major radiological modalities secures its significant advantage in overall average accuracy, confirming the overall effectiveness of our methodology.

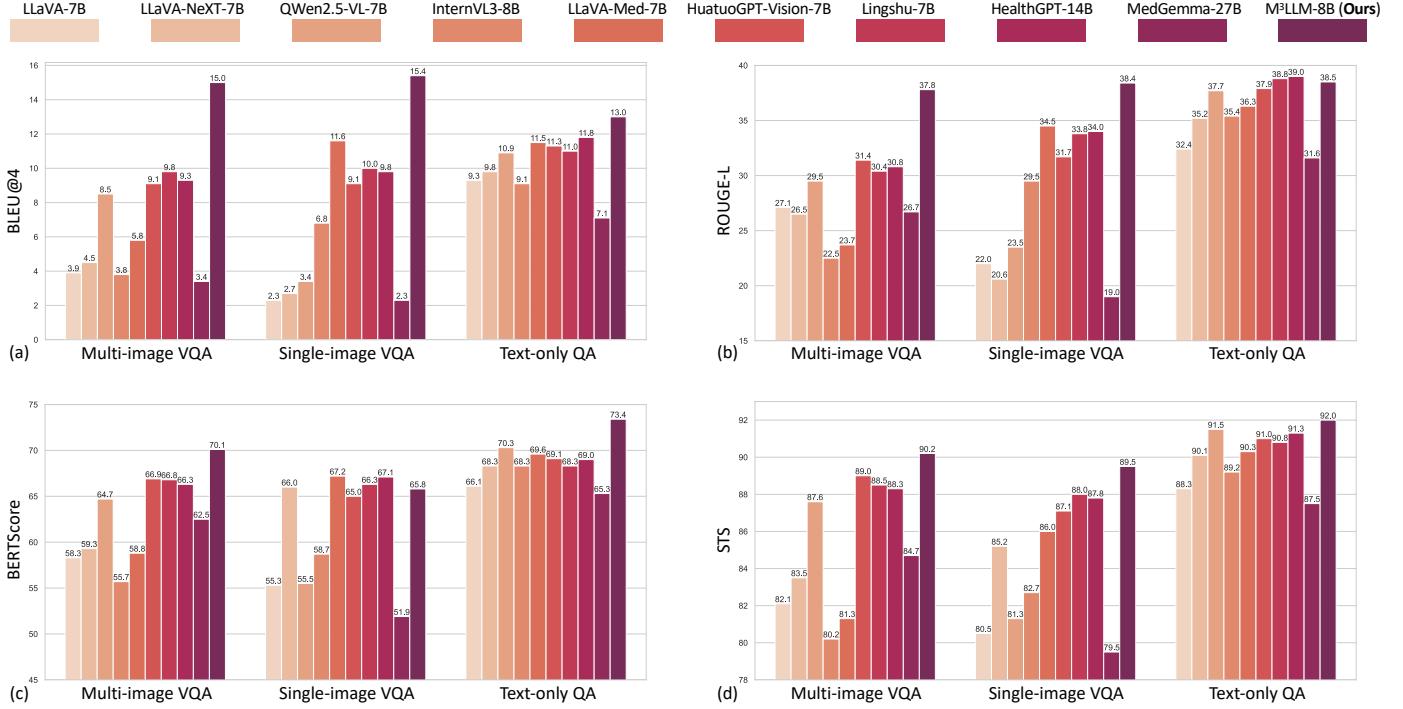


Figure 4: Performance comparison on open-ended text generation tasks within the PMC-MI-Bench. We compare our M³LLM against state-of-the-art general-purpose and specialized medical MLLMs across three question-answering task types: Multi-image VQA, Single-image VQA, and Text-only QA. Performance is evaluated using four standard text generation metrics: (a) BLEU@4, (b) ROUGE-L, (c) BERTScore, and (d) Semantic Textual Similarity (STS). The results consistently demonstrate the superior performance of M³LLM across all evaluated tasks and metrics compared to the baseline models.

Furthermore, MMMU Health & Medicine evaluation in Table 6 confirms consistent superiority across medical specialties, with our M³LLM achieving 62.7% average accuracy compared to the best baseline model (InternVL3-8B ³: 57.3%). In particular, the Basic Medical Science (BMS) performance shows particularly strong improvement (63.3% vs. 56.7% of QWen2.5-VL-7B ²), directly reflecting the clinical reasoning capabilities developed through our comprehensive instruction generation approach. The Clinical Medicine (CM) reaches 70.0% versus 66.7% of the baseline QWen2.5-VL-7B ² and InternVL3-8B ³, demonstrating the enhanced diagnostic reasoning that results from systematic medical knowledge integration. The Public Health (PH) (73.3% vs. 63.3% of InternVL3-8B ³) shows consistent improvements, confirming broad medical knowledge enhancement achieved through our proposed instruction generation paradigm.

3.3 Ablation Study on Context-aware Instruction Tuning

We investigate the performance of our M³LLM to validate the contributions of diverse instructions to substantial performance gains on the PMC-MI-Bench, OmniMedVQA ²⁸, and MMMU-Med ²⁹ datasets. Specifically, we conduct a detailed ablation study across four types of instructions, including the multi-image VQA, single-image VQA, multi-choice VQA, and text-only QA, by leveraging or removing one of these four instruction types in the training set. These experiments demonstrate the effects of our comprehensive instruction generation approach and identify the relative importance of each instruction category.

As illustrated in Table 7, compared to the baseline without instruction tuning (Line 1), we observe that different types of instructions bring significant improvements in Semantic Textual Similarity (STS). On the one

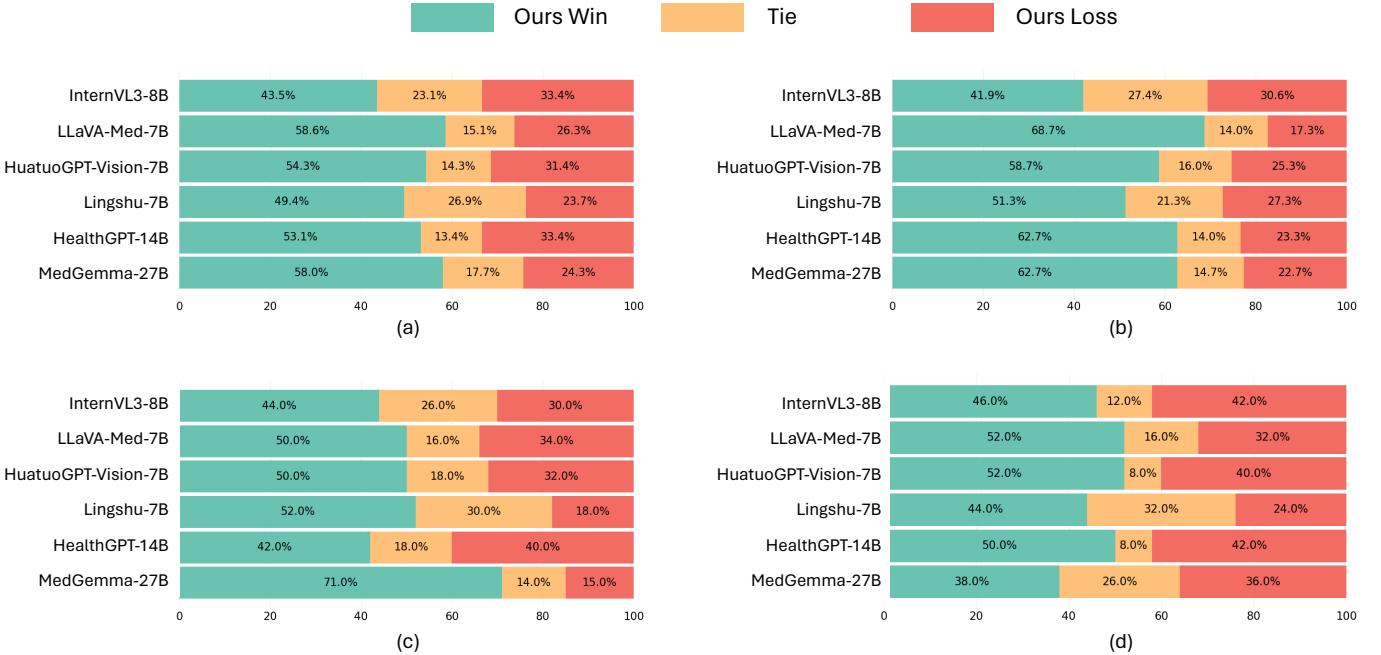


Figure 5: Comparison of LLM-as-a-judge assessment of our M^3 LLM against state-of-the-art MLLMs. We conduct the assessment using GPT-4o as a judge across multiple tasks on the PMC-MI-Bench, including (a) the overall performance, (b) the multi-image VQA, (c) the single-image VQA, and (d) the text-only QA.

hand, by adding each type of PMC-MI instructions to the training set, the MLLM can be improved on the same type of samples on the PMC-MI-Bench. In particular, the instructions of multi-image VQA bring a 3.8% STS performance improvement on the same type of the PMC-MI-Bench (Line 2), the instructions of single-image VQA bring a 1.0% STS performance improvement (Line 3), and the instructions for multi-choice VQA (Line 4) and text-only QA (Line 5) bring a 2.0% STS performance and 4.0% accuracy improvement, respectively. On the other hand, these instructions further improve the performance of other tasks, for example, the multi-image VQA instructions can improve the performance of single-image VQA with a STS increase of 1.9%. This confirms that the generated instructions can provide sufficient medical knowledge to facilitate the model to better complete various types of downstream tasks. Moreover, by leveraging these diverse types of training instructions, the tuned models reveal an impressive advantage over the baseline model on the public single-image benchmarks, with an accuracy increase from 79.0% to 85.7% for OmniMedVQA²⁸ and from 59.3% to 62.7% for MMMU-Med²⁹. These results demonstrate the effectiveness of our designs in the training instruction dataset.

To validate the positive transfer effects across different types of training instructions, we further perform the instruction tuning by excluding one type of instruction samples from the training set (Line 6-9 in Table 7). By comparing the M^3 LLM with all types involved (Line 10), the ablative models confirm that these training data can further promote the performance on different tasks based on other training data, including four tasks on PMC-MI-Bench, as well as OmniMedVQA²⁸ and MMMU-Med²⁹. In particular, single-image VQA can further improve the performance of the model on MMMU-Med by 2.7%. It is worth noting that in the training instructions, multi-choice VQA has a significant performance gain on the public OmniMedVQA and MMMU-Med benchmarks, which shows that the positive transfer of medical knowledge of the same task type is effective. Finally, by comprehensively utilizing the four types of instruction samples in the training set, our M^3 LLM achieves the best performance among these downstream tasks, achieving an increase of 6.8%, 3.9%, 3.1%, and 8.0% in multi-image VQA, single-image VQA, multi-choice VQA, and text-only QA over the baseline (Line 1) on the PMC-MI-Bench, respectively.

Table 5: Comparison of state-of-the-art MLLMs on public OmniMedVQA in terms of different modalities. Specifically, CT denotes Computed Tomography, FP denotes Fundus Photography, MR denotes Magnetic Resonance Imaging, OCT denotes Optical Coherence Tomography, Der denotes Dermoscopy, Mic denotes Microscopy Images, US denotes Ultrasound.

Method	CT	FP	MR	OCT	Der	Mic	X-Ray	US	Avg
LLaVA-7B ¹	30.7	26.1	23.9	24.8	30.7	26.4	23.0	25.1	26.0
LLaVA-NeXT-7B ²⁵	31.7	27.0	24.6	25.7	30.2	26.6	25.8	25.5	27.1
QWen2.5-VL-7B ²	62.5	70.8	66.4	66.2	68.0	70.9	76.8	35.6	64.7
InternVL3-8B ³	78.9	87.5	80.4	77.2	81.8	82.2	87.3	76.4	79.0
LLaVA-Med-7B ⁵	38.7	48.3	39.5	45.7	58.8	49.3	43.4	48.1	46.5
HuatuoGPT-Vision-7B ⁹	70.0	84.2	72.1	85.7	72.5	75.7	81.9	81.2	77.9
Lingshu-7B ²⁶	77.2	88.9	83.3	89.9	83.2	82.7	86.4	80.6	82.8
HealthGPT-14B ¹²	70.3	82.5	83.6	88.0	69.0	72.7	83.7	56.7	75.8
MedGemma-27B ²⁷	75.5	80.8	66.1	76.3	74.0	66.1	78.5	55.8	70.3
M³LLM-8B (Ours)	85.1	85.5	89.3	90.2	79.8	83.6	88.7	78.2	85.7

Table 6: Comparison of state-of-the-art MLLMs on public MMMU-Med in terms of different health and medicine tracks. Specifically, BMS for Basic Medical Science, CM for Clinical Medicine, DLM for Diagnostics and Laboratory Medicine, P for Pharmacy, and PH for Public Health.

Method	BMS	CM	DLM	P	PH	Avg
LLaVA-7B ¹	23.3	20.0	26.7	23.3	23.3	23.3
LLaVA-NeXT-7B ²⁵	20.0	20.0	26.7	33.3	23.3	24.7
QWen2.5-VL-7B ²	56.7	66.7	36.7	56.7	56.7	54.7
InternVL3-8B ³	53.3	66.7	43.3	60.0	63.3	57.3
LLaVA-Med-7B ⁵	33.3	40.0	26.7	40.0	53.3	38.7
HuatuoGPT-Vision-7B ⁹	53.3	70.0	46.7	43.3	40.0	50.7
Lingshu-7B ²⁶	56.7	53.3	60.0	46.7	53.3	54.0
HealthGPT-14B ¹²	50.0	50.0	43.3	46.7	50.0	48.0
MedGemma-27B ²⁷	46.7	53.3	50.0	53.3	43.3	49.3
M³LLM-8B (Ours)	63.3	70.0	53.3	53.3	73.3	62.7

Furthermore, we dive into the multi-image VQA instructions, comprising the VQA regarding the multi-subimage, single-subimage, and the spatial relationship, on the PMC-MI-Bench. Specifically, we implement the ablation study with single-image VQA, multi-choice VQA, and text-only QA of PMC-MI available, by adding or removing one type of the multi-image VQA instructions. The details of these multi-image VQA instructions are presented in Stage 4: Context-Question-Answer Instruction Generation of Section 6.2. From Line 1 to Line 4 in Table 8, we observe that these three types of multi-image VQA instructions contribute to the performance of multi-image VQA samples in PMC-MI-Bench, with the performance increase of 1.7%, 0.8%, and 0.2%, respectively. When these instructions are combined, the M³LLM results in the remarkable performance of 78.2%, 82.5%, 86.4% and 90.0% on the multi-image, single-image, text-only, and multi-choice types of PMC-MI-Bench. These results confirm the impact of the diverse types of instructions on the M³LLM to achieve superior performance on multi-image, single-image, multi-choice, and text-only tasks of PMC-MI-Bench, OmniMedVQA, and MMMU-Med benchmarks.

Table 7: Ablation study of the training data settings on PMC-MI-Bench, OmniMedVQA, and MMMU-Med. Performance is reported using Semantic Textual Similarity (STS) for free-text tasks and Accuracy for the multi-choice task.

Training Data Setting				PMC-MI-Bench				OmniMedVQA	MMMU-Med
Multi-img	Single-img	Text-only	Multi-choice	Multi-img	Single-img	Text-only	Multi-choice		
1				71.4	78.6	83.3	82.0	79.0	59.3
2	✓			75.2	80.5	84.0	86.0	81.4	61.3
3		✓		74.5	79.6	84.1	84.0	82.4	60.0
4			✓	73.9	79.5	85.3	84.0	81.9	60.7
5			✓	73.4	80.5	83.3	86.0	80.2	61.3
6	✓	✓	✓	77.1	80.2	84.9	84.0	84.1	60.7
7	✓	✓		77.3	81.1	84.7	88.0	83.9	61.3
8	✓		✓	75.8	80.7	85.3	86.0	85.2	60.0
9		✓	✓	76.4	81.4	85.4	86.0	84.6	60.0
10	✓	✓	✓	78.2	82.5	86.4	90.0	85.7	62.7

Table 8: Ablation study of different multi-image training settings on the PMC-MI-Bench. Performance is reported using Semantic Textual Similarity (STS) for free-text tasks and Accuracy for the multi-choice task.

Training Data Setting of Multi-image VQA			PMC-MI-Bench			
Multi-subimage	Single-subimage	Spatial relationship	Multi-img	Single-img	Text-only	Multi-choice
1			74.5	80.2	85.3	86.0
2	✓		76.2	80.8	85.3	86.0
3		✓	75.3	81.4	85.8	86.0
4			74.7	80.1	85.6	84.0
5	✓	✓	77.6	82.1	85.7	88.0
6	✓		77.1	80.9	84.9	84.0
7		✓	75.9	81.7	85.5	86.0
8	✓	✓	78.2	82.5	86.4	90.0

3.4 Clinical Validation on Longitudinal Medical Imaging

To evaluate the performance of our M³LLM in clinical scenarios requiring longitudinal reasoning, we conduct experiments using the MIMIC chest X-ray dataset, as shown in Fig. 6. The dataset is divided at the patient level into a training set and a validation set in a 1:1 ratio, ensuring no patient overlap between the two. The training set is used for fine-tuning all MLLMs, while the validation set is employed for performance evaluation. Each data record in the dataset contains two chest X-ray images from different examinations of the same patient. For the disease diagnosis task, the first examination image is used as input to predict whether the patient has a specific disease. For the progression prediction task, both X-ray images are used to determine the progression of a specific disease, categorizing it as improvement, deterioration, or stability. Furthermore, we calculate the accuracy for cases where both disease diagnosis and progression prediction are correct, offering a comprehensive effectiveness measure of the MLLMs in longitudinal reasoning.

