MedLayBench-V: A Large-Scale Benchmark for Expert-Lay Semantic Alignment in Medical Vision Language Models

Han Jang[♣,△,♠,†], **Junhyeok Lee**[♣,♡,♠,†], **Heeseong Eum**[♣,♡,♠], **Kyu Sung Choi**[♣,♡,△,◇,♠,*]

[♣]Seoul National University

[♡]Seoul National University College of Medicine

[△]Department of Radiology, Seoul National University Hospital

[◇]Healthcare AI Research Institute, Seoul National University Hospital

[♠]The Advanced Imaging and Computational Neuroimaging (AICON) Laboratory

{hanjang, jhlee0619, seong6466}@snu.ac.kr, ent1127@snu.ac.kr

Project Page Code Dataset

Abstract

Medical Vision-Language Models (Med-VLMs) have achieved expert-level proficiency in interpreting diagnostic imaging. However, current models are predominantly trained on professional literature, limiting their ability to communicate findings in the lay register required for patient-centered care. While text-centric research has actively developed resources for simplifying medical jargon, there is a critical absence of large-scale multimodal benchmarks designed to facilitate lay-accessible medical image understanding. To bridge this resource gap, we introduce **MedLayBench-V**, the first large-scale multimodal benchmark dedicated to expert-lay semantic alignment. Unlike naive simplification approaches that risk hallucination, our dataset is constructed via a Structured Concept-Grounded Refinement (SCGR) pipeline. This method enforces strict semantic equivalence by integrating Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) with micro-level entity constraints. MedLayBench-V provides a verified foundation for training and evaluating next-generation Med-VLMs capable of bridging the communication divide between clinical experts and patients.

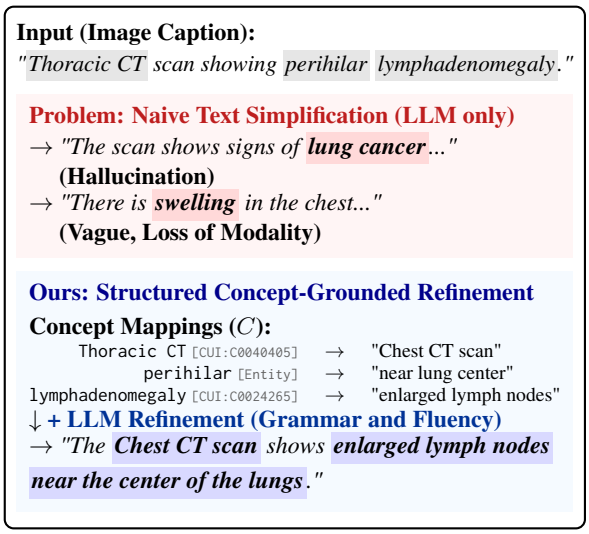


Figure 1: **Motivation.** Our method prevents hallucinations by enforcing Structured Constraints: It explicitly maps extracted Concepts and Entities (e.g., *lymphadenomegaly*) to lay terms, ensuring diagnostic accuracy while preserving specific details.

1 Introduction

Enhancing the linguistic accessibility of clinical documentation has emerged as a paramount objective in biomedical Natural Language Processing (NLP). Driven by the imperative to facilitate patient-centered care, recent research has coalesced around tasks such as Biomedical Lay Summarization (BioLaySumm) and Neural Text Simplification (NTS) (Shardlow and Nawaz, 2019; Yao et al., 2024). Collectively framed as Medical Lay Language Generation (MLLG), these efforts aim to translate highly specialized medical jargon into the accessible lay register. This paradigm shift is epitomized by initiatives like the BioLaySumm shared tasks (Xiao et al., 2025; Goldsack et al., 2024) and recent benchmarks like MedAgentBoard (Zhu et al., 2025a), where MLLG is established as a core competency for medical artificial intelligence (AI). Recent studies attribute success in this domain to the advanced semantic reasoning of Large Language Models (LLMs), which allows them to modify lexical complexity while maintaining semantic invariance, thereby ensuring that core medical facts are preserved despite the stylistic shift (Liao et al., 2025).

While the text-to-text simplification landscape has advanced significantly, the integration of this lay perspective into multimodal systems remains an open challenge. Medical Vision-Language Models (Med-VLMs), such as those trained on RO-

[†]These authors contributed equally to this work.

[*]Corresponding author: ent1127@snu.ac.kr

Findings of the Association for Computational Linguistics: ACL 2026, pages 18375–18394

COv2 (Rückert et al., 2024) or PMC-OA, have achieved expert-level proficiency in interpreting diagnostic imaging (Lozano et al., 2025). However, a critical limitation persists in their current training paradigm. Unlike text-centric LLMs that are becoming increasingly adaptable to the lay register, current Med-VLMs are predominantly optimized for the rigid clinical jargon found in professional literature. As illustrated in Figure 1, this domain-specific optimization creates a significant barrier to usability; while models successfully encode visual features into technical tokens like 'Pneumothorax', their ability to ground the same visual evidence in natural language equivalents like 'Collapsed lung' remains unsupported due to the lack of parallel lay data. This suggests that without a dedicated benchmark to facilitate expert-to-lay alignment, Med-VLMs will remain confined to a specialized lexicon, severely limiting their applicability in patient-centered care.

Overcoming this resource scarcity, however, presents significant methodological challenges. Existing multimodal benchmarks are exclusively populated with expert-level reports and offer no ground truth for lay-accessible descriptions. Furthermore, relying on standard lexical metrics like BLEU (Papineni et al., 2002) is insufficient for validation as they inherently penalize the vocabulary shifts required for simplification (Zhao et al., 2024a). Moreover, constructing a benchmark via naive LLM generation carries the risk of hallucination or the omission of vital quantitative details, which compromises the factual integrity required for medical AI (Liao et al., 2025).

To bridge this divide, we introduce **MedLayBench-V**, the first multimodal benchmark designed to facilitate patient-centric medical image understanding. Drawing inspiration from recent text-centric approaches that leverage structured medical knowledge to enhance summary relevance (Ming et al., 2025), we extend this philosophy to the multimodal domain via a novel **Structured Concept-Grounded Refinement (SCGR)** pipeline. Our approach synergizes macro-level conceptual mapping from the Unified Medical Language System (UMLS) with micro-level entity constraints extracted via Named Entity Recognition (NER) (Bodenreider, 2004). This hybrid strategy ensures that the generated lay captions maintain strict semantic equivalence with the original expert reports while effectively transitioning to the lay register. Using this verified dataset, we establish the first comprehensive baselines for expert-lay alignment, providing a standardized foundation for future research in accessible medical AI.

Our contributions are summarized as follows:

- To the best of our knowledge, we introduce **MedLayBench-V**, the first foundational benchmark encompassing diverse medical imaging modalities specifically curated to bridge the linguistic divide between clinical experts and laypersons.
- We propose the SCGR pipeline, a verifiable framework that extends knowledge-guided text simplification principles to vision-language tasks, ensuring high clinical correctness and hallucination control.
- We establish a comprehensive evaluation protocol for Expert-Lay semantic alignment and provide standardized baselines, offering a robust foundation for future research in patient-centered medical AI.

2 Related Works

2.1 Patient-Centered Clinical Reporting

The complexity of medical documentation creates significant barriers to patient understanding, driving the need for automated systems that can translate clinical narratives into accessible language. To address this, the field has evolved from early Neural Text Simplification (NTS) efforts into the broader paradigm of Medical Lay Language Generation (MLLG) (Shardlow and Nawaz, 2019; Yao et al., 2024). This transition is marked by large-scale community initiatives such as the BioLaySumm shared tasks and the MedAgentBoard benchmark, which provide standardized tasks to bridge the communication gap between experts and laypersons (Xiao et al., 2025; Zhu et al., 2025a).

Within this text-centric landscape, LLMs have achieved remarkable proficiency, effectively balancing lexical simplification with semantic invariance as demonstrated by frameworks (Liao et al., 2025). However, this progress has yet to permeate the multimodal domain. Unlike the thriving domain for text-only models, there is a critical absence of benchmarks designed to evaluate Med-VLMs leaving it unclear whether current SOTA models can successfully ground visual findings in lay-accessible language without compromising factual accuracy.

2.2 Medical Vision-Language Models and Dataset Scarcity

In the multimodal domain, Med-VLMs have achieved expert-level proficiency in interpreting diagnostic imaging (Zhang et al., 2023; Li et al., 2023; Sellergren et al., 2025). These capabilities are predominantly driven by large-scale datasets such as ROCOv2 (Rückert et al., 2024) and BIOMEDICA (Lozano et al., 2025). However, these datasets are exclusively curated from professional biomedical literature, thereby optimizing models strictly for the rigid clinical jargon.

A critical limitation in existing multimodal datasets is the scarcity of parallel multimodal data that pairs medical images with patient-friendly descriptions. While models can successfully align visual features with technical concepts (e.g., "Pneumothorax"), the lack of ground truth for natural language equivalents (e.g., "Collapsed lung") prevents them from learning the lay register. Unlike the text domain where lay benchmarks exist, the vision-language field suffers from this fundamental resource gap, which hinders the development of expert-lay alignment capabilities in VLMs.

