UNIFYING MEDICAL VISION-LANGUAGE LEARNING ACROSS IMAGING MODALITIES IN VIETNAMESE: A MULTIMODAL DATASET AND BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical report generation has advanced rapidly with Vision-Language Models (VLMs), yet progress remains constrained by the scarcity of high-quality paired medical image-report data, particularly for low-resource languages. Existing datasets are largely modality-specific, covering common modalities such as X-ray, CT, or MRI individually rather than jointly, while PET/CT remains underrepresented. Moreover, many datasets are assembled from figure screenshots or re-captured images rather than the native DICOM studies acquired in routine clinical practice, limiting their clinical fidelity and practical value. To address these limitations, we introduce **ViMed-Rad**, a large-scale multimodal radiology dataset jointly covering four major imaging modalities: X-ray, CT, MRI, and PET/CT. ViMed-Rad comprises 93,781 clinical studies paired with Vietnamese reports authored by radiologists, including 62,148 X-ray, 13,247 CT, 12,010 MRI, and 6,376 whole-body PET/CT studies. All studies are native DICOM acquisitions exported from hospital PACS archives, retaining acquisition metadata and spanning diverse anatomical regions and imaging protocols. Building on ViMed-Rad, we establish **ViMed-Bench**, a unified benchmark that couples clinical report generation and report-grounded VQA tasks with a multimodal framework integrating modality-specific visual representations into a shared language generation space, so that a single model serves all four modalities. Benchmarking existing medical VLMs reveals substantial performance gaps across imaging types, particularly for underrepresented modalities, together with a consistent degradation when reports are generated in Vietnamese. Experiments on ViMed-Rad show that the proposed framework accommodates heterogeneous planar, volumetric, and functional-anatomical imaging within one model, outperforming baselines across all four modalities while retaining most of the accuracy of modality-specific counterparts, and providing a scalable foundation for multi-modality medical VLMs rather than separate models for each imaging domain.

1 INTRODUCTION

Medical imaging plays a central role in diagnosis, treatment planning, and disease monitoring. However, shortages of radiologists and increasing imaging demand place substantial pressure on healthcare systems, especially in resource-constrained settings. Our survey at major Vietnamese hospitals illustrates this workload, with reported daily volumes per radiologist of approximately 300 X-ray, 45 MRI, 70 CT, and 20 PET/CT examinations. These challenges motivate AI-assisted tools that support image interpretation and report drafting under clinical oversight.

Medical Vision-Language Models (VLMs) offer a promising foundation for such tools by connecting medical images with clinical language. They support tasks including Visual Question Answering (VQA) and report generation. LLaVA-Med Li et al. (2023) demonstrates biomedical instruction-following capabilities, while RadFM Wu et al. (2025) and M3D Bai et al. (2024) extend vision-language learning to 2D and volumetric imaging. More recent generalist models, including MedVersa Zhou et al. (2025) and MedGemma 1.5 Sellergren et al. (2026), further broaden coverage across imaging modalities and clinical tasks. Progress in medical VLMs has been driven by datasets pairing images with clinical text. MIMIC-CXR Johnson et al. (2024) provides chest radiographs

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

Table 1: Comparison of medical imaging-text datasets across imaging modalities, image format, scale, text supervision, language, and data source. **2D** and **3D** denote planar and volumetric imaging; **EN** and **VI** denote English and Vietnamese.

Dataset	X-ray	CT	MRI	PET/CT	Format	# Samples	Text Type	Lang.	Source
MIMIC-CXR	2D	–	–	–	DICOM	227K	Report	EN	PhysioNet
CT-RATE	–	3D	–	–	NIfTI	25K	Report	EN	CT-RATE
RadGenome-Brain MRI	–	–	3D	–	NIfTI	1K	Report	EN	Public MRI
PMC-OA	2D	2D	2D	2D	JPEG	1,646K	Caption	EN	PMC OA
OmniMedVQA	2D	2D	2D	–	PNG	118K	VQA	EN	Websites
M3D-Data	–	3D	–	–	Numpy	120K	Text & VQA	EN	Websites
MedMD	2D	3D	3D	–	Numpy	16M	Text & VQA	EN	Websites
ViX-Ray	2D	–	–	–	PNG	5.4K	Report	VI	Hospital
ViMed-PET	–	–	–	3D	DICOM	2,757	Report & VQA	VI	Hospital
ViMed-Rad (Ours)	**2D**	**3D**	**3D**	**3D**	**DICOM**	**93K**	**Report & VQA**	**VI**	**Hospital**