The results, summarized in Fig. 6 (a) and (b), demonstrate that M³LLM outperforms all compared MLLMs across both tasks. Specifically, M³LLM achieves 73.9% accuracy in disease diagnosis, and 45.1% in longitudinal progression prediction. These results highlight the model’s superior ability in understanding both single-image and longitudinal relationships in medical imaging. Compared to the second-best model, HuatuoGPT-Vision-7B⁹, M³LLM achieves a 6.1% improvement in disease diagnosis accuracy and a 1.4% improvement in progression prediction accuracy. Notably, M³LLM’s superior performance on these two tasks further under-

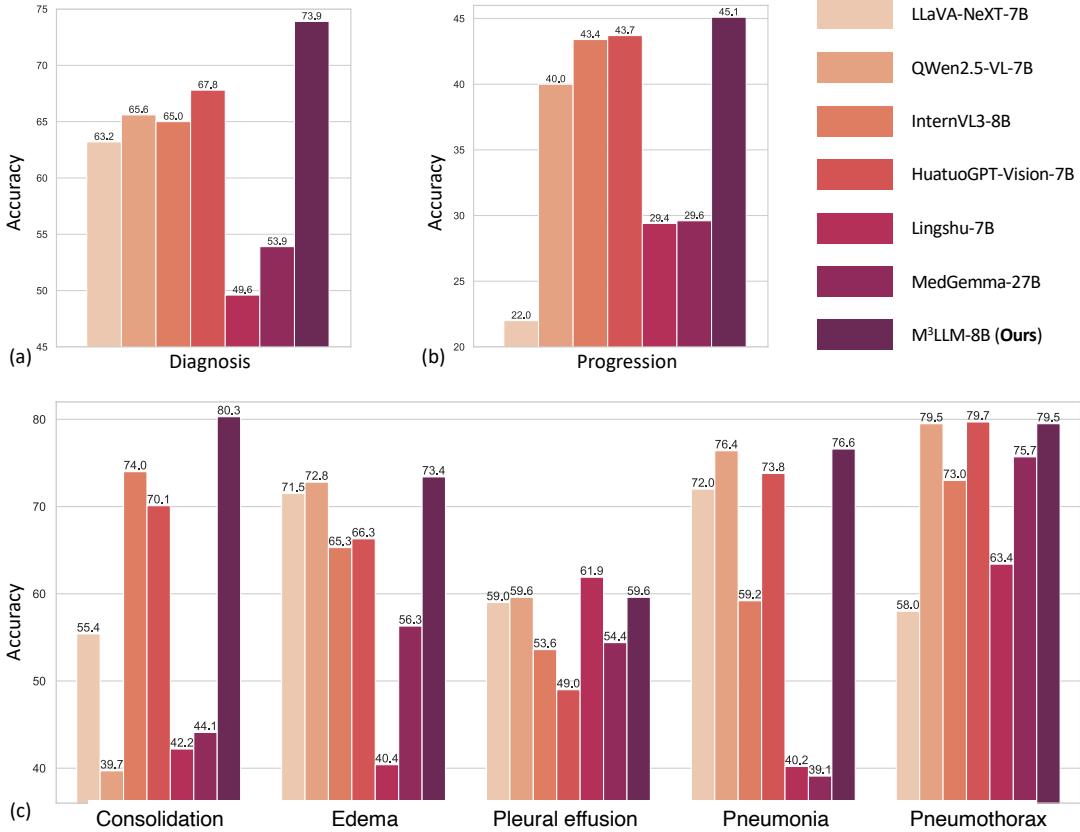


Figure 6: Performance comparison on the MIMIC chest X-ray longitudinal benchmark. We compare the performance of our M³LLM against several state-of-the-art MLLMs on clinical validation tasks requiring longitudinal reasoning. (a) The accuracy comparison for the disease diagnosis task based on the first examination image. (b) The accuracy comparison for the disease progression prediction task (*i.e.*, improvement, deterioration, or stability) using both examination images. (c) Detailed disease diagnosis accuracy comparison across five specific conditions: Consolidation, Edema, Pleural effusion, Pneumonia, and Pneumothorax. The results highlight the superior performance of M³LLM in both diagnosis and progression prediction, as well as its strong generalization across various disease types.

scores its robustness and clinical applicability. To assess the model’s ability to handle different disease types, we analyze its performance across five conditions: consolidation, edema, pleural effusion, pneumonia, and pneumothorax, as presented in Fig. 6 (c). M³LLM achieves the highest accuracy in three out of five disease categories, including consolidation (80.3%), edema (73.4%), and pneumonia (76.6%), and competitive performance on pneumothorax (79.5% vs. the best 79.7%) and pleural effusion (59.6% vs. the best 61.9%). This analysis confirms the model’s capacity to generalize across various diseases and accurately recognize subtle radiological changes associated with each condition. The substantial improvements in disease diagnosis, particularly for diseases like consolidation and pneumonia, underscore the model’s ability to provide timely and precise predictions, which are critical for effective clinical intervention.

Despite the challenges inherent in progression prediction, such as the subtlety and complexity of longitudinal changes, M³LLM consistently outperforms the state-of-the-art MLLMs, reflecting the strength of its spatio-temporal reasoning capabilities. The ability to detect nuanced changes in disease progression, whether indicative of improvement, deterioration, or stability, is vital for longitudinal patient care. The accuracy of MLLMs on the chest X-ray dataset, which requires simultaneous success in both disease diagnosis and progression prediction, further underscores its comprehensive understanding of longitudinal medical imaging. The superior

performance of M³LLM can be attributed to its systematic instruction generation paradigm, which trains the model to effectively capture progression patterns and spatial-temporal relationships. These results demonstrate the potential of M³LLM to improve clinical decision-making in real-world longitudinal patient care.

3.5 Training Data Scale Analysis

We further conduct the systematic evaluation of M³LLM performance across varying training data scales on PMC-MI-Bench, as well as public OmniMedVQA²⁸ and MMMU-Med²⁹ benchmarks. These results in Table 9 demonstrate the relationship between instruction data scale and medical multi-image understanding capabilities, confirming the quality of our instruction generation paradigm and providing insights for optimal deployment strategies. Specifically, the InternVL-3-8B³ without context-aware instruction tuning serves as the baseline and achieves the accuracy of 79.0% and 59.3% on the OmniMedVQA and MMMU-Med, respectively. On this basis, we increase the ratio of the training set from 0% to 100%. We observe that the M³LLM obtains an accuracy increase of 3.1% and 1.4% on OmniMedVQA and MMMU-Med with only 5% of the PMC-MI training set. Further increasing the number of training samples will continue to improve model performance, but the rate of increase will slow significantly. For example, when the number of training samples reaches 30%, the performance of M³LLM reaches 83.6% on OmniMedVQA. As we continue to increase the number of training samples, the performance increase remains relatively stable until the full training dataset achieves 85.7% performance. These experimental results show that our instruction data can effectively bring medical knowledge to M³LLM. The effect is obvious on a small amount of data, while more data leads to better performance on downstream tasks.

3.6 Training Data Quality Assessment

We further conduct the professional medical assessment on the randomly sampled instructions of the PMC-MI dataset to confirm high-quality instruction generation across all stages, as well as substantial inter-annotator agreement supporting reliable quality assessment. In our implementation, we randomly select 140 training samples from five stages, where five samples are selected from each of the six types in each stage, except that spatial relationship and multi-choice VQA instructions do not need to go through the fifth stage of context improvement. Each training sample is evaluated by two medical professionals from the perspectives of correctness, completeness, and clarity. Each item is scored on a 1, 3, or 5 basis, where 5 means the entire sample, including context, question, and answer, is satisfactory, with no error or hallucination, 3 means the sample is generally satisfactory, with one or two minor flaws and no significant error, and 1 means the sample is unsatisfactory, with significant errors. As illustrated in Table 10, our training samples have generally received satisfactory evaluation results, with average scores for correctness, completeness, and clarity exceeding 4 across different stages. During the data preparation phase, our paradigm performs particularly well in the second stage of Medical Knowledge Complementation, achieving an average score of 4.8, indicating that it accurately provides effective and relevant medical knowledge. In contrast, the third stage of Medical Visual Perception Enhancement, proved to be more challenging, with a score of 4.4. This highlights the limitations of current medical MLLMs. Notably, the instructions generated in the fourth stage achieved an average score of 4.6, which was further improved to 4.9 after the fifth stage of Context Refinement. This improvement clearly demonstrates that our pipeline is effective in enhancing the context within the instructions, thereby further improving the overall quality of the instructions.

Inter-annotator agreement analysis. To rigorously quantify the consensus of three medical professionals, we conduct the pilot test to assess the instructions for correctness, completeness, and clarity, and calculate the Intraclass Correlation Coefficient (ICC) among their assessments on 78 randomly-sampled cases, and the overall

Table 9: Analysis of training data scale on PMC-MI-Bench, OmniMedVQA and MMMU-Med.

Training data ratio	PMC-MI-Bench		OmniMedVQA	MMMU-Med
	Open-ended tasks	Multi-choice VQA		
0%	73.2	82.0	79.0	59.3
5%	77.4	86.0	82.1	60.7
10%	77.9	86.0	82.9	61.3
20%	78.5	84.0	83.3	59.3
30%	78.3	86.0	83.8	60.0
50%	79.3	88.0	84.4	60.7
75%	80.1	88.0	84.9	61.3
100%	80.6	90.0	85.7	62.7

Table 10: Manual assessment of training data quality across five stages in the instruction generation paradigm.

Stage	Correctness	Completeness	Clarity	Average
Stage 1	4.8 ± 0.2	4.4 ± 0.4	4.9 ± 0.1	4.7 ± 0.2
Stage 2	4.9 ± 0.1	4.8 ± 0.3	5.0 ± 0.0	4.9 ± 0.1
Stage 3	4.3 ± 0.1	4.5 ± 0.3	4.5 ± 0.0	4.4 ± 0.1
Stage 4	4.6 ± 0.1	4.7 ± 0.2	4.8 ± 0.2	4.7 ± 0.1
Stage 5	4.8 ± 0.1	4.9 ± 0.1	5.0 ± 0.1	4.9 ± 0.1

ICC of 0.816 for all rated aspects indicates excellent reliability. Dimension-specific analyses further confirmed this strong agreement, with an ICC of 0.867 for correctness, 0.751 for completeness, and 0.720 for clarity. This robust statistical consensus is underscored by a high rate of exact agreement (74.8%) and near-perfect agreement within one-score difference (98.3%), confirming a consistent quality assessment across the independent medical professionals. A detailed analysis of the rare disagreements reveals that conflicts primarily involve nuanced edge cases in clinical interpretation rather than fundamental accuracy issues. Most disagreements concern completeness assessments where evaluators differed on the optimal level of detail required for a specific clinical scenario. Correctness disagreements typically arise in cases involving rare pathological conditions or emerging diagnostic criteria, while clarity disagreements focus on the accessibility of technical terminology for different medical specialties.

PMC-MI dataset characteristics. To ensure comprehensive evaluation across diverse medical scenarios, we analyze the M³LLM with dataset characteristics by randomly sampling 1,000 cases from the PMC-MI dataset. Each case is analyzed using GPT-4o to extract key textual information, including image modality and the anatomical system involved.

The dataset encompasses a wide variety of imaging modalities, as depicted in Fig. 7 (a). The most represented categories are microscopy (24.2%) and histopathology (20.9%), reflecting the critical role of detailed cellular and tissue-level imaging in medical diagnostics. Multimodal composite images, which require integration across multiple imaging types, make up 14.2% of the dataset, highlighting the increasing complexity of modern medical imaging scenarios. Advanced radiological modalities, including MRI (10.7%), CT (6.4%), and PET-CT (0.9%), ensure sufficient coverage of cross-sectional imaging commonly used in clinical practice. Other modalities, such as ultrasound (2.3%), X-ray (2.3%), and clinical photography (2.2%), provide additional diversity, ensuring the dataset captures a broad spectrum of real-world medical imaging scenarios.

For the anatomical systems, neurological imaging accounts for the largest proportion (23.4%) as shown in Fig. 7 (b), reflecting the high prevalence of brain and nervous system studies in clinical and research settings. Musculoskeletal and cardiovascular systems are also well-represented, contributing 11.0% and 10.3%, respectively, while gastrointestinal (8.9%) and respiratory (5.3%) systems further ensure diversity. Ophthalmology (5.4%), reproductive systems (5.5%), and dermatology (2.3%) are included as specialized areas, ensuring the evaluation extends to less common but clinically significant domains.

The diversity of the PMC-MI dataset, both in imaging modalities and anatomical systems, ensures that the proposed M³LLM is equipped to handle a wide range of real-world medical applications. It enables the model to excel in single- and multi-image scenarios, integrate information across varied imaging types, and effectively reason about complex longitudinal changes. The inclusion of medical multi-modal multiple-image further guides the model to synthesize information from diverse sources, a critical requirement for addressing complex diagnostic challenges. Together, these attributes make the dataset an invaluable resource for driving advancements in medical image understanding and improving the robustness of AI models in clinical practice.

4 Discussion

This study introduces M³LLM, a multi-modal LLM tailored to address the unique challenges of medical multi-image understanding. Through its five-stage, context-aware instruction generation paradigm, M³LLM demonstrates superior performance in multi-image understanding, single-image understanding, and longitudinal clinical analysis. By leveraging the training on over 237,000 compound figures, M³LLM bridges the gap between complex visual biomedical research and real-world clinical applications.

The M³LLM represents a significant advancement in handling medical multi-image data, a crucial yet underexplored aspect of medical AI. It excels by integrating information across multiple sub-images to capture complex spatial, contextual, and diagnostic relationships, unlike existing models (*e.g.*, LLaVA-Med⁵, Med-Flamingo⁶, HuatuoGPT-Vision⁹ and HealthGPT¹²) primarily focus on single-image tasks³⁰⁻³². This superior capability is evident in its performance on the PMC-MI-Bench for multi-image understanding, where M³LLM achieves a STS of 78.2, significantly outperforming HealthGPT (73.7) and LLaVA-Med (63.0) in Table 1. This effectiveness stems directly from our five-stage, context-aware instruction generation paradigm, whose core distinction lies in the **explicit modeling and learning** of composite reasoning. While advanced MLLMs like InternVL3³ and QWen2.5-VL² possess the architectural capacity for multi-image input, their training lacks *clinically meaningful* fine-tuning specific to medical scenarios and, crucially, does not *systematically enforce* the synthesis of information across images. They may learn implicit associations when presented with multiple images, but they are not *explicitly taught* to analyze the spatial, temporal, or cross-modal relationships that define complex medical cases. Our paradigm directly addresses this gap by creating tasks that require the model to compare sub-images, track changes, or integrate findings from different modalities (*e.g.*, CT and histopathology in Fig. 1). Through this process, we are fundamentally shaping the model’s reasoning capabilities to handle multi-dimensional dependencies explicitly, rather than just fine-tuning for medical content. This methodological advantage is the key driver of M³LLM’s superior performance in complex medical scenarios and ensures its alignment with real-world diagnostic workflows.

Complementing its explicit modeling capabilities, another key strength of M³LLM with multi-image input lies in its ability to perform longitudinal analysis, which is critical for tracking disease progression over time. On the basis of the MIMIC chest X-ray longitudinal dataset, M³LLM demonstrated substantial improvements in predicting disease progression and integrating temporal relationships across imaging studies. For example, M³LLM achieves higher accuracy in identifying both current pathological conditions and future disease trajec-

tories, outperforming baseline models such as HuatuoGPT-Vision⁹ and InternVL3³, which are limited in their ability to integrate sequential data. This capability reflects the benefits of M³LLM’s context-aware training design, which specifically incorporates spatial and temporal reasoning tasks. By enabling dynamic analysis of longitudinal imaging data, M³LLM provides a potential solution for chronic disease management, prognosis, and follow-ups.

Beyond these core advantages in multi-image reasoning, M³LLM further demonstrates the strength across diverse datasets, tasks, and input settings. By training on a large-scale dataset derived from the PubMed Central biomedical literature, M³LLM effectively leverages domain-specific knowledge to handle a wide range of benchmarks from various sources, including MIMIC^{18,19}, OmniMedVQA²⁸, and MMMU-Med²⁹, covering tasks that span from radiology and pathology to clinical question answering. This showcases how biomedical knowledge embedded in PubMed Central data can be utilized to solve problems across distinct domains. Moreover, M³LLM also exhibits strong performance across diverse task types, including single-image VQA, text-only QA, and multi-choice VQA. On diverse benchmarks, M³LLM consistently achieved state-of-the-art results, demonstrating its flexibility in adapting to the requirements of different task formats. Unlike existing MLLMs that often struggle to generalize beyond single-image VQA, the comprehensive instruction tuning pipeline of M³LLM allows it to handle diverse settings effectively. These results highlight M³LLM’s ability to generalize its reasoning capabilities from biomedical literature to a variety of clinical scenarios and task types.

On the basis of technical achievements, the clinical implications of M³LLM underscore the potential to transform real-world healthcare workflows. In practical settings, M³LLM can assist clinicians in synthesizing complex findings from multi-panel imaging studies, such as integrating MRI, CT, and histopathology images to form a unified diagnostic conclusion. This capability reduces the cognitive burden on radiologists and supports faster, more accurate decision-making, particularly in time-sensitive scenarios like emergency care. Additionally, M³LLM’s reliance on routine clinical images and textual data makes it a cost-effective and accessible solution for low-resource healthcare settings, where access to advanced diagnostic tools is often limited. M³LLM is capable of processing free-text health records and dynamic imaging data, which positions it as a practical tool for diverse healthcare environments.

Despite its strengths, this study has limitations that highlight opportunities for further research. First, the performance of M³LLM relies on the diversity and scale of its training data. In scenarios where training data for specific tasks or rare clinical conditions is limited, the model’s performance may degrade accordingly, *e.g.*, the underexplored fundus photography and ultrasound imaging as indicated in Table 5. Addressing this limitation will require curating more diverse datasets, particularly focusing on underrepresented populations, rare diseases, and specialized medical scenarios to ensure robust generalization across all use cases. Second, while M³LLM focuses on visual and textual data, integrating additional clinical modalities such as laboratory test results, patient histories, and treatment response data could further enhance its diagnostic capabilities and provide a more holistic understanding of patient conditions. Third, while traditional metrics like accuracy, BLEU, and ROUGE-L provide useful insights into performance, they may not fully capture the nuances of clinical reasoning and decision-making in clinical practice. Developing domain-specific evaluation benchmarks, validated by medical professionals, will be essential for accurately assessing the model’s utility in real-world clinical workflows. By addressing these limitations, future research can further expand the applicability and impact of M³LLM in diverse medical contexts.

In conclusion, M³LLM represents a significant step forward in medical AI, offering a robust solution for understanding and reasoning over medical multi-images. By addressing the challenges of multi-image analysis and integrating temporal reasoning, M³LLM sets a new benchmark for multimodal medical AI systems. Its scalable, cost-effective, and clinically relevant framework has the potential to transform real-world diagnostic

workflows and improve patient care. Future research should focus on expanding its applications across diverse clinical contexts and incorporating additional data modalities to further bridge the gap between biomedical research and clinical practice.

5 Comprehensive Datasets and Benchmarks

Current evaluation benchmarks for medical MLLMs predominantly focus on single-image scenarios, which inadequately reflect the complexity of real-world clinical practice where physicians integrate multi-modal imaging studies for comprehensive diagnosis. To address this critical gap, we introduce the PMC Multiple Image (PMC-MI) dataset, a large-scale training dataset for developing robust medical multi-image MLLMs, alongside the PMC-MI-Bench, a novel benchmark specifically designed for evaluating medical multi-image understanding capabilities.

PMC-MI Dataset. The PMC-MI dataset comprises instruction sets derived from 237,137 medical compound figures harvested from PubMed literature, each paired with rich contextual information. Each compound figure contains an average of 4.97 sub-figures, representing different aspects of the same medical case and mirroring the multi-perspective analysis required in clinical practice. Furthermore, the average resolution of the compound figure is 705.1 pixels in width and 599.8 pixels in height. The accompanying textual context is substantial, with compound figure captions averaging 102.5 words, associated inline text averaging 188.4 words. As demonstrated in Fig. 7, the dataset encompasses diverse medical specialties, including radiology, histopathology, dermatology, and ophthalmology, with imaging modalities ranging from MRI and CT scans to microscopic and photographic documentation. Through the sophisticated instruction generation paradigm (detailed in Section 6.2), the PMC-MI dataset provides comprehensive training data for developing MLLMs capable of medical multi-image reasoning.