2.3 Limitations of Current Benchmarks

To bridge the expert-lay divide, prior research has predominantly focused on text-to-text simplification strategies. Early approaches relied on rule-based methods or phrase tables to substitute medical jargon with simpler synonyms (Shardlow and Nawaz, 2019). With the advent of LLMs, recent studies have shifted towards generative rewriting, employing models such as GPT-4o to translate clinical notes into patient-friendly language (Yao et al., 2024). However, LLMs frequently generate plausible yet factually incorrect descriptions or omit vital quantitative details to satisfy readability constraints, thereby compromising patient safety in clinical settings (Moor et al., 2023; Zhu et al., 2025b). For instance, a recent prospective trial demonstrated that while LLM-based simplification significantly reduces cognitive workload, it introduced factual errors and omissions in approximately 6–7% of reports, necessitating rigorous verification mechanisms (Prucker et al., 2025).

Recent initiatives, such as the BioLaySumm 2025 Shared Task (Goldsack et al., 2022; Xiao et al., 2025) and Layman's RRG (Zhao et al., 2024a), have begun to incorporate visual modalities to address these grounding issues. Despite these advances, current multimodal benchmarks remain limited in scope, predominantly focusing on specific modalities like Chest X-rays (CXR) with restricted dataset sizes. Furthermore, these datasets typically rely on end-to-end LLM generation for creating lay captions, which can perpetuate the very hallucinations they aim to resolve without rigorous concept-level verification. To facilitate the training of robust, general-purpose Med-VLMs, there is a critical need for a large-scale, diverse benchmark that extends beyond single modalities.

2.4 Evaluation Metrics for Medical Text Generation

Evaluating the quality of MLLG systems remains a persistent challenge due to the inadequacy of existing metrics. Traditional n-gram based metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) measure surface-level overlap. However, they inherently penalize the vocabulary shifts required for simplification, making them unsuitable for expert-to-lay translation tasks (Zhao et al., 2024a; Zhang et al., 2019). Conversely, medically-oriented metrics like Green (Ostmeier et al., 2024) and RaTEScore (Zhao et al., 2024b) focus on clinical factuality and entity extraction.

While effective for expert reports, they do not assess whether the generated text is understandable to a lay audience. Finally, standard readability metrics rely on heuristic formulas (e.g., sentence length) rather than actual comprehensibility, often failing to capture the semantic nuances required for patient education (Yao et al., 2024). Therefore, effective MLLG evaluation requires a comprehensive framework that simultaneously assesses visual grounding, factual correctness, and lay accessibility. However, performing such multi-dimensional evaluation is unfeasible with current VLM datasets due to the critical absence of lay-aligned references. To bridge this gap, we introduce MedLayBench-V, a unified benchmark designed to facilitate this holistic evaluation.

3 Methodology

We introduce **MedLayBench-V**, a large-scale multimodal benchmark designed to bridge the gap between expert clinical jargon and patient-accessible language. To ensure the high semantic fidelity of this benchmark, we propose the Structured Concept-Grounded Refinement (SCGR) pipeline.

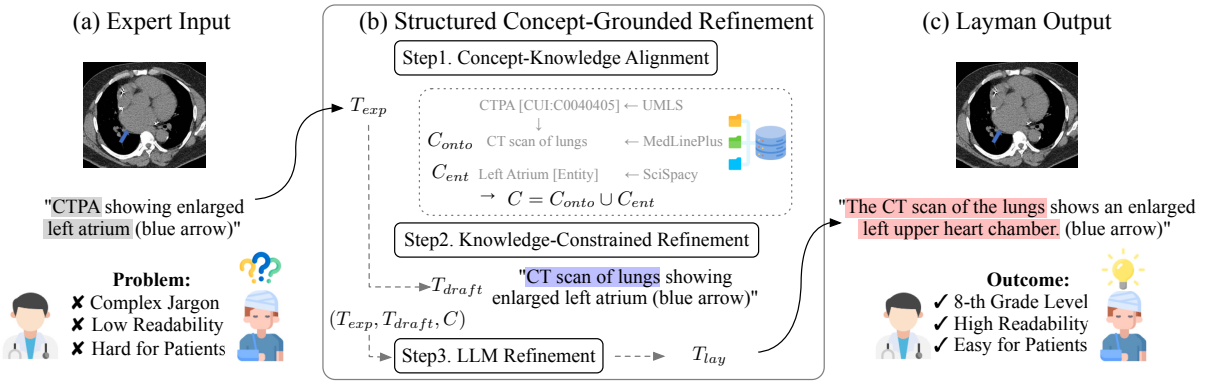


Figure 2: **Overview of the SCGR Framework.** (a) Expert Input extracts technical concepts from the initial jargon-heavy reports. (b) Structured Concept-Grounded Refinement maps terms to lay definitions and employs Llama-3.1-8B (Dubey et al., 2024) to synthesize the final caption, optimizing for syntax and fluency while strictly adhering to factual constraints (Detailed prompt in Appendix A). (c) Layman Output provides a clinically accurate and accessible description.

Crucially, our framework explicitly decouples semantic extraction from stylistic refinement. This separation ensures strict *Semantic Equivalence* between the expert and lay registers, mitigating the hallucinations common in end-to-end generation. The pipeline consists of three distinct stages, **corresponding to Steps 1–3 in Figure 2(b)**: (i) Concept-Knowledge Alignment, (ii) Knowledge-Constrained Refinement, and (iii) LLM Refinement.

3.1 Data Source and Task Definition

We utilize the ROCOv2 dataset (Rückert et al., 2024)[1] as our seed corpus. Derived from the PubMed Central Open Access (PMC-OA) subset (Lin et al., 2023)[2], ROCOv2 is uniquely advantageous for our task as it provides not only diagnostic captions (T_{exp}) but also pre-computed UMLS Concept Unique Identifiers (CUIs) extracted via the MedCAT toolkit (Kraljevic et al., 2021)[3]. These existing annotations serve as a critical foundation for our semantic extraction phase.

Despite the richness of this clinical metadata, the expert descriptions in ROCOv2 remain inherently unintelligible to non-specialists. Our objective is to augment these pairs with layman-accessible descriptions (T_{lay}), creating the first dual-register medical benchmark optimized for patient-centric VLM training and testing.

[1] https://huggingface.co/datasets/eltorio/ROCOv2-radiology

[2] https://pmc.ncbi.nlm.nih.gov/tools/openftlist/

[3] https://github.com/CogStack/MedCAT2

3.2 Concept-Knowledge Alignment

To guarantee that the simplified captions retain the diagnostic precision of T_{exp}, we first extract a set of semantic constraints C. This process integrates high-level ontology mapping with fine-grained entity recognition.

Ontology-Based CUI Mapping. We utilize the UMLS Metathesaurus API (Bodenreider, 2004)[4] to ground clinical terms to CUIs. In contrast to heuristic string matching, direct API querying guarantees precise alignment with standard medical ontologies. This step captures core medical concepts (e.g., `C0040405` → "CTPA"). We denote the set of identified CUIs as C_{onto}, ensuring that the pathology is rigorously anchored to standardized terminology.

Fine-Grained Entity Extraction. We supplement CUIs with a biomedical Named Entity Recognition (NER) model, SciSpacy (Neumann et al., 2019)[5]. This module explicitly extracts quantitative attributes (e.g., lesion sizes) and spatial descriptors (C_{ent}) often missed by high-level mapping. We integrate these two sources to establish the final semantic constraint set C. Formally, this is defined as:

$$C = C_{onto} \cup C_{ent} \tag{1}$$

where C_{onto} represents the high-level ontological constraints anchored to UMLS, and C_{ent} denotes the fine-grained entity constraints extracted via NER.

[4] https://www.nlm.nih.gov/research/umls/

[5] https://allenai.github.io/scispacy/

Table 1: **Linguistic Complexity and Readability Analysis.** Our refinement consistently reduces reading difficulty, improves accessibility, and standardizes vocabulary across the entire dataset.

Linguistic Metric	Train Set ($N = 59,962$)		Validation Set ($N = 9,904$)		Test Set ($N = 9,927$)		Overall (Total) ($N = 79,793$)	
	Expert	**Layman**	**Expert**	**Layman**	**Expert**	**Layman**	**Expert**	**Layman**
Readability Metrics								
FKGL (Kincaid et al., 1975) ↓	13.05	**10.29**	13.29	**10.50**	13.21	**10.53**	13.10	**10.35**
CLI (Coleman and Liau, 1975) ↓	15.73	**9.82**	16.12	**10.06**	16.02	**10.04**	15.82	**9.88**
DCRS (Dale and Chall, 1948) ↓	14.02	**11.73**	14.09	**11.80**	14.02	**11.77**	14.03	**11.74**
SMOG (Mc Laughlin, 1969) ↓	13.71	**12.21**	13.85	**12.35**	13.88	**12.41**	13.75	**12.25**
FRE (Flesch, 1948) ↑	26.44	**56.15**	24.85	**55.00**	25.64	**55.09**	26.14	**55.88**
Lexical Statistics								
Average Sentence Length	23.73	27.81	25.45	28.86	25.61	29.34	24.17	28.13
Vocab Size ↓	36,875	**20,589**	14,877	**9,191**	14,865	**9,238**	44,673	**24,085**

3.3 Knowledge-Constrained Refinement

Leveraging the semantic constraint set C, we synthesize the lay caption T_{lay}. This phase shifts the linguistic register while strictly adhering to the extracted medical facts.