with radiology reports, and CheXpert Plus Chambon et al. (2024) adds reports and demographics at a comparable scale. CT-RATE Hamamci et al. (2026) and Merlin Blankemeier et al. (2026) pair chest CT with reports, and MR-RATE Forithmus (2026) and AutoRG-Brain Lei et al. (2024) extend report supervision to MRI. These resources have enabled substantial methodological progress, but their coverage varies across imaging modalities, clinical settings, languages, and data formats. In practice, clinicians review several modalities within one Picture Archiving and Communication System (PACS) workflow National Electrical Manufacturers Association (2025). A medical VLM should therefore serve all of them in one system rather than one model per modality. This requires data spanning the modalities under consistent tasks and protocols, and reflecting the patient population and reporting language of the target setting Zech et al. (2018); Salmè et al. (2025). Existing resources satisfy neither requirement fully, leaving two gaps: in modality coverage and access to native imaging, and in geographic and linguistic representation.

Fragmented modality coverage and inconsistent access to native imaging. Widely used image-report datasets remain concentrated on individual modalities, particularly chest X-ray Johnson et al. (2024); Demner-Fushman et al. (2016); Bustos et al. (2020) and, increasingly, CT Hamamci et al. (2026); Chen et al. (2024). Resources such as ViMed-PET Nguyen et al. (2025) and PET2Rep Zhang et al. (2026) have begun to cover functional imaging, yet supervision remains uneven across X-ray, CT, MRI, and PET/CT. Native imaging data are also inconsistently available: MIMIC-CXR releases DICOM images Johnson et al. (2024), whereas other resources distribute publication-derived figures Lin et al. (2023); Hu et al. (2024) or preprocessed volumes Hamamci et al. (2026); Bai et al. (2024). Training and evaluating models on these representations may introduce a gap between benchmark conditions and clinical deployment. As summarized in Table 1, the remaining need is therefore not simply for more images, but for a resource that combines broad modality coverage, native imaging data, and consistent evaluation.

Limited geographic and linguistic representation. Widely used report-generation datasets draw on a limited set of clinical settings, and large-scale report supervision remains predominantly English Johnson et al. (2024); Hamamci et al. (2026); Forithmus (2026). Translating existing reports can improve linguistic accessibility, but it does not address differences in patient populations, nor does it reproduce the negation and uncertainty conventions of reports authored in local clinical practice Salmè et al. (2025); Wan et al. (2023). In Vietnamese, existing resources remain modality-specific: ViX-Ray covers chest X-ray Nguyen et al. (2026) and ViMed-PET covers PET/CT Nguyen et al. (2025). A broader resource is therefore needed to study both local adaptation and learning across imaging modalities.

To address these gaps, we introduce **ViMed-Rad**, a large-scale Vietnamese radiology dataset spanning four major imaging modalities: X-ray, CT, MRI, and whole-body PET/CT. ViMed-Rad combines selected public resources with our in-house clinical collection, bringing together 93,781 native-DICOM studies with expert-authored reports and 28.8M slices. To the best of our knowledge, it is the first resource to bring these four modalities together in native DICOM format under common task definitions and evaluation protocols. Building on this resource, we establish **ViMed-Bench**, a benchmark for report generation and three report-grounded VQA tasks of increasing difficulty. It includes a unified framework with modality-specific encoders, a shared projector and language model,

Table 2: Zero-shot performance of medical VLMs on public radiology benchmarks, evaluated against English (EN) references and their Vietnamese (VI) translations. VI cells are shaded by their relative drop from the EN score of the same model and metric: 10-25%, 25-50%, 50-75% and above 75%, with drops below 10% left unshaded.