PMC-MI-Bench. The PMC-MI-Bench serves as our comprehensive evaluation benchmark, featuring 300 carefully curated test cases drawn from the diverse pool of processed compound figures. On average, each benchmark case includes 4.81 sub-figures. Furthermore, the average resolution for the compound figure within these benchmark cases is 688.2 pixels in width and 587.4 pixels in height. Each case is accompanied by substantial textual context comprising 102.2 words in the main caption, and 191.1 words in associated inline text, reflecting the complexity required for rigorous evaluation. The benchmark is structured to assess six distinct aspects of medical understanding, comprising the three specialized sub-tasks of multi-image VQA (holistic multiple sub-images reasoning, focused single sub-image analysis within a compound figure, and spatial relationship assessment), alongside standard single-image VQA, text-only QA, and multi-choice VQA. Each category contains 50 meticulously validated samples, ensuring balanced evaluation across different reasoning capabilities required for clinical practice. Details of the PMC-MI-Bench dataset assessment and proofing are provided in Section 3.6.

Both datasets undergo rigorous validation by medical professionals who verify accuracy and clinical relevance. PMC-MI-Bench specifically receives independent review from two board-certified physicians who assess the diagnostic appropriateness and medical accuracy of each test case. As such, the PMC-MI dataset enables training of MLLMs that can handle the multi-image reality of clinical practice, while PMC-MI-Bench provides standardized evaluation beyond simple classification tasks. Together, they capture longitudinal and multi-modal imaging scenarios essential for modern medical AI systems, offering a comprehensive assessment of MLLMs for clinical deployment in scenarios requiring integrated analysis of multiple imaging studies.

OmniMedVQA Benchmark. The OmniMedVQA²⁸ serves as a broad assessment benchmark for standard

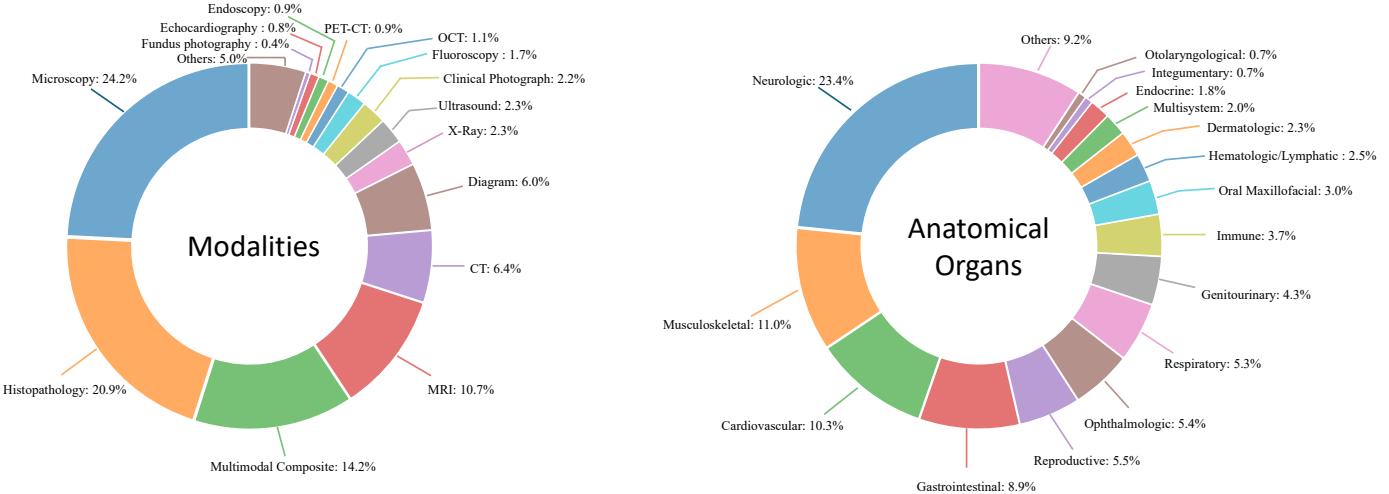


Figure 7: The distribution of the sampled PMC-MI dataset by modality and anatomical system. (a) The distribution illustrates the proportion of different imaging modalities, highlighting the diversity across microscopy (24.2%), histopathology (20.9%), multimodal composite images (14.2%), and radiological imaging such as MRI (10.7%) and CT (6.4%). Other modalities, including ultrasound, X-ray, and clinical photography, ensure comprehensive representation of real-world medical imaging. (b) The distribution shows the distribution of anatomical systems covered in the dataset, with the largest proportions attributed to the neurological system (23.4%), musculoskeletal system (11.0%), and cardiovascular system (10.3%), alongside significant contributions from gastrointestinal, respiratory, and other specialized systems. This distribution ensures comprehensive evaluation across diverse medical domains.

single-image medical VQA tasks. Following the evaluation protocol by MedEvalKit²⁶, we evaluate performance on this benchmark, which comprises 88,996 multi-choice visual questions derived from 82,059 images. It covers eight distinct imaging modalities, including Computed Tomography (CT), Fundus Photography (FP), Magnetic Resonance Imaging (MR), Optical Coherence Tomography (OCT), Dermoscopy (Der), Microscopy (Mic), X-Ray, and Ultrasound (US). The questions span five types, including Modality Recognition, Anatomy Identification, Disease Diagnosis, Lesion Grading, and Other Biological Attributes. The reported average performance across modalities is weighted based on the number of samples within each modality category²⁶.

MMMU-Med Benchmark. The MMMU-Med²⁹, a specialized subset of the larger MMMU dataset, provides a focused evaluation benchmark specifically for assessing single-image understanding capabilities in the medical domain. For our evaluation, we utilize the labeled validation set, which consists of 150 closed-ended multiple-choice visual questions. These questions are evenly distributed across five distinct biomedical subjects, with 30 questions per subject. The subjects of MMMU-Med cover Basic Medical Science (BMS), Clinical Medicine (CM), Diagnostics and Laboratory Medicine (DLM), Pharmacy (P), and Public Health (PH).

MIMIC Longitudinal Chest X-ray Benchmark. For clinical longitudinal validation, we utilize chest X-ray images sourced from the MIMIC database^{18,19}, obtained under appropriate CITI approval and fully de-identified following HIPAA guidelines. The benchmark dataset comprises 1,326 pairs of sequential chest X-ray examinations from individual patients, each paired with ground-truth labels. It specifically focuses on assessing disease progression across five common radiological findings: Consolidation, Edema, Pleural effusion, Pneumonia, and Pneumothorax. For each finding, the progression between the two examinations is categorized into one of three states: Improving, Stable, or Worsening. To ensure a rigorous evaluation that prevents data leakage, the dataset is split into training and test sets at the patient level. This benchmark structure allows us to evaluate the model’s longitudinal reasoning capabilities in a setting that mirrors real-

world diagnostic workflows where clinicians compare serial images to monitor patient status.

6 Methods

6.1 Preparation and Processing of PMC Multiple-Image Dataset

We construct a large-scale training corpus by harvesting the open-access subset of PubMed Central. As of June 18, 2024, the repository contained 6,106,189 papers. From this vast collection, we implement a rigorous three-step filtering pipeline to curate a high-quality dataset specifically for medical multi-image understanding.

Step 1: Preliminary Filtering. We first filter the papers to retain only those with licenses permitting research use, reducing the pool to 5,099,175 articles. To efficiently identify relevant content, we employ a fine-tuned PubmedBERT³³ to classify image-caption pairs based solely on their textual content. This text-based pre-screening allows us to rapidly identify 3.7 million potential medical image-text pairs from the 5.1 million papers. We employ a Vision Transformer (ViT) fine-tuned for compound figure detection to effectively distinguish compound figures from single-panel images and non-medical graphics. As a result, we identify 3,156,144 medical compound figures, excluding 643,401 non-compound or irrelevant images.

Step 2: Medical Content Screening. In the second step, we ensure the medical relevance of the images. We further refine this set to ensure high clinical relevance, and employ a specialized DenseNet-121³⁴, pretrained on the ImageCLEF³⁵, MedICaT³⁶, and DocFigure³⁷ datasets, to distinguish genuine medical imagery from non-medical graphics such as charts and diagrams. As a result, this step retains compound figures only if medical sub-images constitute over 90% of their visual content.

Step 3: Textual Quality Control. In the third step, we apply textual quality controls by establishing minimum length thresholds to guarantee sufficient context for instruction generation. We require compound-level captions to exceed 50 words and individual sub-image captions (if available) to contain at least 10 words.

The application of these three-step criteria systematically refines the initial harvest, resulting in a final, high-quality collection of 237,137 compound figures suitable for the subsequent instruction generation process, forming the basis for both the training dataset and the benchmark.

PMC-MI Dataset Generation. Our instruction generation paradigm comprises five interconnected stages designed to maximize information extraction from these filtered medical compound figures. We employ QWen2.5-32B³⁸ for automatic summarization of inline texts and medical terminology extraction, the advanced medical MLLM HuatuoGPT-Vision-34B⁹ for sub-image analysis, and template-based generation combined with large language model creativity to produce four distinct question types covering comprehensive medical image understanding scenarios. The specific details of this five-stage paradigm are elaborated in Section 6.2.

PMC-MI-Bench Curation Process and Professional Examination. From the larger pool of processed data, we randomly select diverse samples across medical specialties and imaging modalities for benchmark construction. Each potential benchmark sample then undergoes a rigorous preliminary screening for medical relevance, complexity suitable for benchmarking, and educational value. This process prioritizes cases that exemplify different aspects of compound figure understanding (*e.g.*, spatial, temporal, cross-modal analysis) while ensuring balanced representation across six defined question categories. Finally, these screened candidates undergo intensive professional validation distinct from the automated dataset generation process. Two board-certified

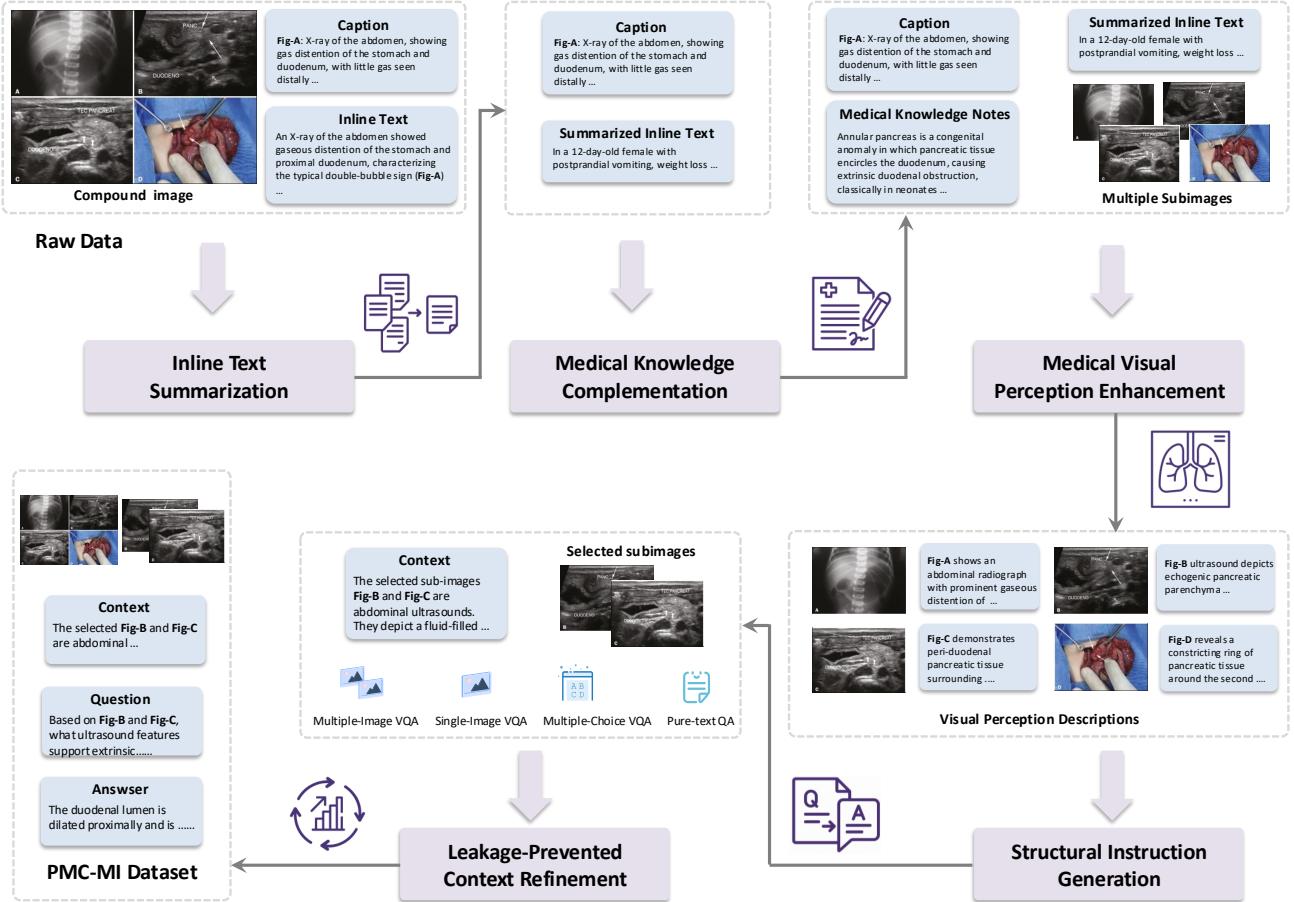


Figure 8: Illustration of the five-stage, context-aware instruction generation paradigm. The paradigm involves the inline text summarization, medical knowledge complementation, medical visual perception enhancement, structural instruction generation, and leakage-prevented context refinement. As such, the paradigm is performed on the collected PMC biomedical literature, and generates the high-quality PMC-MI dataset for training and manually proofed PMC-MI-Bench for evaluation.

medical professionals independently review each candidate, verifying information consistency between constructed contexts and original source materials, ensuring answer accuracy and completeness, and confirming diagnostic appropriateness. Inter-annotator agreement exceeds 85% across all evaluation criteria, with disagreements resolved through expert consultation. This multi-step process ensures PMC-MI-Bench maintains high quality, clinical relevance, and balanced distribution across question categories, medical specialties, and imaging complexity.

Quality Control Measures. Both the training dataset and benchmark employ automated filtering to remove questions with potential answer leakage, factual inconsistencies, or inadequate medical complexity. PMC-MI-Bench additionally undergoes manual verification of each test case to ensure benchmark reliability and clinical relevance.

6.2 Context-Aware Instruction Generation Paradigm

We develop a five-stage, context-aware instruction generation paradigm designed to systematically transform medical compound figures and their associated textual descriptions into comprehensive training data. Cen-

tral to our approach is a *divide-and-conquer* strategy that decomposes the complex challenge of multi-image context-aware instruction generation into a sequence of manageable, specialized sub-tasks. This instruction generation paradigm addresses the fundamental challenge of creating instruction data that captures not only visual content but also the complex clinical reasoning, medical knowledge integration, and multi-image relationship understanding required for effective medical compound figure analysis.

Stage 1: Inline Text Summarization. To address the challenges posed by the often lengthy and verbose inline text from biomedical literature, we employ the QWen2.5-32B³⁸ to systematically analyze and process the inline text regarding each compound figure. Inline texts in biomedical literature often contain essential clinical insights relevant to compound figure analysis and patient diagnosis, yet their verbosity can hinder the learning and understanding processes of MLLMs. To overcome this, our summarization stage goes beyond simple text extraction by condensing and reorganizing these inline references into concise and coherent clinical narratives. This process focuses on highlighting key information directly related to compound figures and patient diagnosis, such as pathological findings, diagnostic workflows, and treatment outcomes, while removing extraneous details. The resulting summarized texts not only preserve medical accuracy but also streamline complex technical descriptions into accessible formats, laying a foundational clinical context for the subsequent instruction generation paradigm. The prompt of this stage is illustrated in Fig. 15, and an example is elaborated in Fig. 26.

Stage 2: Domain-Specific Medical Knowledge Complementation. Building upon Stage 1, this stage utilizes QWen2.5-32B³⁸ to analyze the compound figure captions and the inline text summaries to systematically extract and elaborate on key medical concepts critical to understanding each case. This stage first identifies key medical concepts, such as symptoms, pathologies, diagnostic procedures, and treatment approaches, and then generates comprehensive explanations for each, including their clinical significance, diagnostic criteria, imaging characteristics, and relationships to other medical conditions. This process ensures that the instruction data is enriched with sufficient domain-specific medical knowledge, facilitating accurate clinical reasoning and diagnosis. We manually check sampled training data, confirming that the medical concepts extracted and elaborated by the LLM are both accurate and relevant to the case. By integrating this domain-specific medical knowledge into the paradigm, we provide the rich medical context essential for effective medical compound figure analysis. The prompt of this stage is illustrated in Fig. 16, and an example is elaborated in Fig. 27.

Stage 3: Multi-Modal Medical Visual Perception Enhancement. While Stage 1 and Stage 2 provide rich textual information, the context-aware instructions also require accurate visual knowledge to bridge the gap between text and medical images. To achieve this, it is critical to precisely analyze the content of medical compound figures, which consist of multiple subimages with distinct clinical implications. Given the difficulty current MLLMs face in processing compound figures holistically, we adopt a *divide-and-conquer* strategy, where each subimage is pre-segmented and analyzed individually using HuatuoGPT-Vision-34B⁹. This approach captures detailed visual features such as anatomical structures, pathological findings, imaging artifacts, and diagnostic characteristics, and provides links between visual findings and the textual context established in earlier stages. By synthesizing these subimage descriptions, we create a comprehensive understanding of the compound figure. This stage is essential for enriching our instruction dataset with precise multi-modal knowledge, enabling the M³LLM to reason effectively across both textual and visual domains and enhancing its diagnostic and clinical reasoning capabilities. The prompt of this stage is illustrated in Fig. 17, and an example is elaborated in Fig. 28.