Lexical Alignment and Draft Synthesis. For each concept in C_{onto}, we retrieve patient-friendly definitions by querying the MedlinePlus vocabulary within the UMLS Metathesaurus. Curated by the National Library of Medicine (NLM), MedlinePlus serves as the authoritative bridge between rigorous clinical ontologies and public health literacy (Miller et al., 2000)[6]. By aligning UMLS CUIs directly with MedlinePlus definitions, we ensure that the terminology is not merely simplified but standardized to a trusted lay register. We then construct an intermediate noisy lay draft (T_{draft}) via deterministic dictionary-based substitution. While grammatically noisy, T_{draft} serves as a reliable lexical basis for the subsequent refinement.

Constraint-Guided Linguistic Refinement. To generate the final accessible caption, we employ Llama-3.1-8B-Instruct (Dubey et al., 2024) within a constrained generation framework. We chose Llama-3.1-8B-Instruct for this stage due to its open-weight reproducibility, computational practicality for processing approximately 80K samples, and strong instruction-following capability for constrained text refinement. Since the structured constraints are responsible for preserving semantic fidelity, the LLM's role is limited to grammar and fluency optimization, which does not require a larger model or domain-specific medical knowledge. Our structured prompt incorporates: (1) the source text T_{exp} ensuring factual grounding, (2) a strict constraint set C for hallucination mitigation, and (3) the initial draft T_{draft} to steer vocabulary selection. The objective is to downscale linguistic complexity from a college-level register to a high school level, ensuring the output remains semantically faithful to the clinical findings through explicit constraints. Figure 3 demonstrates qualitative examples of our refinement across different modalities.

[6] https://medlineplus.gov/

Table 2: **Dataset Statistics and Quality Consistency.** We report consistency across Train (N=59,962), Validation (N=9,904), and Test (N=9,927). The **Overall** column represents the weighted average (N=79,793). High clinical correctness (RaTEScore, GREEN) and consistent simplification scores (LENS) across all splits confirm the robust quality of our refinement pipeline.

Metric	Train	Val	Test	Overall
Relevance				
BLEU-4 (Papineni et al., 2002) ↑	20.99	22.32	22.45	**21.34**
ROUGE-L (Lin, 2004) ↑	49.33	50.13	50.40	**49.56**
METEOR (Banerjee and Lavie, 2005) ↑	53.00	53.40	53.56	**53.12**
Readability				
LENS (Maddela et al., 2023) ↑	63.28	62.91	62.94	**63.19**
Radiological Factuality				
RaTEScore (Zhao et al., 2024b) ↑	64.66	64.57	65.09	**64.70**
GREEN (Ostmeier et al., 2024) ↑	69.03	70.14	70.03	**69.29**

4 Experiments

We demonstrate the value of **MedLayBench-V** through a comprehensive analysis of its linguistic properties and quality consistency, followed by a zero-shot downstream benchmark to evaluate current VLMs' capability in handling both expert and layman medical concepts.

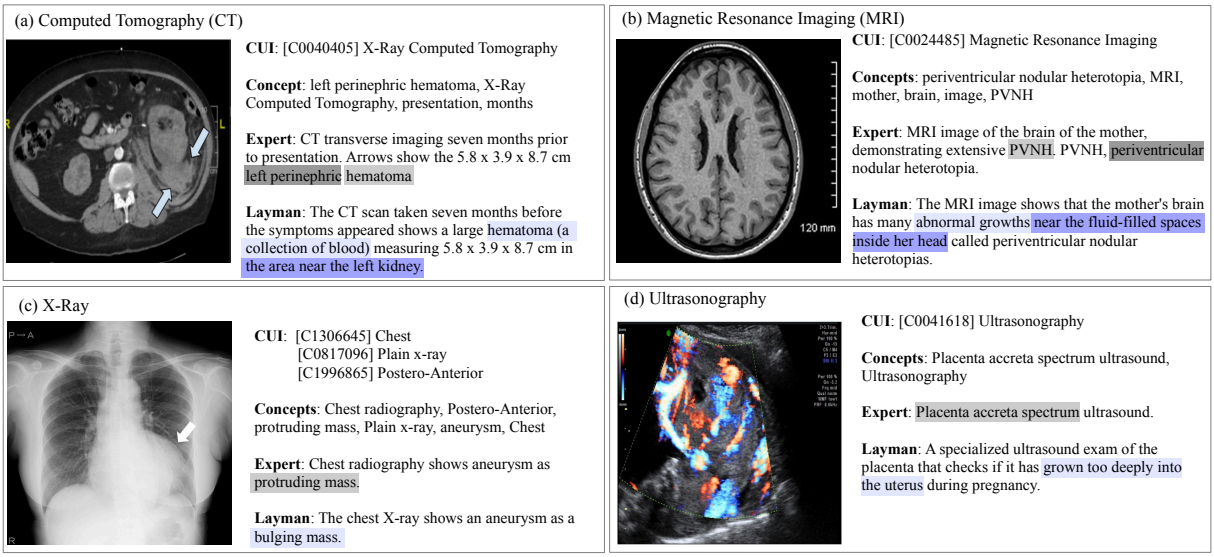


Figure 3: **Qualitative Comparison of Jargon Refinement across Modalities.** The figure illustrates example cases from CT, MRI, X-Ray, and Ultrasound. Highlights indicate the transformation from medical jargon (Original expert-level caption) to patient-friendly language (Layman-level caption). Our method successfully simplifies anatomical terms, structural definitions, and visual descriptions while preserving core medical information. Additional examples are provided in Appendix G.

4.1 Evaluation Metrics

To ensure a comprehensive assessment, we employ metrics across four dimensions: textual similarity, linguistic readability, clinical factuality, and downstream utility.

- **Relevance:** We use standard n-gram metrics to measure the structural similarity and lexical overlap between expert and layman captions. Specifically, we report BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005).
- **Readability:** To quantify the accessibility of the text, we utilize Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Coleman-Liau Index (CLI) (Coleman and Liau, 1975), Dale-Chall Readability Score (DCRS) (Dale and Chall, 1948), Simple Measure of Gobbledygook (SMOG) Index (Mc Laughlin, 1969), and Flesch Reading Ease (FRE) (Flesch, 1948). Additionally, we incorporate LENS (Maddela et al., 2023), a learnable metric specifically optimized for text simplification.
- **Radiological Factuality:** Evaluating the clinical integrity of simplified text is critical. We employ Radiological Report Text Evaluation (RaTEScore) (Zhao et al., 2024b) and Generative Radiology Report Evaluation and Error Notation (GREEN) (Ostmeier et al., 2024). These model-based metrics are designed to detect hallucinations and ensure clinical correctness in radiology reports.
- **Downstream Performance:** To assess whether the simplified text preserves essential semantic information for automated analysis, we evaluate zero-shot text-to-image retrieval performance. We report Recall@K (R@1, R@5, R@10) to measure retrieval accuracy using the generated captions.

4.2 Dataset Statistics and Quality Analysis

We analyze the linguistic characteristics and semantic consistency of MedLayBench-V, which comprises 79,789 image-text pairs across 7 modalities, maintaining the original ROCOv2 configuration (Rückert et al., 2024).

Linguistic Complexity and Accessibility. As presented in Table 1, our refinement pipeline successfully standardizes the linguistic complexity of medical captions.

- **Vocabulary Reduction:** The unique vocabulary size is reduced by 46.1% in the layman version compared to the expert version. This indicates a significant removal of long-tail medical jargon and noisy tokens, streamlining the dataset for generalizable learning.

Table 3: Overall top-K retrieval performance on MedLayBench-V across four modalities (X-Ray, CT, MRI, Ultrasound). **Bold** indicates best performance, underline indicates second best performance. Values are presented as Expert / Layman. All values are in percentage (%).

Model	Image → Text			Text → Image		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
OpenAI-Base	1.23 / 1.08	3.96 / 3.74	6.56 / 6.32	1.57 / 1.54	4.41 / 4.51	7.03 / 7.12
CoCa-Large	2.10 / 2.15	5.70 / 5.71	8.24 / 8.07	3.56 / 3.64	8.78 / 8.84	11.97 / 12.11
LAION-2B	2.28 / 2.33	6.67 / 6.58	9.88 / 9.78	4.31 / 4.29	9.94 / 9.88	13.76 / 13.74
OpenCLIP-Huge	3.33 / 3.44	8.71 / 8.43	12.58 / 12.28	5.17 / 5.15	11.88 / 12.10	16.59 / 16.70
PubMedCLIP	4.61 / 4.26	13.46 / 13.12	20.93 / 20.66	4.85 / 4.71	14.49 / 14.43	21.94 / 21.73
BMC-CLIP	22.69 / 22.42	40.83 / 40.36	50.33 / 49.65	23.04 / 23.21	42.09 / 42.03	52.09 / 51.71
PMC-CLIP	28.98 / 28.38	53.12 / 52.47	64.14 / 63.60	30.90 / 30.24	55.66 / 55.16	66.11 / 65.55
BiomedCLIP	**31.06 / 30.70**	**58.52 / 58.11**	**70.31 / 69.41**	**32.50 / 32.07**	**59.94 / 59.09**	**71.07 / 70.44**

- **Improved Readability:** We observe a consistent drop in grade-level metrics across the entire dataset. Notably, the FKGL drops from 13.10 to 10.35, and the Coleman-Liau Index decreases from a graduate level of 15.82 to 9.88, aligning with the recommended reading level for patient education materials (Rooney et al., 2021).
- **Enhanced Accessibility:** The FRE score more than doubles from 26.14 to 55.88. This shift in text difficulty from very confusing to fairly difficult ensures the content is accessible to a general audience with a standard high school education.

Detailed modality distributions and concept frequency analyses are provided in Appendix B.

Quality Consistency across Splits. Table 2 reports the semantic quality and consistency of our dataset. The relevance metrics, including BLEU-4, ROUGE-L, and METEOR, show minimal variance across training, validation, and test sets, with an overall METEOR of 53.12, confirming that our pipeline produces stylistically consistent outputs regardless of data split. The LENS score, a learnable metric for text simplification, remains stable at 63.19 across all splits, indicating robust rewriting quality throughout the dataset. Most importantly, the clinical correctness scores, RaTEScore and GREEN, demonstrate that our simplification preserves the factual integrity of the original reports, with the test set achieving 65.09 and 70.03 respectively, confirming high clinical safety despite reduced linguistic complexity.

Table 4: **Human Evaluation Results.** Two radiologists (E1, E2) and one lay reader (L) rated 100 SCGR-generated caption pairs on a 5-point Likert scale.

Criterion	E1	E2	L	Avg.
Factual Correctness	$4.67_{\pm 0.68}$	$4.96_{\pm 0.40}$	$4.95_{\pm 0.17}$	4.86
Completeness	$4.66_{\pm 0.68}$	$4.96_{\pm 0.40}$	$4.95_{\pm 0.22}$	4.86
Simplicity	$4.69_{\pm 0.63}$	$4.73_{\pm 0.53}$	$4.54_{\pm 0.62}$	4.65
Fluency	$4.85_{\pm 0.38}$	$4.98_{\pm 0.14}$	$4.96_{\pm 0.20}$	4.93

4.3 Human Evaluation

To validate the SCGR-generated captions beyond automatic metrics, we conducted a human evaluation following Jeblick et al. (2024). Two board-certified radiologists and one lay reader rated 100 randomly sampled caption pairs on a 5-point Likert scale across four criteria: Factual Correctness, Completeness, Simplicity, and Fluency.

As shown in Table 4, all criteria averaged above 4.5, with Factual Correctness and Completeness reaching 4.86, confirming that SCGR preserves clinical integrity. Simplicity scored comparatively lower at 4.65, suggesting room for optimization in certain specialized descriptions.

4.4 Downstream Task: Zero-Shot Retrieval

To evaluate the utility of MedLayBench-V, we conducted a zero-shot Image-Text Retrieval (ITR) experiment. This task measures how well models can align visual features with both *Expert* (original) and *Layman* (refined) textual descriptions. We report the Recall@K metrics for both Image-to-Text and Text-to-Image retrieval in Table 3, with visualizations provided in Figure A4i in Appendix E.

Bootstrap significance testing (n=1,000, two-sided) confirms that all absolute performance differences remain below 1.03%, with detailed results in Appendix D. An embedding-level analysis with t-SNE visualizations further confirms this finding (Appendix C).

Experimental Setup. Following the standard zero-shot retrieval protocol (Radford et al., 2021), we extract image and text embeddings from each dual-encoder model, apply L2-normalization, and compute pairwise cosine similarity across all image-text pairs in the test set (N=9,927). Recall@K is computed by checking whether the ground-truth match appears within the top-K ranked candidates. No fine-tuning or prompt engineering is applied; all models are evaluated using their publicly available pre-trained weights.

Baseline Models. We benchmarked diverse dual-encoder architectures, categorized into general-domain and medical-domain models. For the general domain, we employed OpenAI-CLIP (Radford et al., 2021) and OpenCLIP (Cherti et al., 2023) (trained on LAION-2B (Schuhmann et al., 2022)), along with CoCa (Yu et al., 2022), which integrates contrastive and generative objectives. For the medical domain, we selected models pre-trained on large-scale biomedical image-text pairs to assess the impact of domain adaptation. These include PubMedCLIP (Eslami et al., 2023), BMC-CLIP (Lozano et al., 2025), PMC-CLIP (Lin et al., 2023), and BiomedCLIP (Zhang et al., 2023), which utilize domain-specific encoders aligned with biomedical imagery.

Performance of Medical vs. General VLMs. We observe a clear performance hierarchy based on domain adaptation. While general domain models (e.g., OpenAI-CLIP) struggle with medical contexts (Recall@1 $< 5\%$), medical-specific models show improved alignment. BiomedCLIP achieves state-of-the-art performance, benefiting from its large-scale pre-training on biomedical literature.

Semantic Preservation in Layman Captions. Crucially, our results demonstrate that simplifying the language does not compromise semantic fidelity. As evidenced in Table 3, retrieval performance remains robust across all medical models, exhibiting negligible degradation when transitioning from *Expert* to *Layman* queries. For instance, BiomedCLIP exhibits only a marginal drop in Image-to-Text Recall@1 (31.06% $\rightarrow$ 30.70%).

Table 5: **SCGR Ablation Study.** Averaged R@1 (%) across I2T and T2I. Full per-model results in Appendix E.

Condition	CUI	MedLP	LLM	Avg. R@1
LLM Only	×	×	✓	1.96
LLM + CUI	✓	×	✓	2.08
SCGR (Ours)	✓	✓	✓	**11.26**
Expert	–	–	–	11.44

This explicitly verifies that MedLayBench-V successfully retains the core diagnostic semantics required for visual alignment, proving high readability can be achieved without sacrificing medical accuracy.

Ablation: Impact of Structured Grounding. To isolate each SCGR component, we conducted a systematic ablation (Table 5). Without structured grounding, LLM Only collapses to 1.96 avg R@1, an 83% drop from Expert. CUI extraction alone yields negligible recovery, while full SCGR restores 98.4% of Expert-level performance, confirming knowledge-constrained refinement as the critical component. Per-model breakdown is provided in Appendix E.

4.5 Downstream Task: Zero-Shot Captioning

To further expose the expert-lay register gap, we conducted a zero-shot captioning experiment using both medical and general-domain VLMs (Appendix F). In particular, LLaVA-Med (Li et al., 2023) exhibits severe expert bias with a BERTScore gap of +22.93 between expert and layman prompts, while other models show near-zero gaps, confirming that lay-register adaptability varies significantly across model families.

5 Conclusion

In this work, we introduced MedLayBench-V, the first multimodal benchmark for quantifying the semantic alignment between clinical jargon and lay language. By evaluating state-of-the-art VLMs, we formalized the existence of a representation alignment gap, revealing that current medical models are overfitted to the professional register at the expense of patient accessibility. Our proposed structured concept-grounded refinement pipeline provides a foundational framework for developing next-generation Medical AI that is both clinically accurate and universally understandable.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00251022) (K.S.C.); the SNUH Research Fund (No. 04-2024-0600; No. 04-2025-2060) (K.S.C.); and the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) grant funded by the Ministry of Health&Welfare (No. RS-2024-00439549) (K.S.C.).

Limitations

While MedLayBench-V establishes a foundation for patient-centric AI, we acknowledge limitations regarding the reliance on synthetic data, restriction to English, and modality imbalances inherited from the source. Although our pipeline ensures clinical correctness via structured constraints, synthetic captions may lack the subtle nuances of human-authored text, and validation with diverse patient groups is needed to assess real-world utility.

More importantly, we hypothesize that the representation alignment gap between clinical jargon and lay language may have been obscured by the limited complexity of the current retrieval task. We posit that a distinct gap exists but requires more challenging scenarios to be fully exposed. Consequently, our future work will focus on scaling this benchmark to a wider array of complex downstream tasks. By increasing both the scale and difficulty, we aim to rigorously identify this latent alignment gap and develop robust methodologies to effectively bridge the expert-lay divide.