Stage 4: Context-Question-Answer Instruction Generation. Building on the textual outputs from the previous stages, this stage constructs a diverse and clinically relevant instruction dataset designed to enhance the M³LLM with multi-modal reasoning. The instructions are categorized into four major types, each addressing

specific challenges in medical image analysis and question answering. Note that the ablation study of these four major types is presented in Table 7. (1) **Multi-Image VQA** focuses on improving the M³LLM’s ability to accurately analyze compound figures. This involves providing at least two input images and posing questions that require integration of information across multiple sub-images. Multi-Image VQA is further divided into three distinct subtypes: (a) questions that require synthesizing information from multiple sub-images to provide holistic case assessments, mimicking real-world clinical scenarios; (b) questions focusing on detailed understanding of a single specific sub-image while maintaining awareness of the broader context of the input compound figure; and (c) questions distinguishing spatial relationships between two specific sub-images, such as their relative positioning or alignment. This category is pivotal for enabling the M³LLM to handle the complexity of medical compound figures and is a key driver for improving its diagnostic accuracy in multi-image settings. We further present the ablation study on these three types of multi-image VQA instruction in Table 8. (2) **Single-Image VQA** ensures the model retains its ability to handle simpler but equally important tasks. This focuses on scenarios where only one image is available for analysis, testing the model’s capacity for detailed visual understanding, reasoning, and diagnostic insight from individual medical images. While less complex than composite analysis, this category remains essential for many real-world applications. (3) **Text-only QA** evaluates the M³LLM’s ability to process medical questions in a text-only context. This ensures that the model’s medical knowledge and reasoning capabilities remain robust even without visual input, allowing it to handle a wide range of clinical scenarios where textual information dominates. These tasks test the model’s understanding of medical concepts, clinical reasoning, and its ability to connect textual information with broader clinical knowledge. (4) **Multi-Choice VQA** introduces structured multi-choice questions, which are a common format in public benchmarks. These tasks assess the model’s ability to apply medical knowledge and diagnostic reasoning in a constrained and highly structured format. This category ensures that the M³LLM performs well on widely-used evaluation standards while maintaining consistency across different question formats. By integrating these four instruction types, this stage creates a comprehensive dataset that not only strengthens the M³LLM’s ability to analyze compound figures but also ensures its robustness in single-image reasoning, text-based medical question answering, and structured multi-choice formats. The prompt of this stage is illustrated in Fig. 18, 19, 20, 21, 22 and 23 for different tasks, and an example is elaborated in Fig. 29.

Stage 5: Leakage-Prevented Context Refinement. The final stage prevents the answer leakage issue in context-aware instruction generation. When constructing the paired context, question, and answer using LLMs, the process relies on the same source text, which can lead to the context including key information that reveals or hints at the correct answer. This issue undermines the challenge posed to the model, reducing the effectiveness of training and evaluation by allowing the model to rely on cues rather than genuine reasoning. To resolve this, we implement a rigorous refinement process using advanced language models to systematically review generated instructions. This involves detecting and removing unintended answer-related information in the context through analysis of linguistic patterns, medical terminology, and logical relationships between context and questions. By ensuring the context remains informative yet neutral, this stage preserves the integrity of training data, creating meaningful challenges that reflect authentic clinical reasoning. This not only enhances the model’s training efficacy but also ensures its performance is rooted in true understanding and inference, rather than exploiting unintended context cues. The prompt of this stage is illustrated in Fig. 24, and an example is elaborated in Fig. 30.

In summary, the five-stage, context-aware instruction generation paradigm creates a comprehensive corpus of training data that systematically develops multiple competencies essential for medical multi-image understanding. Each generated instruction pair undergoes final validation to ensure clinical accuracy, educational appropriateness, and alignment with real-world diagnostic workflows. The resulting training data encompasses diverse medical specialties, imaging modalities, and clinical scenarios, providing comprehensive coverage of medical multi-image understanding requirements.

6.3 Medical Multi-Image MLLM Architecture

Our M³LLM adopts a streamlined architecture optimized for medical compound figure understanding. The framework comprises three core components, including a Vision Transformer (ViT)³⁹ for comprehensive medical image feature extraction across multiple sub-images, a connector module consisting of two fully connected layers for visual-to-text alignment, and a LLM³⁸ for sophisticated clinical reasoning and text generation. This architecture maintains computational efficiency while enabling complex multi-image understanding through our innovative instruction generation paradigm. In our implementation, we select the InternVL-3-8B³ as the base model to fine-tune on the PMC-MI dataset, where the InternViT^{3,39} serves as the visual encoder and QWen2.5-7B³⁸ serves as the LLM within the MLLM architecture.

Unlike previous approaches that focus primarily on architectural modifications, our framework achieves multi-image understanding through sophisticated training data preparation and instruction generation. The ViT processes each sub-image within medical compound figures, generating rich visual representations that capture both individual image characteristics and cross-image relationships. The connector module facilitates seamless integration of these multi-perspective visual features with textual medical knowledge, enabling the LLM to perform comprehensive clinical reasoning across multiple imaging modalities.

6.4 Training Methodology and Optimization

Multi-Stage Training Protocol. Our training methodology implements a carefully designed multi-stage protocol that progressively develops the model’s capabilities from basic medical knowledge acquisition to sophisticated multi-image reasoning. The initial training stage focuses on fundamental medical concept understanding using single-image instructions, establishing a solid foundation of medical domain knowledge. Subsequent stages introduce increasingly complex multi-image scenarios, enabling the model to develop cross-image reasoning capabilities while maintaining accuracy in individual image analysis.

Instruction Diversity and Clinical Relevance Optimization. Throughout the training process, we maintain a careful balance between instruction complexity, clinical relevance, and educational value. Our methodology ensures that training data encompasses diverse clinical scenarios, including emergency diagnostics, longitudinal patient monitoring, multi-modal imaging integration, and specialist consultations. This comprehensive approach enables the model to handle the full spectrum of medical multi-image understanding requirements encountered in real clinical practice.

Training Configuration. To finetune our M³LLM, we use the AdamW optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1 \times 10^{-8}$. The initial learning rate is set to 5×10^{-5} , and we apply a cosine decay schedule with a warmup ratio of 0.03 to control the learning rate. Training is conducted for 3 epochs using a per-device batch size of 1 and a gradient accumulation step of 1. For regularization, we use a weight decay of 0.05 and apply gradient clipping with a maximum norm of 1.0. To enhance the training process, mixed-precision training (`bfloat16`) is enabled. The random seed is fixed to 42 to ensure reproducibility. We utilize the AdamW implementation from the HuggingFace Transformers framework and use Weights & Biases to track the training and experimental results.

7 Computing Hardware and Software

We use Python (version 3.12) for all experiments and analyses in this study, which can be replicated using the open-source libraries outlined below. All computations are executed on the Yale Misha high-performance computing platform, utilizing NVIDIA H200 GPUs with mixed precision support to ensure reproducible and scalable experimentation. Experiments are based on the PyTorch framework (version 2.7.0) and torchvision (version 0.22.0), leveraging the NVIDIA CUDA toolkit (CUDA 12.6, cuDNN 9.5, NCCL 2.26) for GPU acceleration, thereby enabling efficient large-scale multimodal model training. For implementation, we utilize the Transformers library (version 4.52.4) and the PEFT library (version 0.17.1) for model configuration and parameter-efficient fine-tuning. Concurrently, DeepSpeed (version 0.17.5) and xFormers (version 0.0.30) are employed for distributed optimization and memory-efficient attention computation. Data preprocessing and analysis are conducted using NumPy (version 2.2.6), Pandas (version 2.3.0), scikit-learn (version 1.7.1), and SciPy (version 1.15.3). Model evaluation involved Hugging Face Datasets (version 3.6.0), Evaluate (version 0.4.5), BERTScore (version 0.3.13), and ROUGE (version 1.0.1) libraries for calculating relevant metrics. For multimodal inference, we integrate vLLM (version 0.9.0.1), OpenCLIP (version 2.32.0), and LLaVA (version 1.7.0.dev0).

8 Evaluation Metrics

Evaluating medical MLLMs presents unique challenges due to the diverse output formats and clinical reasoning requirements. We employ a comprehensive framework for benchmarking and validation that addresses both open-ended text generation and multi-choice question answering scenarios, ensuring robust assessment of model capabilities across different clinical tasks.

8.1 Open-Ended Text Generation Evaluation

Open-ended medical text generation requires precise evaluation beyond simple text matching, as clinical accuracy and completeness are paramount for patient safety. To assess different aspects of response quality, we adopt a multi-faceted strategy that combines string-based metrics, semantic similarity assessment, and LLM-as-a-judge evaluation.

String-Based Metrics. We employ BLEU⁴⁰ and ROUGE⁴¹ to provide a baseline evaluation of linguistic similarity by quantifying n-gram overlap between model outputs and references. BLEU⁴⁰ measures precision, while ROUGE⁴¹, particularly ROUGE-L, emphasizes recall-oriented similarity. In practice, these metrics are limited in capturing the nuanced meanings of medical language, where small variations in terminology (e.g., *myocardial infarction* vs. *heart attack*) can significantly impact clinical interpretation.

Semantic Similarity Assessment. To evaluate deeper semantic alignment beyond string matching, we use two complementary metrics: BERTScore⁴² and Semantic Textual Similarity (STS)⁴³. BERTScore focuses on token-level semantic overlap. It computes contextual embeddings for each token in the prediction and reference text, performing optimal matching to derive the F1 score. This metric excels at assessing content fidelity and coverage, especially for domain-specific terminology, as it effectively handles paraphrasing and synonyms. In contrast, STS evaluates overall semantic equivalence, typically at the sentence level. It encodes each text into a single vector representation and calculates cosine similarity, providing a holistic score that reflects whether two texts convey the same meaning, regardless of wording differences. By combining these metrics, we achieve a comprehensive evaluation: BERTScore provides fine-grained insights into lexical and semantic alignment,

while STS offers a high-level measure of semantic similarity. We use the DeBERTa model⁴⁴ for BERTScore and MiniLM⁴⁵ for STS. This ensures a robust assessment of generated text, capturing nuanced semantic differences that string-based metrics might overlook.

LLM-as-a-judge Assessment. To achieve a comprehensive evaluation, we employ the LLM-as-a-judge approach⁴⁶ to leverage the evaluative LLM capability to provide a scalable assessment. This approach supplements traditional string-based and semantic metrics by incorporating human-like judgment to evaluate the nuanced quality of generated medical text. Specifically, we utilize GPT-4o⁴⁷ as the LLM judge to compare the outputs of our M³LLM against state-of-the-art MLLMs. For each sample in the evaluation dataset, we provide the LLM judge with a manually-proofed reference answer, alongside the outputs from both M³LLM and the competing MLLM, with the prompt illustrated in Fig. 25. Following the assessment protocol⁴⁸, the LLM judge compares the two generated outputs and determines which one more closely aligns with the reference answer, by assigning one of three possible outcomes for each sample: *win* for M³LLM, *lose* for the competing MLLM, or *tie* when neither output demonstrates a clear advantage. To quantify the overall performance, we calculate the average scores across the entire dataset by aggregating the *win*, *tie* or *lose* results.

8.2 Multi-Choice VQA Evaluation

For multi-choice VQA tasks, we evaluate the performance of state-of-the-art MLLMs using accuracy as the primary metric, adhering to the implementation protocols established in Lingshu²⁶. This evaluation assesses the model’s capability to accurately interpret multi-choice instructions and select the correct response from a set of options. We construct standardized inputs where visual features are prefixed to text embeddings containing the specific question and candidate options. To ensure rigorous evaluation, we tailor the system instructions to the specific requirements of each benchmark. For our PMC-MI-Bench, we utilize the prompt: *You are a medical expert who is good at solving medical multi-choice tasks. Please answer with the option letter only. The Question is: <Question>. The candidate options are: <Options>*. For public benchmarks such as OmniMedVQA and MMMU-Med, we adopt the default system prompt from the MedEvalKit codebase²⁶: *Answer with the option’s letter from the given choices directly. The Question is: <Question>. The candidate options are: <Options>*. Final predictions are determined by comparing the option letter generated by the model with the correct option letter from the ground truth. A prediction is considered correct only if the generated option letter matches the ground truth exactly, and these results are used to calculate accuracy.

9 Data Availability

The PMC-MI dataset, utilized for training M³LLM, and PMC-MI-Bench, designed for comprehensive evaluation, are publicly available for access and download via OneDrive ([link](#)). Detailed information on the usage can be found on the Hugging Face repository (<https://huggingface.co/datasets/KerwinFu/M3LLM-PMC>).

10 Code availability

The framework used for this study can be found at GitHub (<https://github.com/franciszchen/M3LLM>). The analysis framework to evaluate all results, generate all plots, and perform all statistical analyses can be found at Google Drive ([the link](#)). All code uses Python (v3.12), PyTorch (v2.7.0), torchvision (v0.22.0), Transformers (v4.52.4), PEFT (v0.17.1), DeepSpeed (v0.17.5), xFormers (v0.0.30), NumPy (v2.2.6), Pandas (v2.3.0), scikit-learn (v1.7.1), SciPy (v1.15.3), Hugging Face Datasets (v3.6.0), Evaluate (v0.4.5), BERTScore (v0.3.13), ROUGE (v1.0.1), vLLM (v0.9.0.1), OpenCLIP (v2.32.0), and LLaVA (v1.7.0.dev0).

11 Funding

This study is supported by the National Institutes of Health National Library of Medicine under Award Number R01LM014604 and R00LM014024.

12 Competing Interests

The authors declare no competing interests.

13 Online Content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgments, peer review information, details of author contributions and competing interests, and statements of data and code availability are available at the link.

14 References

1. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. *NeurIPS* **36**, 34892–34916 (2023).
2. Bai, S. *et al.* Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).
3. Chen, Z. *et al.* Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271* (2024).
4. Liu, F. *et al.* Application of large language models in medicine. *Nature Reviews Bioengineering* 1–20 (2025).
5. Li, C. *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS* **36**, 28541–28564 (2023).
6. Moor, M. *et al.* Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367 (PMLR, 2023).
7. Wu, C. *et al.* Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications* **16**, 7866 (2025).
8. Zhang, K. *et al.* A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine* **30**, 3129–3141 (2024).
9. Chen, J. *et al.* Towards injecting medical visual knowledge into multimodal LLMs at scale. In *EMNLP*, 7346–7370 (2024).
10. Xie, Y. *et al.* Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. In *ICLR* (2025).
11. Zhu, X., Hu, Y., Mo, F., Li, M. & Wu, J. Uni-med: a unified medical generalist foundation model for multi-task learning via connector-moe. In *NeurIPS*, vol. 37, 81225–81256 (2024).
12. Lin, T. *et al.* Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *ICML* (2025).
13. McKinney, S. M. *et al.* International evaluation of an ai system for breast cancer screening. *Nature* **577**, 89–94 (2020).
14. Ferber, D. *et al.* In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications* **15**, 10104 (2024).
15. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
16. Xing, W., Zhu, T., Wang, J. & Liu, B. A survey on mllms in education: application and future directions. *Future Internet* (2024).
17. Sabuncu, M. R. *et al.* Event time analysis of longitudinal neuroimage data. *NeuroImage* **97**, 9–18 (2014).
18. Johnson, A. E. *et al.* Mimic-iv, a freely accessible electronic health record dataset. *Scientific data* **10**, 1 (2023).
19. Bannur, S. *et al.* Learning to exploit temporal structure for biomedical vision–language processing. In *CVPR*, 15016–15027 (2023).
20. Sala, E. *et al.* Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clinical radiology* **72**, 3–10 (2017).
21. Rougier, N. P., Droettboom, M. & Bourne, P. E. Ten simple rules for better figures (2014).

22. Takaya, Y. & Ito, H. New horizon of fusion imaging using echocardiography: its progress in the diagnosis and treatment of cardiovascular disease. *Journal of echocardiography* **18**, 9–15 (2020).
23. Tae, W.-S., Ham, B.-J., Pyun, S.-B. & Kim, B.-J. Current clinical applications of structural mri in neurological disorders. *Journal of Clinical Neurology* **21**, 277 (2025).
24. Langlotz, C. P. *et al.* A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 nih/rsna/acr/the academy workshop. *Radiology* **291**, 781–791 (2019).
25. Li, F. *et al.* Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895* (2024).
26. Xu, W. *et al.* Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044* (2025).
27. Sellergren, A. *et al.* Medgemma technical report. *arXiv preprint arXiv:2507.05201* (2025).
28. Hu, Y. *et al.* Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *CVPR*, 22170–22183 (2024).
29. Yue, X. *et al.* Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 9556–9567 (2024).
30. Lin, W. *et al.* Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 525–536 (Springer, 2023).
31. Zhang, S. *et al.* Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915* (2023).
32. Lozano, A. *et al.* Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19724–19735 (2025).
33. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* **3**, 1–23 (2021).
34. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).
35. Ionescu, B. *et al.* Overview of the imageclef 2024: Multimedia retrieval in medical applications. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 140–164 (Springer, 2024).
36. Subramanian, S. *et al.* Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000* (2020).
37. Jobin, K., Mondal, A. & Jawahar, C. Docfigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 1, 74–79 (IEEE, 2019).
38. Team, Q. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115* (2025).
39. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR* (2021).
40. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318 (2002).

41. Banerjee, S. & Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, 65–72 (2005).
42. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
43. Reimers, N. & Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP* (2019).
44. He, P., Liu, X., Gao, J. & Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR* (2021).
45. Wang, W. *et al.* Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*, vol. 33, 5776–5788 (2020).
46. Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, vol. 36, 46595–46623 (2023).
47. Hurst, A. *et al.* Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
48. Zhang, H. *et al.* HuatuoGPT, towards taming language model to be a doctor. In *Findings of The 2023 Conference on Empirical Methods in Natural Language Processing* (2023).

Input

<Fig-0>

<Fig-1>

<Fig-2>

<Fig-3>

Reference Answer

In LDLR mice fed a high cholesterol diet (Fig-1 and Fig-2), the ultrastructural changes include severely degenerated microvessels with disrupted endothelial cell integrity, vacuolation, narrowed lumens, and significant lipid deposits. In contrast, LDLR-/- mice fed a low cholesterol diet (Fig-3) exhibit thickened basal laminae, expanded perivascular spaces, and degenerating capillary walls. These differences suggest that high cholesterol diets exacerbate vascular degeneration more severely than low cholesterol diets, even in genetically deficient mice.

MedGemma-27B

Based on the provided images, here's a comparison of the ultrastructural changes in microvessels: LDLR mice fed a high cholesterol diet (Fig-1 and Fig-2): **Fig-1A:** Shows a cross-section of a vessel with significant lipid deposition (marked by the white arrow) within the vessel wall, particularly in the medial layer. The intima appears thickened. **Fig-1B:** Shows a longitudinal section, revealing the vessel wall structure and the presence of lipid deposits (marked by the white arrow) within the medial layer. **Fig-1C:** Shows a cross-section with a prominent lipid-laden macrophage (foam cell, marked by the white arrow) within the intima. The vessel wall appears thickened. **Fig-2B:** Shows a cross-section of a vessel with extensive lipid deposits (marked by the white arrows) within the medial layer and potentially within the intima. The vessel wall appears significantly thickened. **LDLR-/- mice fed a low cholesterol diet (Fig-3):** **Fig-3C:** Shows a cross-section of a vessel. There are some cellular infiltrates (marked by the white arrows) within the vessel wall, but there is no obvious evidence of extensive lipid deposition or foam cell formation within the intima or media. **The vessel wall appears relatively normal in thickness compared to the high-cholesterol diet groups.** **Comparison: Lipid Deposition:** LDLR mice on a high cholesterol diet (Fig-1 and Fig-2) exhibit prominent lipid deposition within the vessel walls (media and intima), often forming foam cells (lipid-laden macrophages). In contrast, LDLR-/- mice on a low cholesterol diet (Fig-3) show minimal to no evidence of significant lipid deposition within the vessel walls. **Foam Cell Formation:** Foam cells are clearly visible in the intima.