Finally, while frontier models such as GPT, Gemini, and Claude may already possess expert-to-lay conversion capabilities, evaluating such ability requires a standardized resource with ontology-grounded references. MedLayBench-V serves this role by providing paired dual-register data for reproducible comparison across model families, analogous to how ImageNet remains a shared evaluation standard beyond its original difficulty level.

Despite these limitations, we believe MedLayBench-V represents a meaningful step toward closing the communication gap between clinical AI and patients, contributing to equitable and accessible healthcare. We encourage the community to extend this benchmark to multilingual settings, additional imaging modalities, and more diverse downstream tasks such as visual question answering and medical report generation.

References

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

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the biolaysumm 2024 shared task on the lay summarization of biomedical research articles. *arXiv preprint arXiv:2408.08566*.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. *arXiv preprint arXiv:2210.09932*.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, and 1 others. 2024. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European radiology*, 34(5):2817–2825.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of

new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, and 1 others. 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Weibin Liao, Tianlong Wang, Yinghao Zhu, Yasha Wang, Junyi Gao, and Liantao Ma. 2025. Magical: Medical lay language generation via semantic invariance and layperson-tailored adaptation. *arXiv preprint arXiv:2508.08730*.

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

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

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

Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, and 1 others. 2025. 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*, pages 19724–19735.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. Lens: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Naomi Miller, Eve-Marie Lacroix, and Joyce EB Backus. 2000. Medlineplus: building and maintaining the national library of medicine's consumer health web service. *Bulletin of the Medical Library Association*, 88(1):11.

Shufan Ming, Yue Guo, and Halil Kilicoglu. 2025. Towards knowledge-guided biomedical lay summarization using large language models. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 285–297.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and 1 others. 2024. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 374–390.

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

Philipp Prucker, Keno K Bressem, Jan Peeken, Mateo Jukic, Alexander W Marka, Maximilian Strenzke, Su Hwan Kim, Christian J Mertens, Dominik Weller, Tristan Lemke, and 1 others. 2025. A prospective controlled trial of large language model–based simplification of oncologic ct reports for patients with cancer. *Radiology*, 317(2):e251844.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Michael K Rooney, Gaia Santiago, Subha Perni, David P Horowitz, Anne R McCall, Andrew J Einstein, Reshma Jagsi, and Daniel W Golden. 2021. Readability of patient education materials from high-impact medical journals: a 20-year analysis. *Journal of patient experience*, 8:2374373521998847.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis,

Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Matthew Shardlow and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William K Cheung, and 1 others. 2025. Overview of the biolaysumm 2025 shared task on lay summarization of biomedical research articles and radiology reports. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 365–377.

Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, and Hong Yu. 2024. Readme: Bridging medical jargon and lay understanding for patient education through data-centric nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12609–12629.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Kun Zhao, Chenghao Xiao, Sixing Yan, Haoteng Tang, William K Cheung, Noura Al Moubayed, Liang Zhan, and Chenghua Lin. 2024a. X-ray made simple: Lay radiology report generation and robust evaluation. *arXiv preprint arXiv:2406.17911*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024b. Ratescore: A metric for radiology report generation. *arXiv preprint arXiv:2406.16845*.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. 2025a. Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. *arXiv preprint arXiv:2505.12371*.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong QingqingLong, Yefeng Zheng, and Xian Wu. 2025b. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769.

A Implementation Details and Prompts

In this section, we provide a comprehensive breakdown of the SCGR pipeline's implementation. The core of our approach lies in the rigorous separation of semantic extraction and stylistic refinement, as detailed in Algorithm 1. To ensure that the LLM adheres strictly to clinical facts while simplifying the syntax, we engineered a specific prompt template shown in Figure A1. By explicitly defining the system role as a "Medical Text Simplifier" and enforcing a JSON output format, we enable reliable automated parsing at scale. The "Critical Instructions" block serves as a safeguard against common pitfalls such as hallucinations or the use of subjective pronouns (e.g., "your body"), ensuring the output remains objective and professional.

Algorithm 1: SCGR framework

Input : Set of Expert Captions $\mathcal{T}_{exp} = \{T_{exp}^{(1)}, ..., T_{exp}^{(N)}\}$
Output : Set of Layman Captions $\mathcal{T}_{lay}$

Initialize $\mathcal{T}_{lay} \leftarrow \emptyset$
foreach $T_{exp} \in \mathcal{T}_{exp}$ **do**
  // Step 1: Hybrid Concept Extraction
  $C_{onto} \leftarrow$ ExtractCUIs(T_{exp}) // MedCAT
  $C_{ent} \leftarrow$ ExtractEntities(T_{exp}) // SciSpacy
  $C \leftarrow C_{onto} \cup C_{ent}$
  // Step 2: Knowledge Retrieval & Drafting
  $T_{draft} \leftarrow T_{exp}$
  foreach $c \in C_{onto}$ **do**
    $def \leftarrow$ RetrieveLayDef(c) // MedlinePlus
    $T_{draft} \leftarrow$ Substitute(T_{draft}, c, def)
  end
  // Step 3: Constrained Refinement (LLM)
  $P \leftarrow$ ConstructPrompt(T_{exp}, C, T_{draft})
  $T_{lay} \leftarrow$ Generate(P) // Llama-3
  // Step 4: Quality Verification
  if *CheckFactuality*(T_{lay}, T_{exp}) **then**
    $\mathcal{T}_{lay}$.add(T_{lay})
  end
end
return $\mathcal{T}_{lay}$

B Detailed Dataset Statistics

MedLayBench-V encompasses a diverse range of medical imaging modalities, mirroring real-world clinical distributions. As summarized in Table A1, Computed Tomography (CT) and X-Ray constitute the majority of the dataset, reflecting their prevalence in diagnostic radiology. Table A2 further breaks down the top co-occurring concepts for each modality, confirming that our extraction pipeline correctly identifies modality-specific anatomical structures (e.g., "left ventricle" in Ultrasound, "coronary artery" in Angiography). Additionally, Figure A2 illustrates the long-tail distribution of both UMLS concepts and raw terms. This indicates that while a few common concepts dominate the distribution (head), the dataset also preserves a vast array of rare, specific medical conditions (tail), which is crucial for comprehensive evaluation of medical VLMs.

SCGR Instruction Prompt Template

[System Role]
You are a precise **Medical Text Simplifier**. Rewrite the report for a high school student using the provided Concepts. **[Critical Instructions]**

1. **Source of Truth:** Trust the `Original Caption` completely. Ignore hallucinations in the Draft.
2. **Objective Tone: No 'you'/'your'.** Use 'the patient' or 'the body'.
3. **Strict Format:** Return **ONLY** the refined sentence. No "Note:" or explanations.
4. **No Hallucinations:** Do not invent words. Keep unclear terms in parentheses.

[User Input Template]

Original (Fact): "{Expert Caption (T_{exp})}"
Concepts: [{Verified UMLS Concepts (C)}]
Draft (Ref): "{Noisy Layman Draft (T_{draft})}"
[Structured Output]

```
{
   "layman_caption": "The CT scan shows an enlarged heart..."
}
```

Figure A1: **Prompt Construction for SCGR.** The prompt enforces strict adherence to the *Original Caption* as the source of truth while utilizing the *Draft* only for stylistic reference. The output is constrained to an objective, third-person tone.

Table A1: Distribution of Imaging Modalities. The number of image-caption pairs for each modality as reported in the original ROCOv2 dataset (Rückert et al., 2024).

Code	Modality Name	Count
DRCT	Computed Tomography (CT)	27,747
DRXR	X-Ray (Plain Radiography)	21,997
DRMR	Magnetic Resonance Imaging (MRI)	12,657
DRUS	Ultrasonography	11,429
DRAN	Angiography	4,799
DRCO	Combined Modality	728
DRPE	Positron Emission Tomography (PET)	432
Total		**79,789**

Table A2: Detailed Top 5 Concepts Distribution per Modality. The frequency of the top 5 co-occurring concepts extracted from the text context for each major imaging modality.

(a) Computed Tomography (CT)

Rank	Concept	Freq
1	X-Ray Computed Tomography	27,747
2	CT scan	3,474
3	abdomen	2,869
4	arrow	2,802
5	image	1,669

(b) Magnetic Resonance Imaging (MRI)

Rank	Concept	Freq
1	Magnetic Resonance Imaging	12,659
2	arrow	1,246
3	image	957
4	patient	712
5	brain	661

(c) Ultrasonography

Rank	Concept	Freq
1	Ultrasonography	11,422
2	arrow	892
3	image	633
4	left ventricle	610
5	left atrium	564

(d) Plain X-Ray

Rank	Concept	Freq
1	Plain x-ray	21,936
2	Anterior-Posterior (AP)	9,606
3	Chest	7,196
4	Postero-Anterior (PA)	4,302
5	Bone structure of cranium	3,973

(e) Angiography

Rank	Concept	Freq
1	angiogram	4,766
2	arrow	435
3	Coronary angiography	228
4	right coronary artery	219
5	stenosis	195

(f) Positron-Emission Tomography (PET)

Rank	Concept	Freq
1	Positron-Emission Tomography	432
2	uptake	98
3	increased	64
4	image	40
5	patient	34

C Semantic Preservation Analysis

To empirically validate that our simplification process preserves the underlying medical semantics, we analyzed the embedding space of various Vision-Language Models. Figure A3 visualizes the t-SNE projections of image-text embeddings for both Expert (original) and Layman (refined) captions. Across different architectures (OpenAI-CLIP, BiomedCLIP, PMC-CLIP), we observe that the distributions of Expert and Layman embeddings are nearly isomorphic. Furthermore, the high cosine similarity (≈ 0.99) and low Euclidean distance distributions confirm that the transition to lay language does not shift the semantic vector significantly. This serves as strong evidence that MedLayBench-V successfully lowers the linguistic barrier without compromising the diagnostic information required for downstream evaluation.

D Bootstrap Significance Test

To verify that the performance differences between Expert and Layman captions are not attributable to sampling variance, we conducted bootstrap significance testing (n=1,000, two-sided) on the Overall Recall@K delta, as summarized in Table A3.

General-domain models show largely non-significant differences ($p > .05$), consistent with their low baseline where minor fluctuations are indistinguishable from noise. Medical-domain models exhibit statistically significant drops ($p < .05$) across all metrics, with the largest delta observed for BiomedCLIP at R@10 (-0.75%). Nevertheless, all $|\Delta|$ remain below 1.03%, confirming the degradation is statistically detectable but practically negligible for retrieval.

Table A3: **Bootstrap Significance Test.** Δ denotes Layman $-$ Expert (%). * indicates $p < .05$.

Model	R@1		R@5		R@10	
	Δ	p	Δ	p	Δ	p
General-Domain Models						
OpenAI	−0.07	.020*	−0.07	.222	−0.13	.148
CoCa	−0.03	.590	−0.08	.188	−0.12	.056
LAION	−0.04	.404	−0.05	.248	−0.22	.032*
OpenCLIP	+0.04	.980	−0.13	.202	−0.16	.076
Medical-Domain Models						
PubMedCLIP	−0.20	.006*	−0.28	.004*	−0.32	.008*
BMC-CLIP	−0.10	.036*	−0.45	.002*	−0.57	.001*
PMC-CLIP	−0.66	.001*	−0.67	.001*	−0.64	.001*
BiomedCLIP	−0.41	.001*	−0.65	.001*	−0.75	.001*

Table A4: **Per-Model SCGR Ablation Results.** Averaged R@1 (%) across I2T and T2I for each ablation condition. Condition definitions follow Table 5.

Condition	Pipeline Steps			General-Domain				Medical-Domain				Avg.
	CUI	MedLP	LLM	OpenAI	CoCa	LAION	CLIP	PubMed	BMC	PMC	BioMed	
LLM	×	×	✓	0.50	0.77	0.74	1.03	1.18	3.32	4.19	3.96	1.96
CUI	✓	×	✓	0.41	0.78	0.93	1.17	1.24	3.58	4.51	4.00	2.08
SCGR (Ours)	✓	✓	✓	0.99	2.08	2.52	3.38	3.99	20.53	27.27	**29.30**	**11.26**
Expert	–	–	–	1.06	2.11	2.56	3.34	4.19	20.63	27.93	29.71	11.44

E Ablation Study

Table A4 extends the averaged ablation results in Table 5 with a per-model breakdown. The LLM Only condition shows uniformly poor performance across all models, with medical-domain models suffering disproportionately larger gaps relative to Expert. Adding CUI extraction alone provides marginal gains, confirming that ontological grounding is insufficient without lexical substitution via MedlinePlus. Full SCGR recovers near-Expert performance consistently, with BiomedCLIP showing the largest absolute improvement from 3.96 to 29.30.

Figure A4i visualizes the retrieval performance across all models, confirming negligible gaps between Expert and SCGR-generated Layman captions. In contrast, Figure A4ii illustrates that removing structured grounding causes severe degradation, with BiomedCLIP I2T R@1 collapsing from 31.1% to 5.3%. We identify two dominant failure modes of the Naive LLM. First, it tends to over-simplify specific pathologies into vague terms (e.g., "pneumothorax" → "lung problem"), losing discriminative features. Second, it hallucinates plausible but incorrect details to fill narrative gaps. These findings confirm that explicit knowledge grounding, as provided by SCGR, is essential for high-quality medical lay language generation.

F Zero-Shot Captioning Analysis

To complement the retrieval-based evaluation, we conducted a zero-shot image captioning experiment to directly assess whether current VLMs can adapt their output register. Two medical-domain models, LLaVA-Med (Li et al., 2023) and MedGemma 1.5 (Sellergren et al., 2025), and two general-domain models, LLaVA-v1.5 (Liu et al., 2024) and Qwen2-VL (Wang et al., 2024), each received dual prompts per image on 1,000 test pairs: (A) "Describe this medical image in one sentence using clinical terminology" and (B) "Describe this medical image in one sentence using simple language that a patient with no medical background can understand." We report BERTScore (Zhang et al., 2019) against Expert and Layman references respectively, along with FKGL to measure readability shift.

Table A5: **Zero-Shot Captioning Results.** BERTScore (DeBERTa-xlarge-MNLI) against register-matched references. Δ = Expert − Layman; positive indicates expert-register bias.

Model	BS_{Exp}	BS_{Lay}	Δ	$FKGL_{Exp}$	$FKGL_{Lay}$
Medical-Domain					
LLaVA-Med	55.04	32.12	**+22.93**	7.2	4.1
MedGemma 1.5	64.31	65.11	−0.80	13.6	6.5
General-Domain					
LLaVA-v1.5	61.13	63.05	−1.92	4.8	5.0
Qwen2-VL	63.05	65.28	−2.23	15.0	9.0

As shown in Table A5, LLaVA-Med (Li et al., 2023) shows a severe expert bias (Δ=+22.93) despite producing syntactically simpler outputs (FKGL 7.2→4.1), indicating the bottleneck lies in vocabulary register rather than syntactic complexity. The remaining models exhibit near-zero gaps (Δ=−0.80 to −2.23) with notable readability shifts, suggesting that lay-register adaptability varies across VLM families. This heterogeneity motivates the need for a standardized benchmark like MedLayBench-V to systematically evaluate and improve expert-lay alignment.

G Extended Qualitative Analysis

To further demonstrate the robustness and versatility of the SCGR pipeline, we provide an extended set of qualitative examples across diverse imaging

modalities. Figure A5 and Figure A6 illustrate how our pipeline handles specific linguistic challenges, ranging from simplifying complex vascular anatomy in CT/MRI to interpreting acoustic artifacts in ultrasound. Each example highlights the transformation from the original expert report (Expert) to the generated patient-friendly caption (Layman). Key medical terms are highlighted in grey while their simplified explanations are highlighted in blue to visualize the semantic alignment.