M³LLM-8B

In the LDLR mice fed a high cholesterol diet (Fig-1 and Fig-2), the microvessels exhibit significant ultrastructural changes, including the presence of numerous lipid droplets and vacuoles within the endothelial cells and pericytes. This indicates increased lipid accumulation and potential disruption of the vascular barrier function. In contrast, LDLR-/- mice fed a low-cholesterol diet (Fig-3) show abnormal microvascular architecture characterized by a thickened basal lamina, expanded perivascular spaces, and degenerating capillary walls. These features indicate that, although lipid accumulation is reduced compared to high-cholesterol LDLR mice, structural deterioration of the vessel wall is still present under low-cholesterol conditions.

Figure 9: Comparative case study of M³LLM-8B and MedGemma-27B on the multi-image VQA (w.r.t. multiple sub-images). The case requires the model to interpret transmission electron micrographs from LDLR mice fed a high-cholesterol diet and LDLR-/- mice on a low-cholesterol diet. M³LLM identifies key ultrastructural changes, including disrupted endothelial integrity, lipid accumulation, and narrowed lumens in high-cholesterol samples, as well as thickened basal lamina and degenerating capillary walls in low-cholesterol samples. In contrast, MedGemma's descriptions contain significant errors, such as misclassifying lipid vacuoles and hallucinating structural annotations, demonstrating its limitations in analyzing complex microvascular pathologies.

30

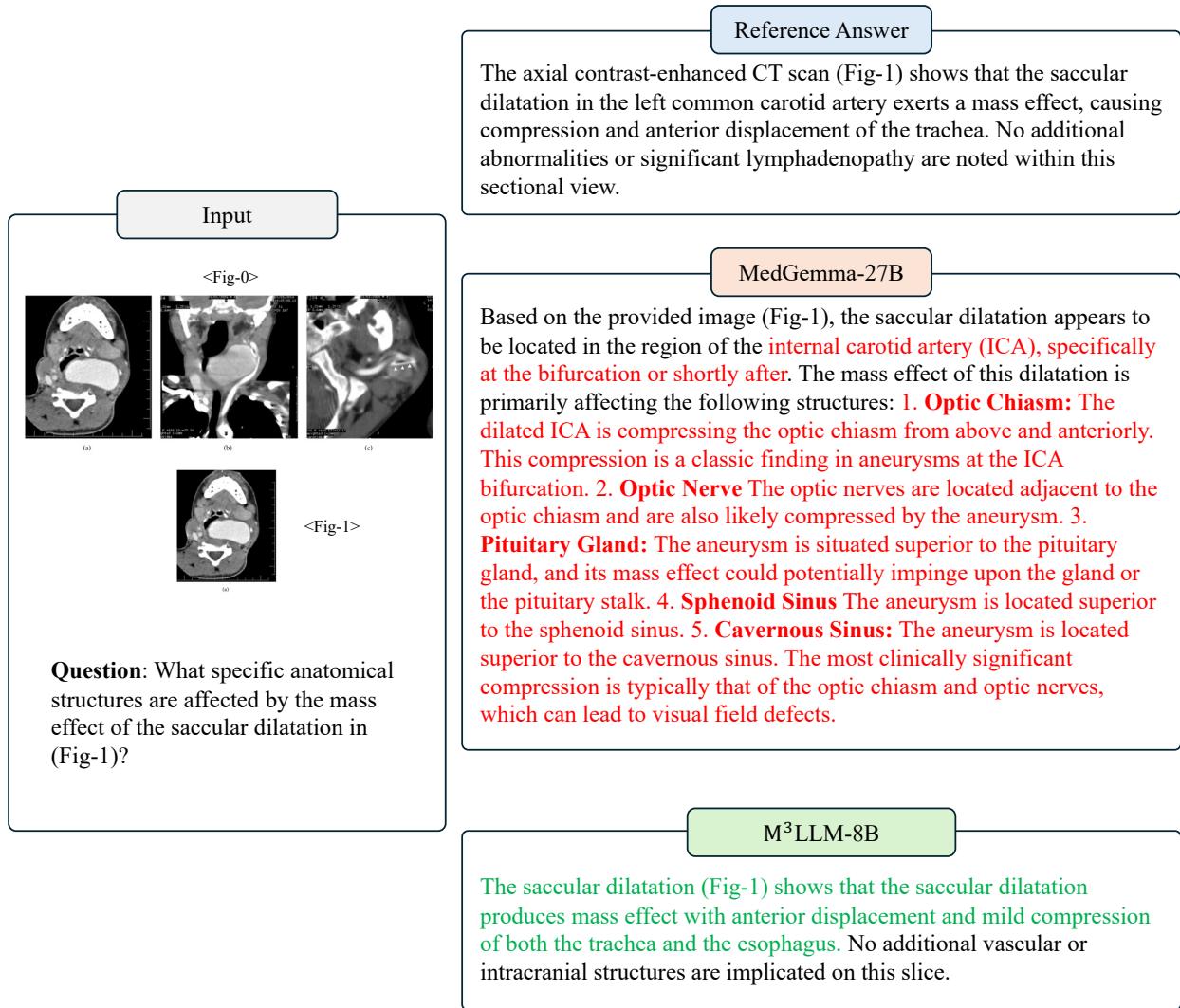


Figure 10: Comparative case study of M³LLM-8B and MedGemma-27B on the multi-image VQA (w.r.t. single sub-image within compound figure). The case requires identifying anatomical structures affected by a mass effect in a neck CT image. The response of M³LLM, identifying the trachea and esophagus, aligns with the reference answer. MedGemma's response, however, contains significant factual errors, hallucinating anatomical structures not relevant to this image slice and demonstrating a critical risk of error in visual reasoning.

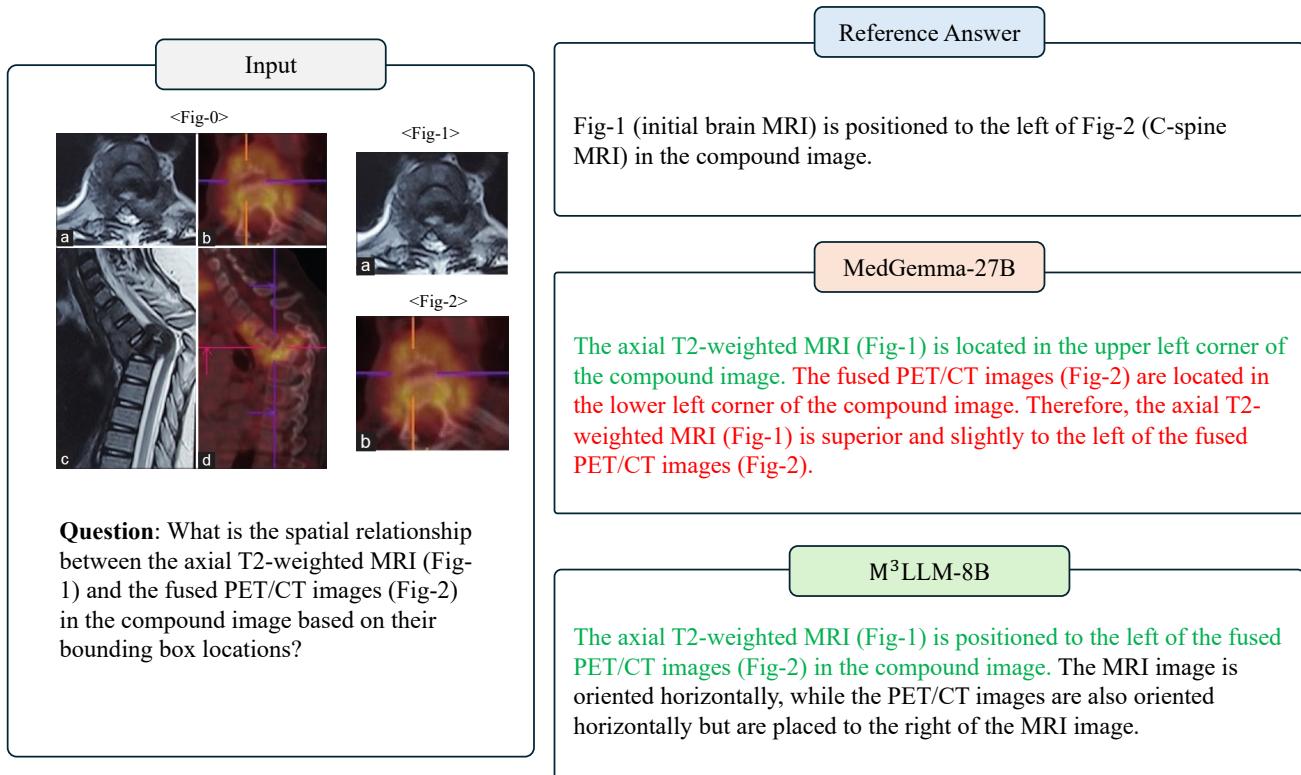


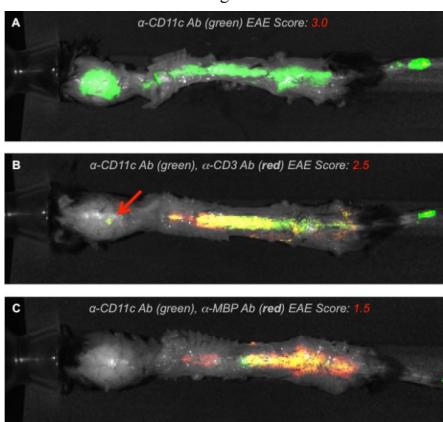
Figure 11: Comparative case study of M³LLM-8B and MedGemma-27B on the multi-image VQA (w.r.t. spatial relationship). The case tests the model's ability to determine the relative positions of sub-images in a compound figure. M³LLM accurately describes the left-right spatial relationship between the MRI and the PET/CT. The response from MedGemma, however, contains a fundamental error in spatial localization, indicating a deficiency in its ability to comprehend compound figure structures.

Reference Answer

The distribution of CD11c+ dendritic cells correlates with the severity of EAE. In mice with more severe EAE, there is increased accumulation of DCs in CNS lesions, particularly around blood vessels and demyelinating areas marked by anti-MBP antibodies. This suggests that DCs migrate to CNS lesions in response to inflammation, potentially contributing to antigen presentation and disease progression.

Input

<Fig-0>



Question: How does the distribution of CD11c+ dendritic cells correlate with the severity of experimental autoimmune encephalomyelitis (EAE) in the compound figure?

MedGemma-27B

The compound figure shows the distribution of CD11c+ dendritic cells (green) in the spinal cord at different stages of experimental autoimmune encephalomyelitis (EAE) using **two-photon microscopy**. The **red channel represents the EAE score**. **Fig-0A (EAE Score: 3.0):** Shows a relatively high density of green (CD11c+) cells, particularly concentrated in the area indicated by the red arrow in Fig-0B. The EAE score is 3.0, indicating a moderate to severe stage of the disease. **Fig-0B (EAE Score: 2.5):** The red arrow points to a region with a high concentration of green (CD11c+) cells. The EAE score is 2.5, which is also a moderate stage. **Fig-0C (EAE Score: 1.5):** Shows a lower density of green (CD11c+) cells compared to Figs-0A and 0B. The EAE score is 1.5, indicating a milder stage of the disease. **Correlation:** Based on the images, there appears to be a positive correlation between the density/concentration of CD11c+ dendritic cells and the severity of EAE. Higher EAE scores (3.0 and 2.5) are associated with a higher density of CD11c+ cells, particularly in specific regions (indicated by the arrow). A lower EAE score (1.5) is associated with a lower density of CD11c+ cells. Therefore, the distribution of CD11c+ dendritic cells seems to correlate with the severity of EAE, with higher severity being associated with a greater presence of these cells in the spinal cord.

M³LLM-8B

The distribution of CD11c+ dendritic cells correlates with the severity of experimental autoimmune encephalomyelitis (EAE) as follows: **Panel A (EAE Score: 3.0):** Shows a **higher density** of CD11c+ dendritic cells, indicating a more severe EAE condition. **Panel B (EAE Score: 2.5)** Demonstrates a **moderate density** of CD11c+ dendritic cells, reflecting a less severe EAE condition. **Panel C (EAE Score: 1.5):** Exhibits the **lowest density** of CD11c+ dendritic cells, corresponding to the mildest EAE condition. **This pattern suggests that increased infiltration of CD11c+ dendritic cells is associated with more severe EAE.**

Figure 12: Comparative case study of M³LLM-8B and MedGemma-27B on the single-image VQA. The case requires the model to correlate the intensity of the CD11c+ signal with EAE disease severity across three microscopy images. M³LLM successfully identifies the positive correlation and visual gradient between signal intensity and the disease score. MedGemma, however, misidentifies this group of images as two-photon microscopy, highlighting M³LLM's superior comprehensive analysis capabilities for compound figures.

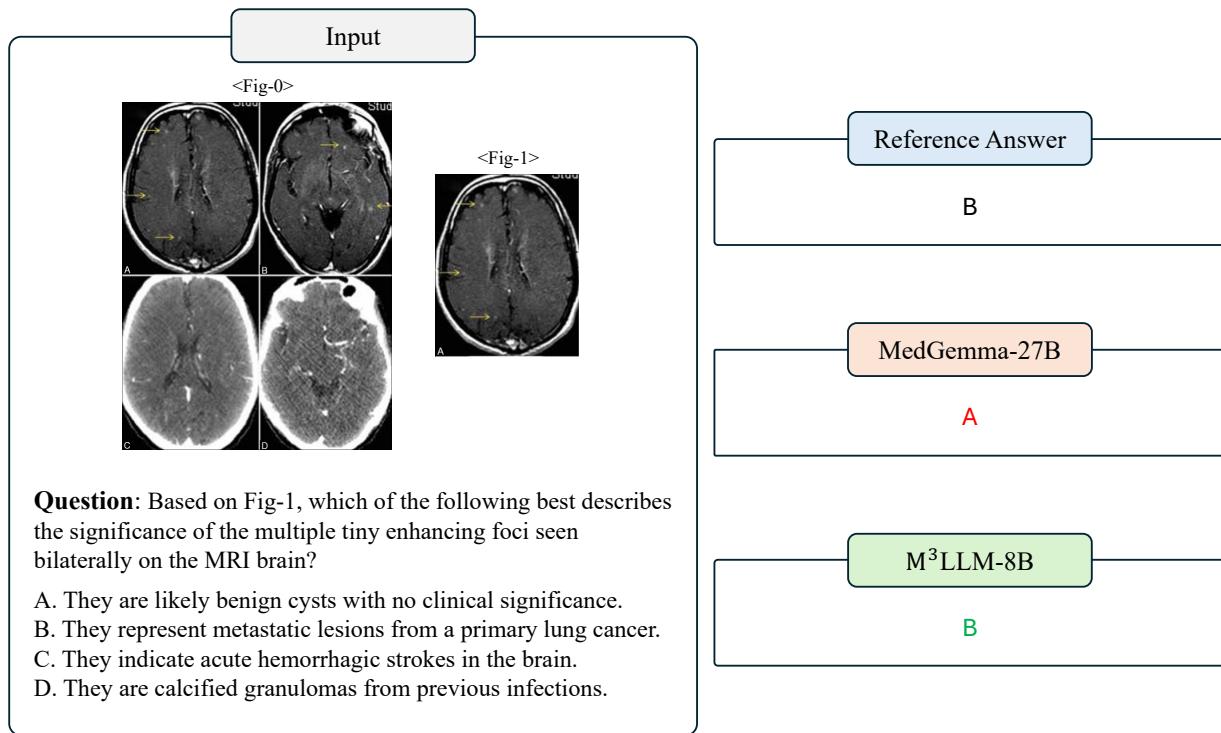


Figure 13: Comparative case study of M³LLM-8B and MedGemma-27B on the multi-choice VQA. The case requires the model to select a diagnosis from multiple options based on brain imaging. M³LLM successfully chooses the correct option **B**, whereas MedGemma selects the incorrect option **A**. This comparison demonstrates M³LLM's superior performance in a structured diagnostic question-answering format.

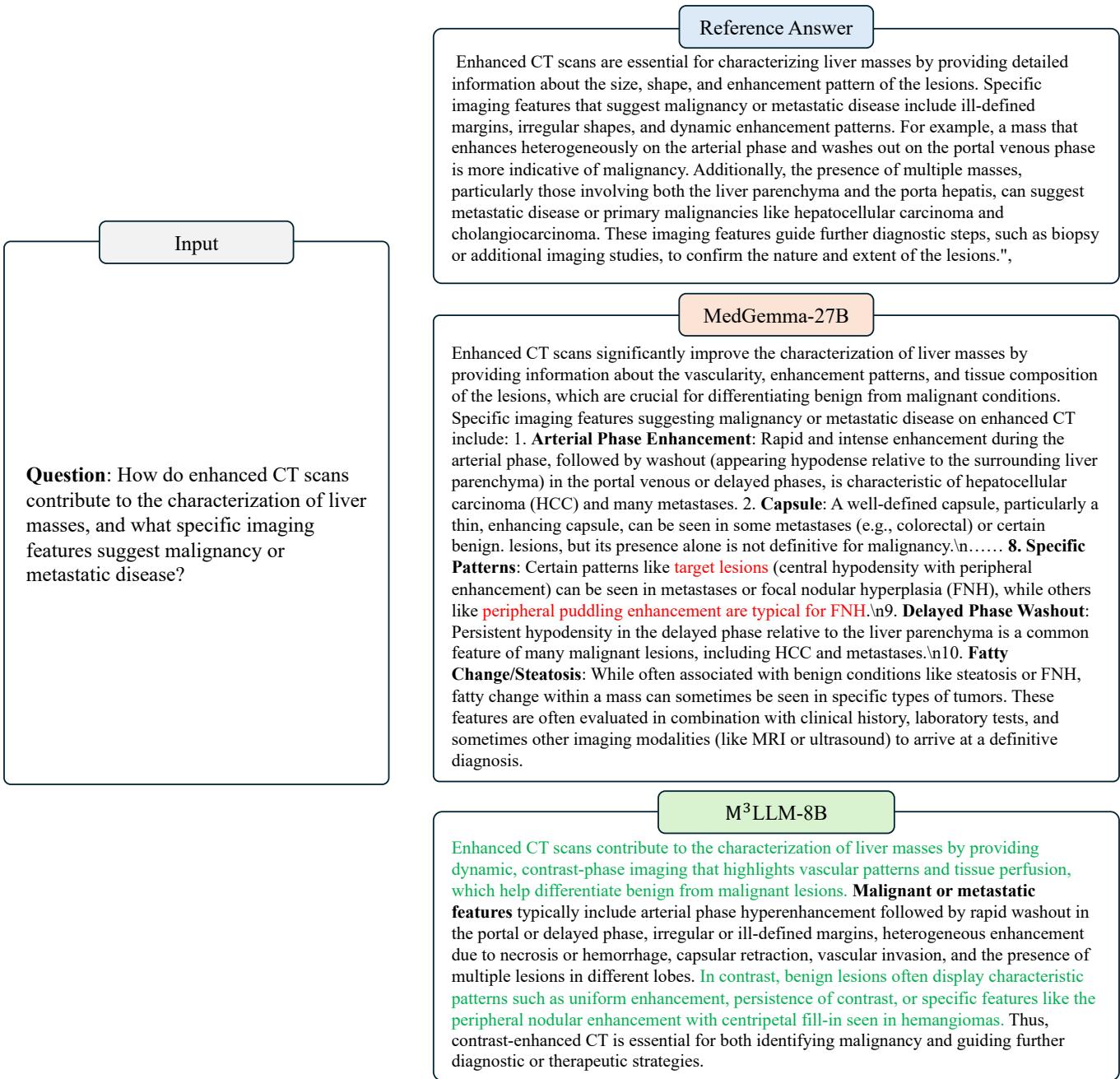


Figure 14: **Comparative case study of M³LLM-8B and MedGemma-27B on the text-only QA.** The case requires the models to answer a technical question about CT features of liver masses without visual input. M³LLM provides an accurate description consistent with the reference answer. The response from MedGemma, however, includes multiple errors and imprecise statements, such as conflating the typical features of different lesions. This highlights M³LLM's superior performance in text-only medical reasoning.