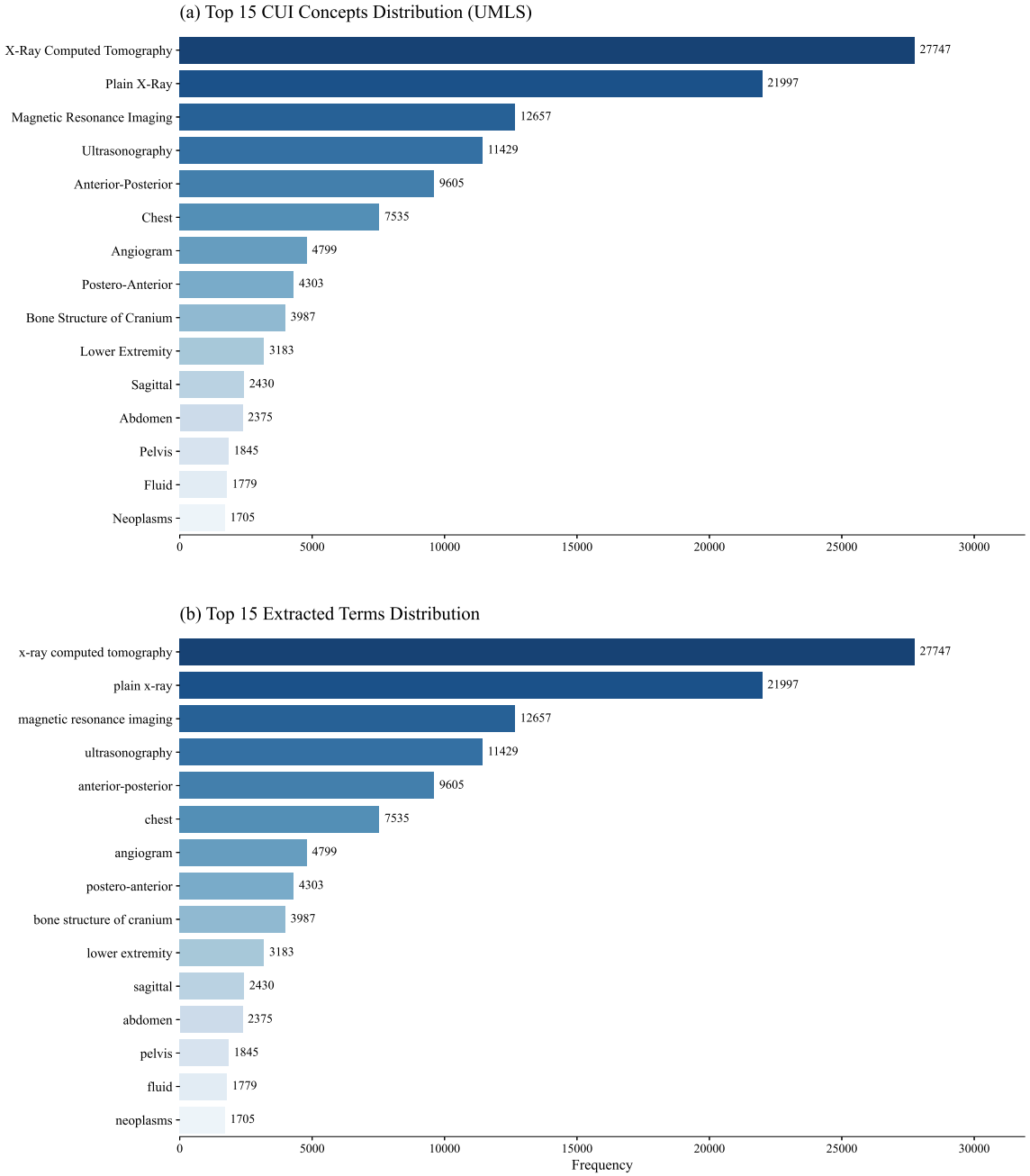


Figure A2: **Distribution of Top 15 Concepts and Terms.** (a) The frequency of Unique Medical Language System (UMLS) Concept Unique Identifiers (CUIs) mapped from the dataset. (b) The frequency of raw extracted terms directly from the captions. Both distributions illustrate the long-tail nature of medical findings in the dataset.

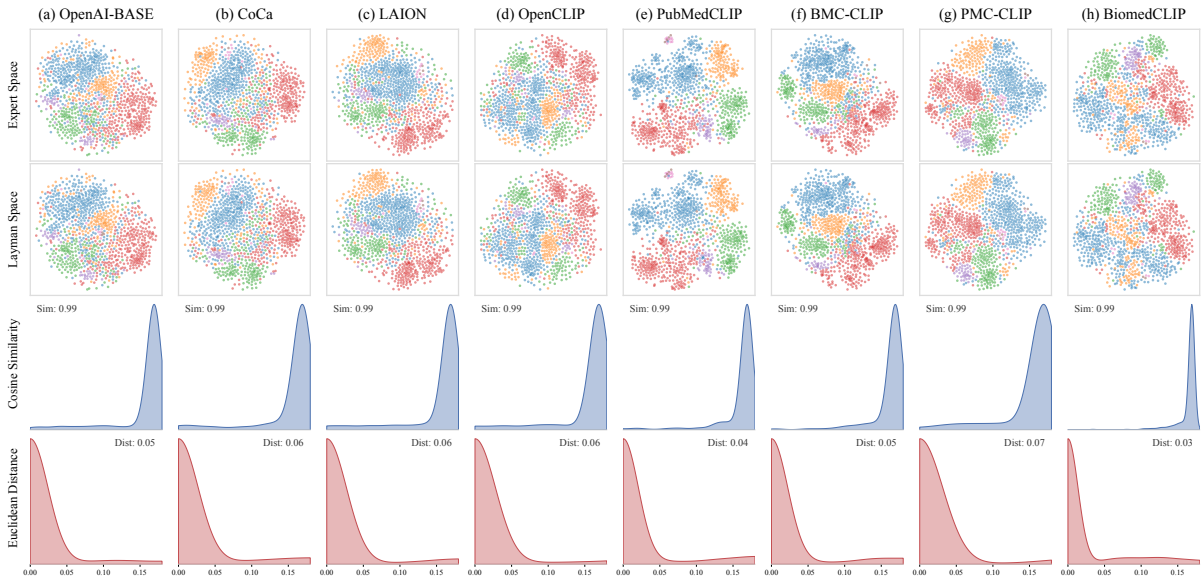
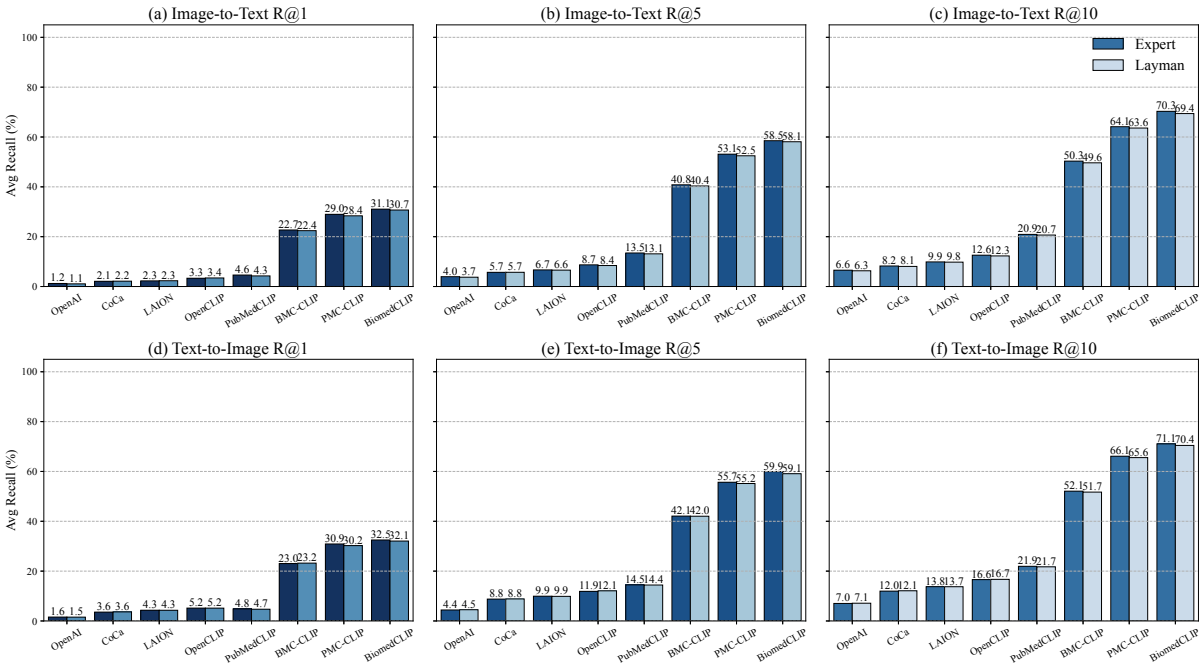
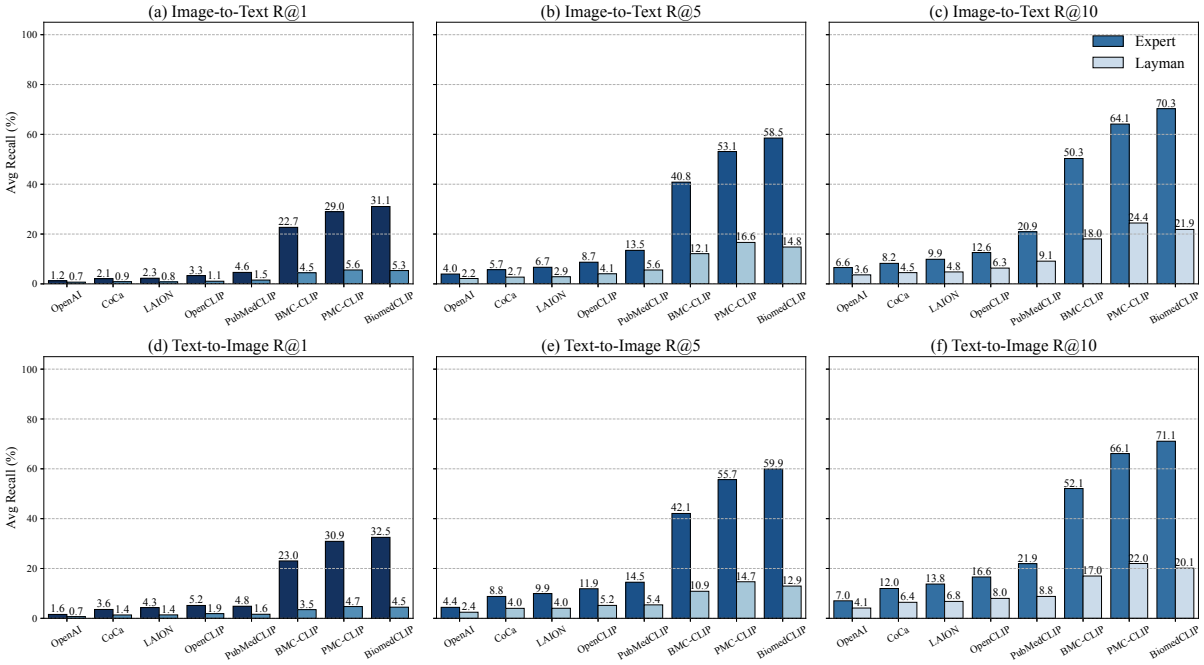


Figure A3: **Embedding space visualization across different CLIP models.** Each column represents a different model. Rows 1–2: t-SNE projections of Expert and Layman embeddings, colored by modality. Row 3: Cosine similarity distribution. Row 4: Euclidean distance distribution. High similarity (Sim $\approx$ 0.99) and low distance (Dist $\approx$ 0.05–0.07) confirm semantic preservation across all models.