Prompt for Stage 1: Inline Text Summarization

Role

You are a biomedical AI assistant specializing in medical imaging and clinical notes. Your primary task is to summarize inline text from PubMed biomedical research papers, focusing on key medical observations, analyses, and conclusions related to patient conditions and diagnostic findings.

Task

Input Description

The input consists of inline text, which refers to multiple text excerpts from a research paper where a compound medical image is mentioned. These excerpts describe the observations, analyses, or clinical interpretations of the compound image and are often scattered across the text. Inline text provides fragmented but relevant details about the image that need to be consolidated into a coherent summary.

Specific Task

Summarize the key points from the inline text, focusing on:

- Observations directly related to patient conditions, medical findings, and diagnostic conclusions.
- Insights that highlight the clinical or research significance of the compound image.
- Excluding irrelevant, repetitive, or vague information. The summary must be clear, concise, and strictly limited to 100 words.

Objective

1. **Accuracy:** Extract only medically relevant details from the inline text. Avoid adding unverified or speculative information.
2. **Clarity:** Ensure the summary is concise and easy to understand. Use formal, standardized medical terminology.
3. **Relevance:** Focus on essential observations, analyses and conclusions related to the compound image. Exclude redundant or non-contributory content.
4. **Brevity:** Eliminate redundancy or unnecessary complexity. Consolidate information into a logical and compact form.

Instructions

1. **Focus on Key Medical Details:** Identify and include only details related to the patient conditions, diagnostic findings, or medical significance of the compound image.
2. **Eliminate Irrelevant Content:** Remove vague, redundant, or non-medical details that do not contribute to understanding the compound image.
3. **Maintain Clarity and Precision:** Use clear, professional language suitable for academic and clinical audiences.
4. **Word Limit:** Ensure the summary does not exceed 100 words.

Output Format

Inline Summary: A concise summary of the inline text, focusing on the most important medical observations, analyses, and conclusions related to the compound image.

Example

Input:

Inline Text: "The compound image highlights axial and coronal CT views of a pulmonary nodule located in the left upper lobe. The nodule is well-circumscribed, suggesting a benign etiology. Surrounding lung parenchyma appears normal.", "Further analysis of the pulmonary nodule indicates no evidence of calcification or ground-glass opacity. These findings reduce the likelihood of malignancy.", "Although the nodule's smooth margins suggest benignity, clinical follow-up is recommended to rule out potential malignant transformation."

Output:

Inline Summary: Axial chest CT demonstrates a focal consolidation in the right lung, suggestive of a potential infectious process.

Figure 15: The prompt design for Stage 1 of the instruction generation paradigm: Inline Text Summarization. This prompt guides the LLM in condensing fragmented and verbose inline text into concise clinical narratives. This stage establishes a standardized textual foundation for the subsequent context-aware instruction generation process.

Prompt for Stage 2: Medical Knowledge Complementation

Role

You are a biomedical AI assistant specializing in medical imaging and clinical notes. Your primary task is to generate concise and accurate medical knowledge for clinical, academic, and AI training purposes in English.

Task

Input Description

The input includes:

1. **Caption:** Describes the visual contents of medical images (e.g., CT, MRI, histology).
2. **Inline Summary:** A concise summary of key medical observations, analyses, and conclusions related to the compound image.

Specific Task

1. Extract medically significant keywords from both the input Caption and Inline Summary.
2. Generate concise medical knowledge focusing on the clinical or diagnostic relevance of the extracted keywords.

Objective

1. **Accuracy:** Ensure all extracted keywords and medical knowledge are medically accurate and supported by established guidelines or literature.
2. **Conciseness:** Limit the medical knowledge to 5 sentences, focusing on the most clinically relevant details.
3. **Professionalism:** Use formal, standardized medical terminology.
4. **Relevance:** Align the medical knowledge with the extracted keywords and clinical context.

Instructions

1. **Extract Keywords:** Identify key medical terms directly related to the Caption and Inline Summary, such as diseases, imaging modalities, anatomical structures, or findings.
2. **Generate Knowledge:** Create concise medical knowledge based on the extracted keywords, emphasizing their clinical or diagnostic importance.

Output Format

Keywords: A list of extracted keywords.

Medical Knowledge: A concise explanation of the keywords, limited to 5 sentences.

Example

Input:

- **Caption:** CT scan of the chest showing a solitary pulmonary nodule in the upper lobe of the right lung.
- **Inline Summary:** Axial and coronal CT scans show a well-circumscribed pulmonary nodule in the right upper lobe, with no calcification or ground-glass opacity, reducing the likelihood of malignancy. Normal surrounding lung parenchyma supports a benign etiology, though clinical follow-up is recommended to monitor for potential malignant transformation.

Output:

- **Keywords:** ["pulmonary nodule", "right upper lobe", "benign etiology", "CT imaging"]
- **Medical Knowledge:** A pulmonary nodule is a small, round lesion within the lung that can represent benign or malignant conditions. CT imaging is essential for evaluating nodules, providing detailed information about their size, margins, and internal characteristics. Smooth margins and the absence of calcification or ground-glass opacity suggest a benign etiology. However, clinical follow-up is recommended to rule out malignancy, as some nodules may transform over time. The location in the right upper lobe and normal surrounding lung parenchyma further support the benign diagnosis.

Figure 16: The prompt design for Stage 2 of the instruction generation paradigm: Medical Knowledge Complementation. This prompt guides the LLM to extract key medical concepts from captions and summaries, and then generate explanatory background knowledge around them. This stage adds necessary medical depth and context to the descriptive source texts, forming a basis for complex downstream tasks.

Prompt for Stage 3: Visual Perception Enhancement

Role

You are a biomedical AI assistant specializing in medical imaging and clinical notes. Your task is to generate concise and detailed descriptions of medical images for clinical and AI training purposes.

Task

Input Description

The input includes:

1. **Image Captions:** Original and rewritten captions describing the medical image.
2. **Corresponding Image.**

Specific Task

Generate a professional description of the image, focusing on visible findings, their clinical relevance, and additional insights based on visual perception.

Objective

1. **Accuracy:** Ensure all details are medically accurate and based on visible features or provided references.
2. **Conciseness:** Limit the description to 5 sentences, focusing on the most relevant and visually significant findings.
3. **Professionalism:** Use formal medical language and maintain scientific rigor.
4. **Visual Focus:** Provide insights grounded in the image's visual features, avoiding redundancy with the captions or medical knowledge.

Instructions

1. **Use Reference:** Use the captions and medical knowledge for context but focus on describing what is visually observable in the image. Avoid repeating or paraphrasing the provided references.
2. **Focus on Visible Details:** Highlight significant findings visible in the image, such as anatomical features, abnormalities, or patterns. Provide additional insights where possible.
3. **Highlight Clinical Relevance:** Emphasize visually derived details that are clinically meaningful and add diagnostic value.
4. **Avoid Hallucination:** Do not infer or assume details beyond what is visible in the image or explicitly supported by the references.

Output Format

Visual Perception Description: A concise and professional description of the medical image, limited to 5 sentences, emphasizing visual findings and clinical relevance.

Example

Input:

- **Original Caption:** CT scan of the chest showing a solitary pulmonary nodule in the right upper lobe.
- **Image Input**

Output:

Visual Perception Description: The axial chest CT shows a solitary, well-defined pulmonary nodule in the right upper lobe with smooth margins and no visible calcifications. The nodule measures approximately 1.2 cm and is located adjacent to the pleural surface. No surrounding ground-glass opacity, lymphadenopathy, or pleural effusion is observed. The proximity of the nodule to the pleura may warrant further evaluation for any subtle pleural involvement. These features suggest a likely benign etiology, though clinical correlation is recommended.

Figure 17: The prompt design for Stage 3 of the instruction generation paradigm: Medical Visual Perception Enhancement. This prompt directs the MLLM to generate precise, visually grounded descriptions focusing on observable anatomical structures and pathological findings. By enforcing strict adherence to each panel of the compound figure, this stage effectively bridges the gap between raw visual content and textual medical knowledge, ensuring the fidelity of subsequent multi-modal reasoning.

Prompt for Stage 4: Structural Instruction Generation - Multi-image VQA w.r.t. Multiple Sub-images

Role

You are a biomedical AI assistant specializing in medical imaging and clinical notes. Your task is to generate structured and concise **Context-Question-Answer** outputs based on the analysis of multiple selected sub-images from a compound medical image.

Task

Data Description

The input includes:

1. Compound Image Information:

- **Index:** The unique numeric identifier of the compound image (e.g., Fig-0).
- **Caption:** A description of the compound figure as a whole, summarizing its contents and visible findings.
- **Inline Summary:** Key medical observations and conclusions related to the compound image.
- **Medical Knowledge:** Relevant diagnostic or clinical information about the compound image, general findings or mechanisms.

2. Selected Sub-Images:

- A list of indices for the selected sub-images (e.g., Fig-1, Fig-2).

3. Specific Details for Each Selected Sub-Image:

- **Index:** The numeric identifier of the sub-image.
- **Caption:** A description of the visual content of the sub-image.
- **Visual Perception Description:** Observations of the sub-image's visual features, including key findings and structures.
- **Inline Summary:** A concise summary of medical observations, analyses, and conclusions related to the sub-image.

Specific Task

Generate a structured output consisting of:

1. **Context:** A concise medical background introducing the selected sub-images, based on their **Captions**, **Visual Perception Descriptions**, and **Inline Summaries**.
2. **Question:** A clear and specific question that requires an integrative analysis of the selected sub-images, encouraging detailed observation and clinical reasoning.
3. **Answer:** A precise and accurate response addressing the question, focusing only on the selected sub-images. The outputs must focus **only on the selected sub-images** and avoid interference from other sub-images or the compound image as a whole.

Objective

1. **Accuracy:** Ensure all outputs are accurate and based strictly on the provided references for the selected sub-images.
2. **Clarity:** Ensure the **Question** and **Answer** are logically connected and specific to the selected sub-images.
3. **Relevance:** Focus on the selected sub-images, avoiding interference from other sub-images or the compound image.

Instructions

1. **Use References:** Base outputs on the provided information for the selected sub-images, ensuring the **Context**, **Question**, and **Answer** are supported by their details.
2. **Focus on Selected Sub-Images:** Ensure all outputs focus only on the selected sub-images, avoiding references to unselected sub-images or the compound image as a whole.
3. **Integrate Sub-Image Information:** Synthesize information from the selected sub-images to create a cohesive **Context**, **Question**, and **Answer** that encourages integrative thinking.

Output Format

- **Context:** A concise and relevant medical background introducing the selected sub-images.
- **Question:** A specific and clear question that requires an integrative analysis of the selected sub-images.
- **Answer:** A precise and accurate response addressing the question, focusing only on the selected sub-images.

Example

Input:

Compound Image:

- **Index:** 0
- **Caption:** A compound figure of chest CT imaging showing multiple views of a pulmonary nodule.
- **Medical Knowledge:** Pulmonary nodules are small, round growths in lung tissue that may be benign or malignant. Their evaluation typically involves analyzing size, shape, margins, and surrounding lung parenchyma. Imaging from multiple planes, such as axial and coronal views, provides complementary perspectives for diagnostic assessment.

Selected Sub-Images:

- **Indices:** Fig-1, Fig-2
- **Specific Sub-Image Details:**
- **Index:** Fig-1
- **Caption:** An axial CT image of the chest demonstrating a solitary pulmonary nodule in the left upper lobe.
- **Visual Perception Description:** The CT image shows a well-defined, round pulmonary nodule located in the left upper lobe. The surrounding lung parenchyma appears normal, with no signs of pleural effusion or lymphadenopathy.
- **Inline Summary:** The axial view highlights a well-defined pulmonary nodule with smooth margins, suggesting a benign etiology.
- **Index:** Fig-2
- **Caption:** A coronal CT image showing the same pulmonary nodule in the left upper lobe from a different plane.
- **Visual Perception Description:** The coronal CT image demonstrates the nodule's position relative to the surrounding lung structures, confirming its well-defined borders and normal adjacent lung parenchyma.
- **Inline Summary:** The coronal view provides complementary information about the nodule's location and its relationship to nearby lung structures, supporting the interpretation of a benign lesion.

Output:

- **Context:** The selected sub-images (Fig-1 and Fig-2) highlight axial and coronal CT views of a solitary pulmonary nodule located in the left upper lobe. Axial and coronal imaging provide complementary perspectives for evaluating the nodule's size, margins, and its relationship to adjacent lung structures. Together, these views confirm the nodule's well-defined appearance and normal surrounding lung parenchyma, which are key features in assessing its clinical significance.
- **Question:** How do the axial and coronal CT views (Fig-1 and Fig-2) contribute to the evaluation of the pulmonary nodule in the left upper lobe?
- **Answer:** Axial and coronal CT views provide complementary information for assessing the pulmonary nodule. The axial view characterizes the nodule's shape, size, and margins, while the coronal view offers insights into its location and relationship to surrounding lung structures. Together, these perspectives confirm the nodule's well-defined borders and the absence of abnormalities in the adjacent lung parenchyma, supporting the interpretation of a benign lesion.

Figure 18: The prompt design for Stage 4 of the instruction generation paradigm: Structural Instruction Generation for multi-image VQA (w.r.t. multiple sub-images). This prompt demonstrates how information from multiple sub-images is automatically converted into a structured training sample comprising a context, question, and answer. As a core part of Stage 4, it is designed to create complex instructions that train the model to perform integrative reasoning across different views.

Prompt for Stage 4: Structural Instruction Generation - Multi-image VQA w.r.t. Single Sub-image in Compound Figure

Role

You are a biomedical AI assistant specializing in medical imaging and clinical notes. Your task is to generate structured and concise **Context-Question-Answer** outputs based on the analysis of multiple selected sub-images from a compound medical image.

Task

Data Description

The input includes:

1. Compound Image Information:

- **Index:** The unique numeric identifier of the compound image (e.g., Fig-0).
- **Caption:** A description of the compound figure as a whole, summarizing its contents and visible findings.
- **Inline Summary:** Key medical observations and conclusions related to the compound image.
- **Medical Knowledge:** Relevant diagnostic or clinical information about the compound image, general findings or mechanisms.
- 2. Selected Sub-Image:**
- The index for the selected sub-image (e.g., Fig-1).
- 3. Specific Details for Selected Sub-Image:**
- **Index:** The numeric identifier of the sub-image.
- **Caption:** A description of the visual content of the sub-image.
- **Visual Perception Description:** Observations of the sub-image's visual features, including key findings and structures.
- **Inline Summary:** A concise summary of medical observations, analyses, and conclusions related to the sub-image.

Specific Task

Generate a structured output consisting of:

1. **Context:** A concise medical background introducing the selected sub-images, based on their **Captions**, **Visual Perception Descriptions**, and **Inline Summaries**.
2. **Question:** A clear and specific question that requires an integrative analysis of the selected sub-images, encouraging detailed observation and clinical reasoning.
3. **Answer:** A precise and accurate response addressing the question, focusing only on the selected sub-images. The outputs must focus **only on the selected sub-image**.

Objective

1. **Accuracy:** Ensure all outputs are accurate and based strictly on the provided references for the selected sub-image.
2. **Clarity:** Ensure the **Question** and **Answer** are logically connected and specific to the selected sub-image.
3. **Relevance:** Focus on the selected sub-image.

Instructions

1. **Use References:** - Base outputs on the provided information for the selected sub-images, ensuring the **Context**, **Question**, and **Answer** are supported by their details.
2. **Focus on Selected Sub-Image:** - Ensure all outputs focus only on the selected sub-images, avoiding references to unselected sub-image or the compound image as a whole.
3. **Integrate Sub-Image Information:** - Synthesize information from the selected sub-image to create a cohesive **Context**, **Question**, and **Answer** that encourages integrative thinking.

Output Format

- **Context:** A concise and relevant medical background introducing the selected sub-image.
- **Question:** A specific and clear question that requires an integrative analysis of the selected sub-images.
- **Answer:** A precise and accurate response addressing the question, focusing only on the selected sub-image.

Example

Input:

Compound Image:

- **Index:** 0
- **Caption:** A compound figure of chest CT imaging showing multiple views of a pulmonary nodule.
- **Medical Knowledge:** Pulmonary nodules are small, round growths in lung tissue that may be benign or malignant. Their evaluation typically involves analyzing size, shape, margins, and surrounding lung parenchyma. Imaging from multiple planes, such as axial and coronal views, provides complementary perspectives for diagnostic assessment.

Selected Sub-Image:

- **Index:** Fig-1
- **Specific Sub-Image Details:**
- **Index:** Fig-1
- **Caption:** An axial CT image of the chest demonstrating a solitary pulmonary nodule in the left upper lobe.
- **Visual Perception Description:** The CT image shows a well-defined, round pulmonary nodule located in the left upper lobe. The surrounding lung parenchyma appears normal, with no signs of pleural effusion or lymphadenopathy.
- **Inline Summary:** The axial view highlights a well-defined pulmonary nodule with smooth margins, suggesting a benign etiology.

Output:

- **Context:** This axial CT image (Fig-1) shows a solitary pulmonary nodule in the left upper lobe. Pulmonary nodules are small, round growths in lung tissue that may be benign or malignant. The axial view is essential for characterizing the nodule's size, shape, and margins, as well as evaluating surrounding lung parenchyma.
- **Question:** What are the key visible imaging features of the pulmonary nodule in (Fig-1)?
- **Answer:** The axial CT image (Fig-1) demonstrates a well-defined, round pulmonary nodule located in the left upper lobe. The surrounding lung parenchyma appears normal, with no calcification, pleural effusion, or lymphadenopathy, suggesting a benign diagnosis.

Figure 19: The prompt design for Stage 4 of the instruction generation paradigm: Structural Instruction Generation for multi-image VQA (w.r.t. single sub-image within compound figure). This prompt is designed to convert detailed information from a single sub-image into a training sample. This task ensures that while the model learns complex cross-image reasoning, it does not lose its foundational ability to conduct in-depth analysis of individual images.

Prompt for Stage 4: Structural Instruction Generation - Multi-image VQA w.r.t. Spatial Relationship

Role

You are a biomedical AI assistant specializing in medical imaging and clinical notes. Your task is to generate structured and concise **Question-Answer** outputs based on the relative positions of two selected sub-images from a compound medical image. You already pre-computed the relative position of two figures, please rephrase this positional information into a professional and medically accurate QA pair.

Task

Data Description

The input includes:

1. **Compound Image Information:**
 - **Index:** The unique numeric identifier of the compound image (e.g., Fig-0).
 - **Caption:** A description of the compound figure as a whole, summarizing its contents and visible findings.
 - **Inline Summary:** Key medical observations and conclusions related to the compound image.
 - **Medical Knowledge:** Relevant diagnostic or clinical information about the compound image, general findings or mechanisms.
2. **Selected Sub-Images:**
 - A list of indices for the selected sub-images (e.g., Fig-1, Fig-2).
3. **Specific Details for Each Selected Sub-Image:**
 - **Index:** The numeric identifier of the sub-image.
 - **Caption:** A description of the visual content of the sub-image.
 - **Visual Perception Description:** Observations of the sub-image's visual features, including key findings and structures.
 - **Bounding Box Information:** The coordinates of the central point of the sub-image bounding box in the compound image, provided as (x,y).
 - **Precomputed relative position:** The precomputed relative position of two selected figures.

Specific Task

Generate a structured output consisting of:

1. **Question:** A clear and specific question about the spatial relationship (literal physical placement) between the two selected sub-images.
2. **Answer:** A precise and accurate response describing the relative positions of the two selected sub-images. Please use the given center points of the given two subfigures. The outputs must focus **only on the spatial relationship between the two selected sub-images**.

Objective

1. **Accuracy:** Ensure all outputs are accurate and based strictly on the provided references for the two selected sub-images.
2. **Clarity:** Write all outputs in a clear, concise, and professional manner.
3. **Relevance:** Focus on the spatial relationship between the two selected sub-images.

Instructions

1. **Use References:** - Base outputs on the provided information for the two selected sub-images, ensuring the **Question** and **Answer** are supported by their details.
2. **Focus on Selected Sub-Images:** - Ensure all outputs focus only on the two selected sub-images.
3. **Answer Separation:** - please rephrase this positional information into a professional and medically accurate QA pair. No need to mention the reason, directly state the position.

Output Format

- **Question:** A specific and clear question about the spatial relationship between the two selected sub-images.
- **Answer:** A precise and accurate response describing the relative positions of the two selected sub-images.

Example

Input:

Compound Image:

- **Index:** 0
- **Caption:** A compound figure of chest CT imaging showing multiple views of a pulmonary nodule.
- **Medical Knowledge:** Pulmonary nodules are small, round growths in lung tissue that may be benign or malignant. Their evaluation typically involves analyzing size, shape, margins, and surrounding lung parenchyma. Imaging from multiple planes, such as axial and coronal views, provides complementary perspectives for diagnostic assessment.

Selected Sub-Images:

- **Indices:** Fig-1, Fig-2
- **Specific Sub-Image Details:**
 - **Index:** Fig-1
 - **Visual Perception Description:** The CT image shows a well-defined, round pulmonary nodule located in the left upper lobe. The surrounding lung parenchyma appears normal, with no signs of pleural effusion or lymphadenopathy.
 - **Bounding Box Information:** (100, 150)
 - **Index:** Fig-2
 - **Caption:** A coronal CT image showing the same pulmonary nodule in the left upper lobe from a different plane.
 - **Visual Perception Description:** The coronal CT image demonstrates the nodule's position relative to the surrounding lung structures, confirming its well-defined borders and normal adjacent lung parenchyma.
 - **Bounding Box Information:** (250, 100)

Output:

- **Question:** What is the spatial relationship between Fig-1 and Fig-2 of the pulmonary nodule in the compound image based on their bounding box locations?
- **Answer:** Fig-1 is positioned to the left and below Fig-2 in the compound image.

Figure 20: The prompt design for Stage 4 of the instruction generation paradigm: Structural Instruction Generation for multi-image VQA (w.r.t. the spatial relationship). This prompt demonstrates how raw positional data of sub-images (from bounding box information) is automatically converted into a structured question-answer pair. As a part of Stage 4 of the instruction generation paradigm, it is designed to create training samples that specifically target the spatial reasoning capabilities.

Prompt for Stage 4: Structural Instruction Generation - Single-image VQA

Role

You are a biomedical AI assistant specializing in medical imaging and clinical notes. Your task is to generate structured and concise **Context-Question-Answer** outputs based on the analysis of compound medical images for clinical education and AI training.

Task

Data Description

The input includes:

1. Compound Image Information:

- **Index:** The unique numeric identifier of the compound image (e.g., Fig-0).
- **Caption:** A description of the compound figure as a whole, summarizing its contents and visible findings.
- **Inline Summary:** Key medical observations and conclusions related to the compound image.
- **Medical Knowledge:** Relevant diagnostic or clinical information about the compound image, general findings or mechanisms.

Specific Task

Generate a structured output consisting of:

1. **Context:** A concise medical background introducing the compound image, based on its **Caption**, **Inline Summary**, and **Medical Knowledge**.
2. **Question:** A clear and specific question that encourages detailed clinical analysis of the compound image.
3. **Answer:** A precise and accurate response, focusing solely on the compound image as a whole. The outputs should focus on the **compound image as a whole**, avoiding sub-image-specific details unless they are critical.

Objective

1. **Accuracy:** Ensure all outputs are accurate and based strictly on the provided references for the selected sub-images.
2. **Clarity:** Ensure the **Question** and **Answer** are logically connected and specific to the compound image.
3. **Relevance:** Focus on the compound image, avoiding interference from other sub-images or the compound image.

Instructions

1. **Use References:** Base outputs on the provided Compound Image Information, ensuring they address the compound image as a whole.
2. **Focus on the Compound Image:** Ensure all outputs provide insights into the compound image's overall findings, clinical importance, or diagnostic significance. Avoid focusing on specific sub-images or their details unless explicitly relevant to the compound image.
3. **Answer Separation:** Ensure the **Answer** provides unique information not repeated in the **Context** or **Question**.

Output Format

- **Context:** Concise medical background for the compound image.
- **Question:** A specific clinical question about the compound image.
- **Answer:** A precise response based on the provided textual information.

Example

Input:

Compound Image:

- **Index:** 0
- **Caption:** A compound figure of chest CT imaging showing multiple views of a pulmonary nodule.
- **Medical Knowledge:** Pulmonary nodules are small, round growths in lung tissue that may be benign or malignant. Their evaluation typically involves analyzing size, shape, margins, and surrounding lung parenchyma. Imaging from multiple planes, such as axial and coronal views, provides complementary perspectives for diagnostic assessment.

Output:

- **Context:** This compound figure includes multiple CT views of a pulmonary nodule in the left upper lobe. Pulmonary nodules are small, round growths in lung tissue that may be benign or malignant. Their evaluation typically involves analyzing size, shape, margins, and the condition of surrounding lung parenchyma. Imaging from multiple planes, such as axial and coronal views, provides detailed information about the nodule's characteristics and its relationship to surrounding structures.
- **Question:** What is the clinical significance of using multiple CT views to evaluate pulmonary nodules in Fig-0?
- **Answer:** Multiple CT views, such as axial and coronal planes, provide complementary perspectives for assessing pulmonary nodules. They help determine the nodule's size, shape, boundaries, and its relationship to nearby tissues. This approach aids in distinguishing between benign and malignant nodules and guides further diagnostic or treatment decisions.

Figure 21: The prompt design for Stage 4 of the instruction generation paradigm: Structural Instruction Generation for single-image VQA. This prompt is designed to convert comprehensive information about an entire compound figure into a structured training sample with Context, Question and Answer. This task specifically trains the model to develop a global and holistic understanding of a multi-panel figure, rather than analyzing its individual components.

Prompt for Stage 4: Structural Instruction Generation - Text-only QA

Role

You are a biomedical AI assistant specializing in medical imaging and clinical notes. Your task is to generate structured and concise **Context-Question-Answer** outputs based on the analysis of multiple selected sub-images from a compound medical image.

Note: The model will **not be shown the image**. All outputs must be based solely on the provided textual information (caption, summary, and medical knowledge).

Task

Data Description

The input includes:

1. Compound Image Information:

- **Index:** The unique numeric identifier of the compound image (e.g., Fig-0).
- **Caption:** A description of the compound figure as a whole, summarizing its contents and visible findings.
- **Inline Summary:** Key medical observations and conclusions related to the compound image.
- **Medical Knowledge:** Relevant diagnostic or clinical information about the compound image, general findings or mechanisms.

Specific Task

Generate a structured output consisting of:

1. **Context:** Summarize the **Caption** and **Inline Summary** into a concise medical background. The **Context** should focus on the key findings and clinical significance described in the input text and **exclude any Medical Knowledge details**.
2. **Question:** Create a **self-contained and insightful question** based on **Medical Knowledge**. The question should relate to the clinical concepts or keywords described in the context, but it must expand the discussion beyond the context by leveraging general medical knowledge.
3. **Answer:** Provide a **precise and detailed response** to the question, expanding on the medical knowledge or principles relevant to the case.

Objective

1. **Accuracy:** Ensure all outputs are medically accurate and based strictly on the provided references. Avoid speculative or unrelated details.
2. **Clarity:** Write in a clear, concise, and professional manner.
3. **Relevance:** Focus on meaningful and clinically significant concepts.
4. **Distinct Context, Question, and Answer:** The **Context** should summarize the **Caption** and **Inline Summary**, while the **Question** and **Answer** should focus on expanding the discussion using **Medical Knowledge**. The **Answer** must introduce new content that is not already present in the **Context**.

Instructions

1. **Summarize for Context:** Use the **Caption** and **Inline Summary** to create a concise, text-based summary. Avoid including any general medical knowledge or broader clinical concepts in this section.
 2. **Expand with the Question:** Identify keywords or concepts in the context and use **Medical Knowledge** to construct an insightful question. Ensure the question is **self-contained** and does not depend on the context to be understood.
 3. **Answer Separately:** Provide a detailed and accurate answer to the question, focusing on relevant medical knowledge. Avoid duplicating any information already present in the context.
 4. **Examples of Good and Bad Practices:**
 - **Good:** What are the common risk factors for pulmonary nodule malignancy, and how are they assessed clinically?
 - **Bad:** What are the key observations in the compound figure?
 - **Bad:** What does the image suggest about the lesion's malignancy?
- Reason:** These imply the model has visual access, which it does not.

Output Format

- **Context:** A concise summary of the **Caption** and **Inline Summary**.
- **Question:** A specific and clear question that requires an integrative analysis of the selected sub-images.
- **Answer:** A precise and detailed response to the question, introducing information not present in the context.

Example

Input:

Compound Image:

- **Index:** 0
- **Caption:** A compound figure of chest CT imaging showing multiple views of a pulmonary nodule.
- **Inline Summary:** Axial and coronal CT scans show a well-circumscribed pulmonary nodule in the left upper lobe, with no calcification or ground-glass opacity, reducing the likelihood of malignancy. Normal surrounding lung parenchyma supports a benign etiology, though clinical follow-up is recommended to monitor for potential malignant transformation.
- **Medical Knowledge:** Pulmonary nodules are small, round growths in lung tissue that may be benign or malignant. Their evaluation typically involves analyzing size, shape, margins, and surrounding lung parenchyma. Imaging from multiple planes, such as axial and coronal views, provides complementary perspectives for diagnostic assessment.

Output:

- **Context:** Pulmonary nodules are small, round growths in lung tissue that may be benign or malignant. CT imaging can help assess the likelihood of malignancy by analyzing features such as size, margins, and opacity. A well-circumscribed nodule with no calcification or ground-glass opacity, as described in this case, is less likely to be malignant.
- **Question:** What are the risk factors for pulmonary nodule malignancy, and what imaging features are most commonly associated with malignant nodules?
- **Answer:** The axial CT image (Fig-1) demonstrates a well-defined, round pulmonary nodule located in the left upper lobe. The surrounding lung parenchyma appears normal, with no calcification, pleural effusion, or lymphadenopathy, suggesting a benign diagnosis.

Figure 22: The prompt design for Stage 4 of the instruction generation paradigm: Structural Instruction Generation for text-only QA. This prompt is designed to create a training sample based solely on textual information, as the model does not receive any image input for this task. This task specifically targets the linguistic and knowledge-based reasoning skills, ensuring it can perform accurate medical Q&A even without visual aids.

Prompt for Stage 4: Structural Instruction Generation - Multiple-choice VQA

Role

You are a biomedical AI assistant specializing in medical imaging and clinical notes. Your task is to generate structured and concise Four-Choice Question-Answer outputs based on the analysis of medical images for clinical education and AI training.

Task

Data Description

The input includes:

1. Compound Image Information:

- **Index:** The unique numeric identifier of the compound image (e.g., Fig-0).
- **Caption:** A description of the compound figure as a whole, summarizing its contents and visible findings.
- **Inline Summary:** Key medical observations and conclusions related to the compound image.
- **Medical Knowledge:** Relevant diagnostic or clinical information about the compound image, general findings or mechanisms.

Specific Task

Generate a structured output consisting of:

1. **Question:** A clear and specific four-choice question about the compound image or selected sub-image(s) that encourages detailed analysis and clinical reasoning.
2. **Options:** Four plausible answer options, with one correct option and three incorrect distractors. The distractors should be relevant but clearly incorrect based on the input data.

Objective

1. **Accuracy:** Ensure all outputs are medically accurate and based strictly on the provided references for the compound image or sub-image(s).
2. **Clarity:** Write all outputs in a clear, concise, and professional manner.
3. **Relevance:** Focus on clinically significant findings or diagnostic insights from the input data.
4. **Option Plausibility:** Ensure the distractors are relevant to the context and plausible enough to challenge the reader, but clearly incorrect upon analysis. The correct answer will have an equal 25% probability of distribution across A, B, C, or D in the Options list.

Instructions

1. **Use References:** Base outputs on the provided information for the compound image or sub-image(s), ensuring the **Question**, and **Options** are supported by their details.
2. **Focus on Clinical Value:** Ensure the question and options are clinically relevant and encourage critical thinking about the image data.
3. **Create Plausible Distractors:** Distractors should be related to the same clinical or diagnostic context but be clearly incorrect upon closer analysis.

Output Format

- **Question:** A specific and clear four-choice question that encourages detailed analysis and clinical reasoning.
- **Options:** Four plausible answer options, with one correct answer and three incorrect distractors.
- **Correct Answer:** Indicate the correct option (e.g., A, B, C, or D).

Example

Input:

Compound Image:

- **Index:** 0
- **Caption:** A compound figure of chest CT imaging showing multiple views of a pulmonary nodule.
- **Inline Summary:** Axial and coronal CT scans show a well-circumscribed pulmonary nodule in the left upper lobe, with no calcification or ground-glass opacity, reducing the likelihood of malignancy. Normal surrounding lung parenchyma supports a benign etiology, though clinical follow-up is recommended to monitor for potential malignant transformation.
- **Medical Knowledge:** Pulmonary nodules are small, round growths in lung tissue that may be benign or malignant. Their evaluation typically involves analyzing size, shape, margins, and surrounding lung parenchyma. Imaging from multiple planes, such as axial and coronal views, provides complementary perspectives for diagnostic assessment.

Output:

- **Question:** Based on Fig-1, which of the following features most strongly supports a benign diagnosis for the pulmonary nodule?
- **Options:** A. Presence of ground-glass opacity.
B. Smooth, well-defined borders.
C. Enlarged mediastinal lymph nodes.
D. Irregular shape with spiculated margins.
- **Correct Answer:** B

Figure 23: The prompt design for Stage 4 of the instruction generation paradigm: Structural Instruction Generation for multi-choice VQA. This prompt is designed to convert information about a compound figure into a structured multi-choice question. It specifically emphasizes the need to generate three *plausible distractors* to ensure the quality and challenge of the questions, thereby creating data suitable for both training and standardized evaluation.

Prompt for Stage 5: Leakage-Prevented Context Refinement

Role

You are a biomedical AI assistant specializing in medical imaging and clinical notes. Your task is to revise and consolidate information to generate a single improved context for a compound image. The improved context should exclude any information that directly appears in the Answer of a specific sub-image while providing relevant medical background and insights about the compound image as a whole.

Task

Data Description

The input includes:

1. Compound Image Information:

- **Index:** The unique numeric identifier of the compound image (e.g., Fig-0).
- **Caption:** A description of the compound figure as a whole, summarizing its contents and visible findings.
- **Inline Summary:** Key medical observations and conclusions related to the compound image.
- **Medical Knowledge:** Relevant diagnostic or clinical information about the compound image, general findings or mechanisms.
- **Visual Perception Description:** Highlights visible findings in the compound image.

2. Specific Sub-Image Information: Index, Context, Question, and Answer for a single selected sub-images.

Specific Task

Generate an improved comprehensive context for the compound image by using:

- The compound image's Inline Summary, Medical Knowledge, and Visual Perception Description.
- The information of the selected sub-image.

Requirement: Exclude any content that directly appears in the Answer of the specific sub-image.

Objective

1. **Accuracy:** Ensure the improved context is medically accurate and avoids including any information from the Answer of the sub-image.
2. **Clarity:** Keep the consolidated context concise, clear, and professionally written.
3. **Relevance:** Focus on overall findings, mechanisms, or clinical significance of the compound image without introducing unrelated or speculative details.
4. **Answer Separation:** Ensure the improved context does not leak any information that directly appears in the Answer of the sub-image.

Instructions

1. **Use References:** Base the improved context on the provided sub-image contexts, but avoid including any information that overlaps directly with the Answer of the sub-image. Consolidate relevant medical background or broader insights related to the compound image.
2. **Remove Answer Information:** Exclude any content from the sub-image context that overlaps with the sub-image Answer. Provide general information about the compound image without revealing specific details from the Answer.
3. **Focus on Compound Image:** Ensure the improved context provides information that is relevant to the compound image as a whole.
4. **Ensure Clarity and Conciseness:** Write the improved context in a clear, concise, and professional manner, adhering to medical accuracy and educational value.