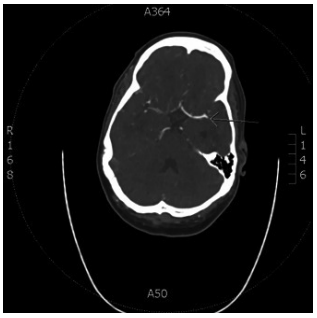
(i) **Zero-Shot Retrieval Performance.** Recall@K results for Image-to-Text (a–c) and Text-to-Image (d–f) tasks. Dark and light bars denote Expert and Layman queries, respectively.



(ii) **Impact of Naive LLM-only Simplification.** Recall@K results using layman captions generated without structured grounding. BiomedCLIP I2T R@1 collapses from 31.1% to 5.3%.

Figure A4: **Retrieval Performance and Ablation Visualization.** (i) SCGR preserves semantic fidelity with negligible gaps between registers. (ii) Naive LLM simplification causes severe semantic drift across all models.

(a) Computed Tomography (CT)



CUI: [C0040405] X-Ray Computed Tomography

Concept: X-Ray Computed Tomography, anterior cerebral artery, internal carotid artery, occlusive thrombi, M2 segment, angiogram

Expert: CT brain perfusion and angiogram on day 16, showing left MCA and ACA territories with occlusive thrombi demonstrated within the inferior division of the M2 segment of the left MCA and within the callossomarginal branch of the left ACA. There was also a complete left ICA occlusion.

Layman: A CT scan showed that there were blood clots blocking part of two main arteries supplying oxygen-rich blood to one side of the brain - specifically, the lower section of the middle cerebral artery's smaller branches and a small branch off the front part of another major artery called the anterior cerebral artery. Additionally, the main pipeline bringing oxygenated blood from the neck to this area had completely blocked.

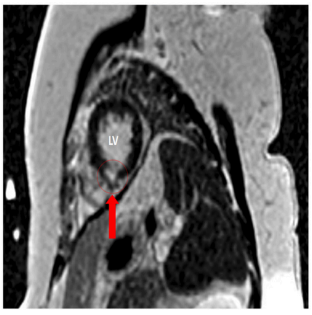
CUI: [C0040405] X-Ray Computed Tomography

Concept: X-Ray Computed Tomography, right temporalis muscle, rim-enhancing, intramuscular, white arrows, hypodensity

Expert: On coronal cut, there is presence of rim-enhancing intramuscular hypodensity within the bulky (white arrows) and thickened right temporalis muscle with enhancement of the right temporalis muscle.

Layman: There is an area on a CT scan where part of the right temple muscle appears swollen and has some empty space inside it that shows up as darker than usual, indicated by the thicker-than-normal right temple muscle marked by the white arrows.

(b) Magnetic Resonance Imaging (MRI)

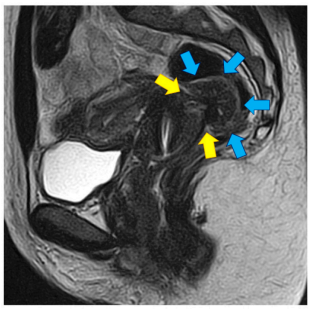


CUI: [C0024485] Magnetic Resonance Imaging

Concept: Inferoseptal late gadolinium enhancement, Magnetic Resonance Imaging, myocardial fibrosis, left ventricle, regional scar

Expert: Inferoseptal late gadolinium enhancement on MRI indicating regional scar and/or myocardial fibrosis. LV, left ventricle.

Layman: The heart's MRI shows signs of scarring or thickening inside one part of the muscle that pumps blood out of the heart, specifically the area between two chambers called the septum. This can be due to damage from past injury or disease affecting the main pumping chamber known as the left ventricle.



CUI: [C0024485] Magnetic Resonance Imaging

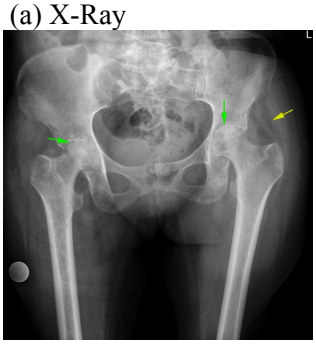
Concept: Magnetic Resonance Imaging, T2-weighted sagittal view, anterior rectal wall, posterior part, yellow arrow, infiltrates

Expert: Pelvic MRI with a T2-weighted sagittal view of a DE nodule (blue arrow) invading the anterior rectal wall. The nodule infiltrates the anterior rectal wall at the level of posterior part of the cervix (yellow arrow).

Layman: A pelvic Magnetic Resonance Image shows a growth (DE nodule indicated by the blue arrow) that has invaded the front part of the rectum's outer layer. This growth also affects the area near the back portion of the uterus (cervix), which is located behind it.

Figure A5: **Qualitative Analysis on Cross-Sectional Modalities.** Comparison of expert and layman descriptions for (a) CT and (b) MRI.

(a) X-Ray



CUI: [C1306645] Anterior-Posterior
[C0030797] Pelvis
[C1999039] Plain x-ray

Concept: pelvic X-ray shows sclerosis, inflammatory process, Anterior-Posterior, gluteal fat plans, Soft tissue edema, femoral head AVN, Plain x-ray

Expert: AP view pelvic X-ray shows sclerosis and collapse of bilateral femoral head in favor of femoral head AVN (green arrow). Soft tissue edema and displaced gluteal fat plans are in favor of joint fluid and possible inflammatory process (Yellow arrow)

Layman: A pelvis X-ray taken from the front shows hardening and shrinking of both hip bones at their top parts due to bone damage called avascular necrosis (AVN) (indicated by the green arrow). There's also swelling around these areas caused by extra liquid in the joints and possibly an inflamed condition (marked by the yellow arrow).

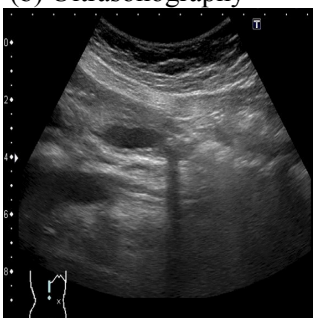
CUI: [C1306645] Anterior-Posterior
[C0023216] Extremity
[C1999039] Plain x-ray

Concept: femoral neck fracture, Anterior-Posterior, tumoral calcinosis, calcific deposits, Lower Extremity, left hip joint, calcified mass

Expert: Preoperative plain radiographs of the femoral neck fracture with tumoral calcinosis. The presence of a multiple calcified mass was observed around the left hip joint. The femur neck fracture is not presented clearly because of calcific deposits.

Layman: The preoperative X-rays show that there's a problem near the break in the upper part of the thigh bone on one side. A collection of hardened calcium is seen surrounding the main joint on this leg. The break itself isn't visible due to these hard calcium buildups.

(b) Ultrasonography



CUI: [C0041618] Ultrasonography

Concept: Abdominal ultrasonography, distal bile duct, acoustic shadow, Ultrasonography, echoes

Expert: Abdominal ultrasonography revealed strong echoes in the distal bile duct with a 9 mm acoustic shadow.

Layman: The ultrasound scan of the abdomen showed bright spots inside the lower part of the tube that carries digestive juices from the liver to the small intestine, blocking some sound waves so they appear as dark areas on the image.

CUI: [C0041618] Ultrasonography

Concept: rectus abdominus muscle, Rectus sheath block, peritoneal cavity, posterior sheath, local anesthetic, Ultrasonography

Expert: Rectus sheath block. Arrow pointing to needle tip injecting local anesthetic into posterior sheath of rectus abdominus between the rectus abdominus muscle (superficial) and peritoneal cavity (deep).

Layman: The doctor is performing a procedure called a Rectus Sheath Block by inserting a small amount of medicine through a thin needle into the space behind the main abdominal muscles, specifically between these muscles and the inner lining that surrounds organs inside the abdomen.

Figure A6: **Qualitative Analysis on Cross-Sectional Modalities.** Comparison of expert and layman descriptions for (a) X-Ray and (b) Ultrasonography.