Output Format

- **Improved Context:** A consolidated and comprehensive context for the compound image, excluding information that overlaps with the Answer of the sub-image.

Example

Input

Compound Image:

- **Index:** 0
- **Caption:** A compound figure of chest CT imaging showing multiple views of a pulmonary nodule.
- **Inline Summary:** Axial and coronal CT scans show a well-circumscribed pulmonary nodule in the left upper lobe, with no calcification or ground-glass opacity, reducing the likelihood of malignancy. Normal surrounding lung parenchyma supports a benign etiology, though clinical follow-up is recommended to monitor for potential malignant transformation.
- **Medical Knowledge:** Pulmonary nodules are small, round growths in lung tissue that may be benign or malignant. Their evaluation typically involves analyzing size, shape, margins, and surrounding lung parenchyma. Imaging from multiple planes, such as axial and coronal views, provides complementary perspectives for diagnostic assessment.
- **Visual Perception Description:** The compound figure contains multiple sub-images showing different CT views of a solitary pulmonary nodule in the left upper lobe. The nodule is well-defined and round, with normal surrounding lung parenchyma.

Specific Sub-Image:

- **Index:** 1
- **Context:** This axial CT image depicts a solitary pulmonary nodule in the left upper lobe. Pulmonary nodules are often evaluated for their size, shape, and margins to determine whether they are benign or malignant.
- **Question:** What are the key visible features of the pulmonary nodule in (Fig-1)?
- **Answer:** The image shows a well-defined, round pulmonary nodule located in the left upper lobe. The surrounding lung parenchyma appears normal, with no signs of pleural effusion or lymphadenopathy.

Output

- **Improved Context:** This compound figure includes multiple CT views of a pulmonary nodule located in the left upper lobe. Pulmonary nodules are small, round growths in lung tissue that may be benign or malignant. Their evaluation typically involves analyzing their size, shape, borders, and the condition of surrounding lung parenchyma. Different imaging planes, such as axial and coronal views, provide complementary perspectives on the nodule's characteristics and its relationship to surrounding structures, aiding in diagnostic assessment.

Figure 24: The prompt design for Stage 5 of the instruction generation paradigm: Leakage-Prevented Context Refinement. This prompt guides the AI model to revise the generated context by removing specific details that overlap with the answer for a particular sub-image. This ensures the context provides general background without inadvertently revealing the solution, thereby enhancing the quality and challenge of the training instructions.

Prompt for LLM-as-a-judge

Role

You are an impartial medical QA evaluator.

Task

Given a **Question**, a **Reference Answer** (ground truth), and **two model predictions (A and B)**, determine which prediction is **more aligned** with the reference.

Your evaluation should prioritize the following aspects:

1. Correctness:

- Whether the prediction is factually correct and consistent with the reference answer.
- Whether it avoids false or misleading medical statements.

2. Completeness:

- Whether the prediction covers the essential clinical points required to match the reference.

3. Clarity:

- Whether the prediction is understandable and unambiguous.

If A and B are **equally correct and complete**, return "Tie".

Do **not** judge writing style, verbosity, or elegance beyond clarity.

Focus strictly on factual and clinical alignment with the reference answer.

Output

Return **ONLY** valid JSON using **exactly** this format:

```
{  
    "winner": "A" | "B" | "Tie",  
    "rationale": "Short explanation (1–3 sentences) identifying correctness differences."  
}
```

Figure 25: The LLM-as-a-judge evaluation prompt for pairwise medical QA comparison. The figure illustrates the structured prompt used to evaluate two model predictions against a reference answer in medical visual question answering. The prompt instructs the judge model to prioritize factual correctness, with completeness and clarity as secondary criteria, and to output a strict JSON object specifying the winner and a brief rationale. This standardized judging prompt enables consistent and reproducible comparison.

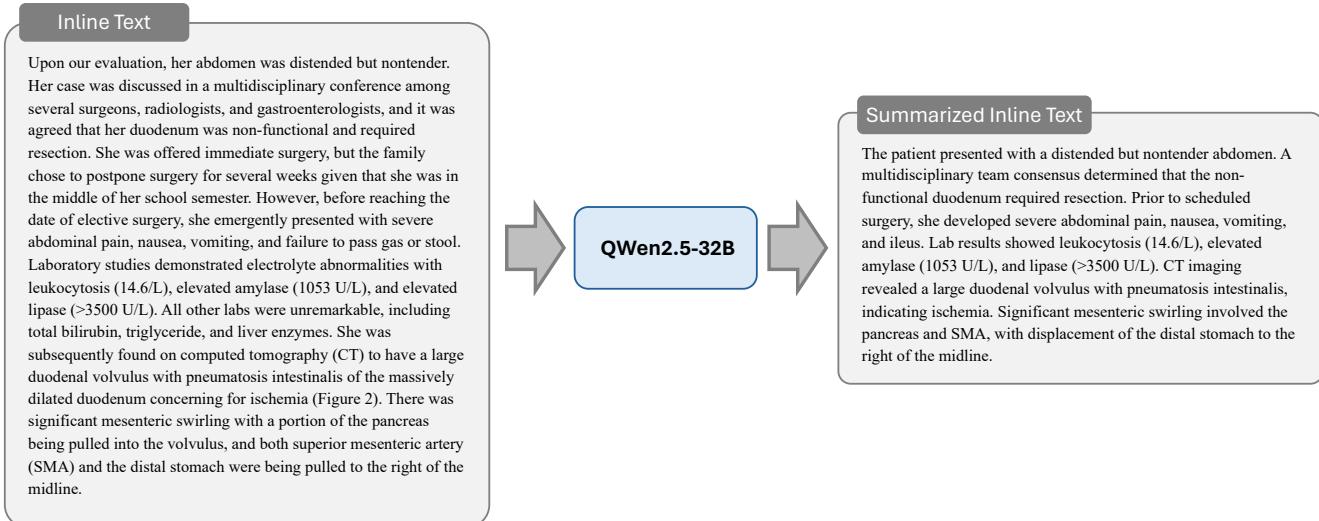


Figure 26: Example of the Stage 1: Inline Text Summarization process. This figure illustrates the transformation of a raw, complex block of inline text (left) into a refined and more structured summary (right). This initial stage of the instruction generation paradigm is designed to standardize the textual input for subsequent automated processing.

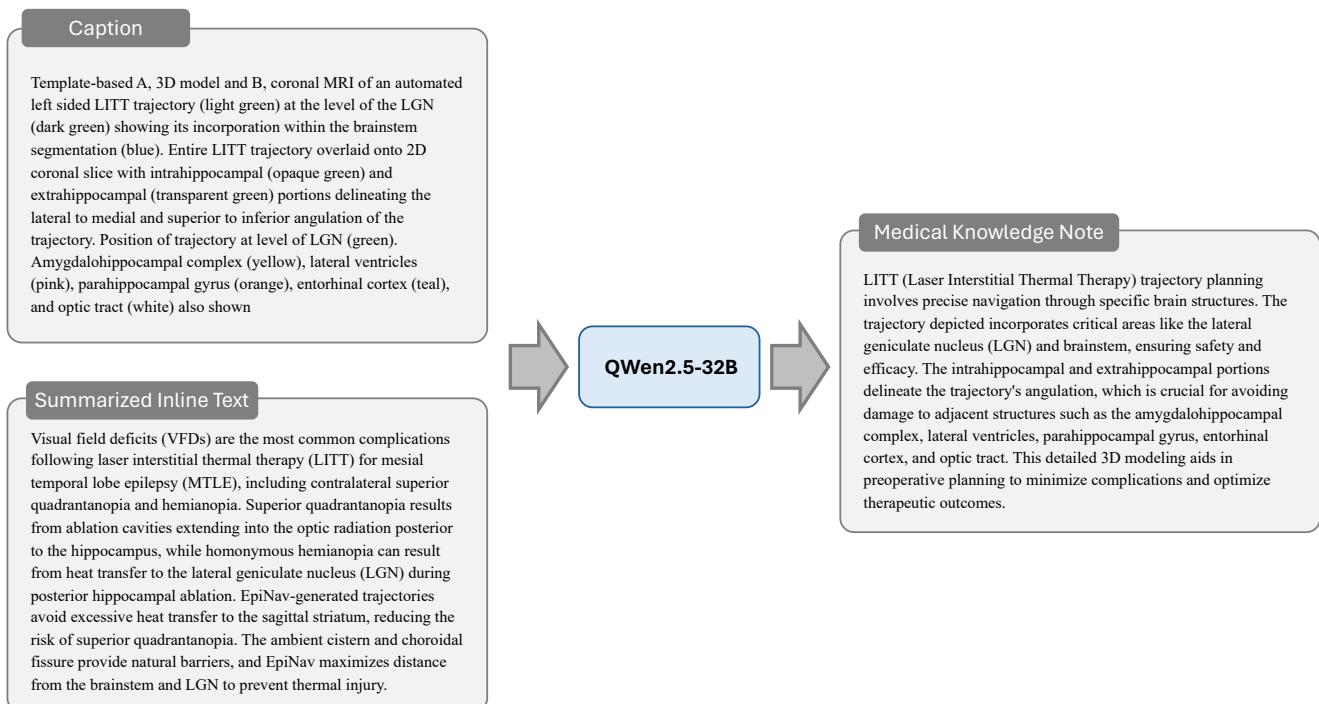


Figure 27: Example of the Stage 2: Medical Knowledge Complementation process. This figure illustrates how an explanatory knowledge text (right) is automatically distilled and generated from a descriptive caption and summary (left). This process enriches the dataset with critical medical background information to support deeper reasoning tasks.

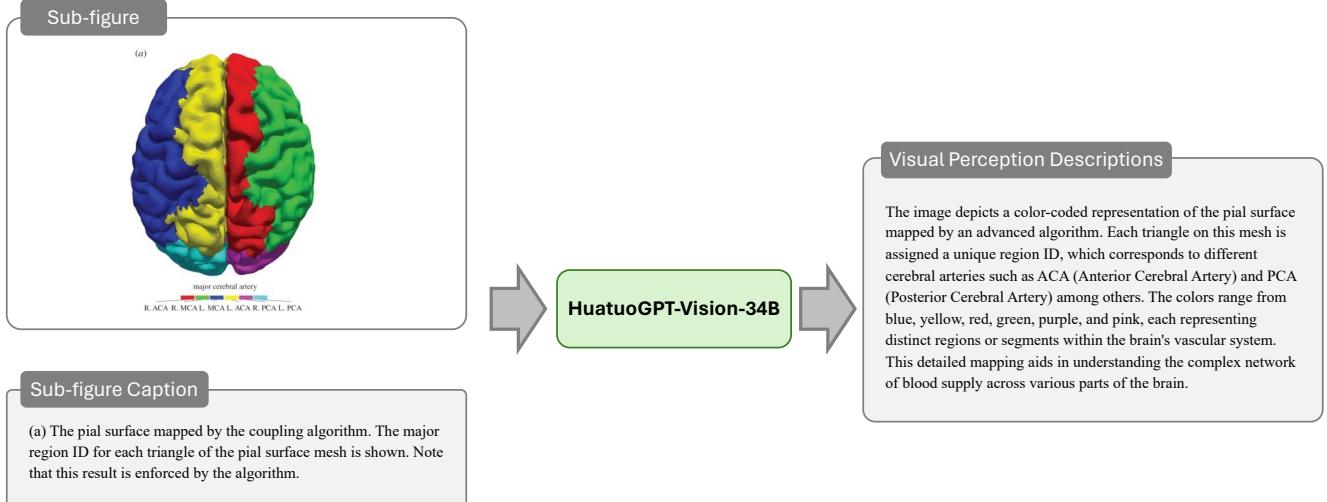


Figure 28: Example of the Stage 3: Medical Visual Perception Enhancement process. This figure illustrates how a multi-modal model is used to transform a sub-image and its basic textual descriptions (left) into a richer and more detailed visual analysis text (right). This process provides a deeper semantic interpretation of the visual features within the image.

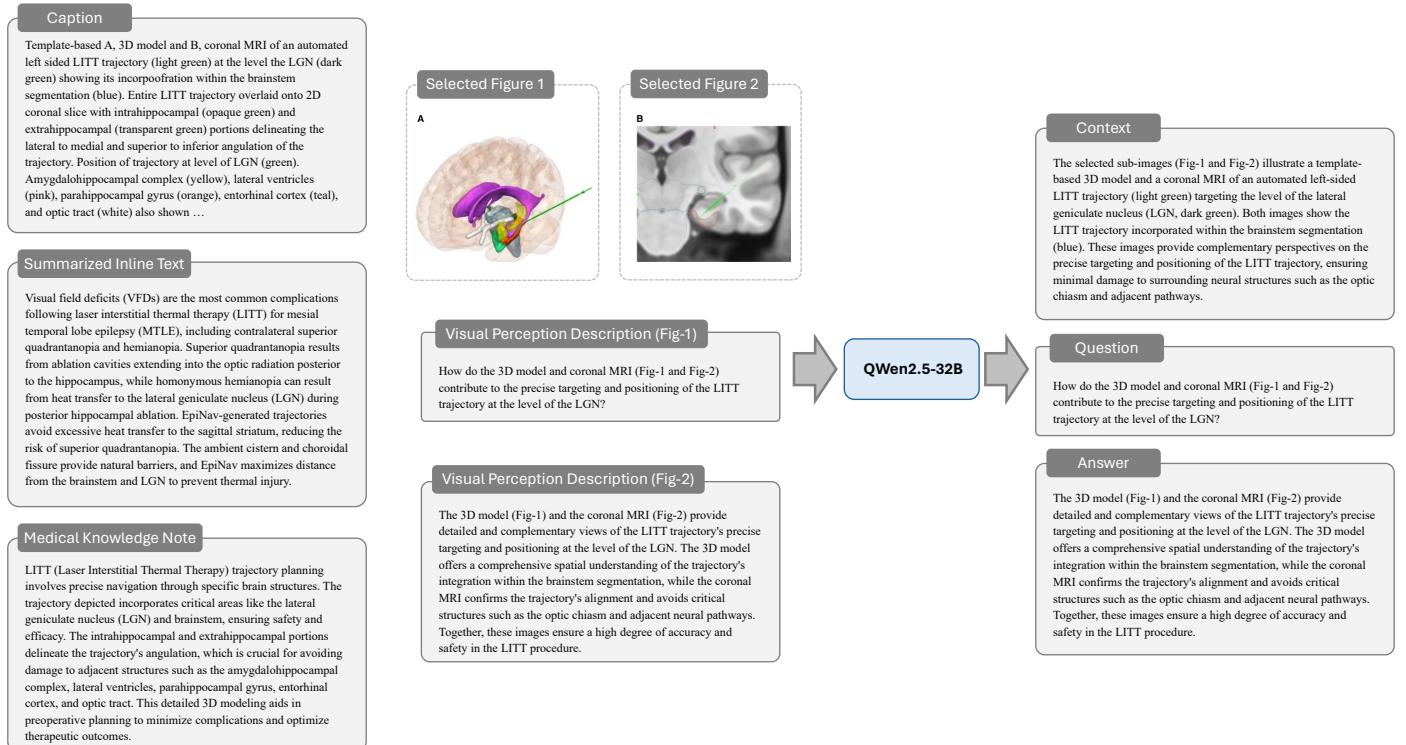


Figure 29: Example of the Stage 4: Structural Instruction Generation process. This figure demonstrates how all the information generated in the preceding three stages (left) is integrated to produce a final, structured *Context-Question-Answer* instruction (right) for model training. This is the key step where processed data is converted into a final training sample.

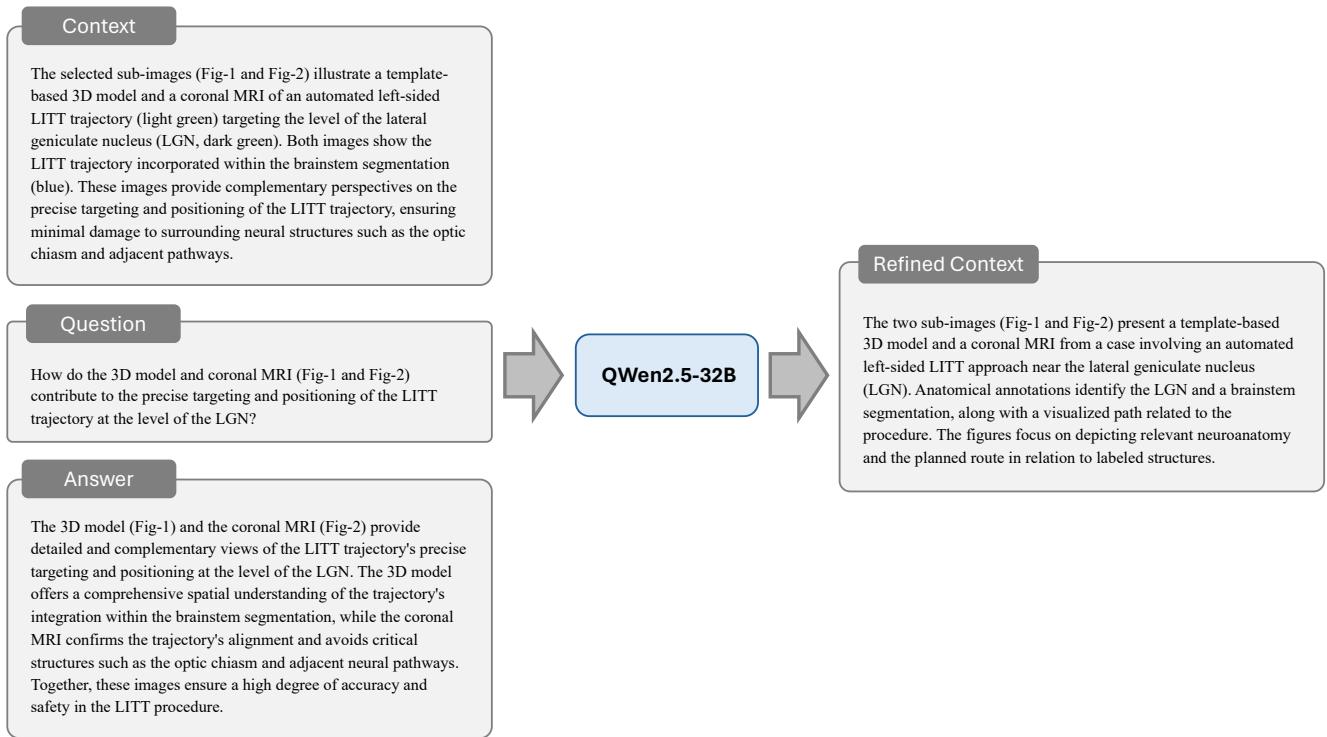


Figure 30: Example of the Stage 5: Leakage-Prevented Context Refinement process. This figure demonstrates the final quality control step. The model reviews the auto-generated context (left) and removes specific details that might leak the answer, producing a more neutral and challenging refined context (right). This process is designed to prevent the model from learning to exploit "cheating" cues in the context, thereby improving the quality of the training.