			IU X-ray				MIMIC-CXR				ViXray			
	Model	**Lang.**	B-4	R-1	R-L	BERT	B-4	R-1	R-L	BERT	B-4	R-1	R-L	BERT
2D Imaging	RadFM	EN	4.88	36.34	27.10	76.56	1.13	15.29	11.02	64.20	0.10	6.98	6.33	59.68
		VI	0.45	10.01	9.93	63.66	0.12	4.23	3.96	56.13	0.22	7.43	7.07	59.84
	LLaVA-Med	EN	0.29	5.88	3.93	63.39	0.63	18.32	10.55	66.22	0.60	17.31	11.05	66.81
		VI	0.26	5.44	5.16	59.39	0.31	6.85	5.56	59.62	0.40	9.61	8.36	59.72
	MedVersa	EN	5.34	40.30	28.57	77.50	3.72	31.45	20.72	73.06	0.90	19.59	13.64	70.19
		VI	0.48	11.20	11.01	64.26	0.24	6.82	6.46	61.42	0.36	8.10	7.84	60.72
	MedGemma 1.5	EN	4.89	35.60	26.62	75.74	2.69	27.80	18.03	70.14	0.51	15.94	9.47	66.86
		VI	0.46	10.85	10.45	63.63	0.82	13.33	8.88	63.32	0.43	15.57	10.12	65.81
			CT-RATE (CT)				**BIMCV-R (CT)**				**MR-RATE (MRI)**			
	Model	**Lang.**	B-4	R-1	R-L	BERT	B-4	R-1	R-L	BERT	B-4	R-1	R-L	BERT
3D Volumetric	RadFM	EN	1.33	23.10	19.14	64.43	0.13	11.95	9.02	57.56	0.10	9.33	6.15	62.07
		VI	0.02	2.50	2.44	54.87	0.04	9.99	7.47	56.90	0.04	1.44	1.27	55.72
	LLaVA-Med	EN	1.06	20.45	12.44	67.10	0.35	15.13	9.00	62.73	0.06	12.45	9.19	63.81
		VI	0.60	8.38	4.23	61.10	0.23	8.92	4.72	59.90	0.01	1.70	1.30	58.41
	M3D	EN	1.52	26.02	12.47	70.09	0.14	17.95	8.91	66.78	1.36	19.57	13.08	68.53
		VI	0.11	3.42	3.42	53.55	0.11	12.45	5.40	65.47	0.21	1.57	4.82	61.84
	MedVersa	EN	0.11	15.85	10.56	66.61	0.06	0.64	0.54	49.71	0.01	4.51	3.88	58.81
		VI	0.01	0.60	0.49	56.75	0.05	0.42	0.43	48.32	0.01	0.13	0.13	52.53

B-4, R-1, R-L and BERT denote BLEU-4, ROUGE-1, ROUGE-L and BERTScore, and higher is better.

and progressive training for single- and joint-modality learning. Using these, we evaluate the model gain from local supervision, and whether one model can serve four modalities without substantial loss relative to modality-specific counterparts. The main contributions of this work are summarized as follows:

- We introduce **ViMed-Rad**, a Vietnamese radiology dataset of 93,781 native DICOM studies with expert-authored reports, covering X-ray (62,148), CT (13,247), MRI (12,010), and whole-body PET/CT (6,376), totaling 28.8M slices. To our knowledge, it is the first to combine these four modalities in native DICOM under common task definitions and evaluation protocols.
- We establish **ViMed-Bench**, which couples report generation and report-grounded VQA tasks of increasing difficulty under fixed patient-level splits. It includes a unified adaptation framework with modality-specific encoders, a shared projector and language model, and progressive fine-tuning for single- and joint-modality learning.
- We benchmark medical VLMs alongside GPT-4o and DeepSeek-v4.1 in the same setting, using text generation and radiologist-designed clinical metrics for findings, attributes, negation, and uncertainty. The evaluation quantifies language gaps, the benefit of local fine-tuning, and the cost of unification: one model retains most clinical accuracy of four modality-specific counterparts.

2 Related Work

Medical Image-Text Datasets. Existing resources differ in the text they pair with an image. Caption-based datasets Rückert et al. (2024); Lin et al. (2023) contain millions of pairs, but figure captions lack the structured findings and diagnostic reasoning found in clinical reports. Medical VQA datasets Lau et al. (2018); Liu et al. (2021); Hu et al. (2024) instead supply targeted questions, which are flat sets of localized pairs with no graded reasoning difficulty. Report-based resources provide the long-form clinical language required for report generation, with MIMIC-CXR Johnson et al. (2024) for chest X-ray, CT-RATE Hamamci et al. (2026) for CT, and MR-RATE Forithmus (2026) for MRI, while BIMCV-R Chen et al. (2024) relies on GPT-4 translations that may introduce artifacts. Beyond English, ViX-Ray Nguyen et al. (2026) and ViMed-PET Nguyen et al. (2025) cover only Vietnamese chest X-ray and PET/CT. ViMed-Rad pairs native Vietnamese reports across four modalities with report-grounded VQA tasks ordered by reasoning difficulty.

Table 3: Demographic and imaging statistics of the proposed ViMed-Rad multimodal dataset. X-ray, CT, and MRI data are collected in 2025, while PET/CT data span 2017-2025, excluding 2022. A case denotes one study, **Demo.** denotes demographics, and SD denotes standard deviation. Dashes (–) indicate statistics that do not apply to the modality.

	Statistic	X-ray	CT	MRI	PET/CT
	Cases	62,148	13,247	12,010	6,376
	Patients	54,077	11,830	11,410	6,309
Demo.	Age (mean ± SD)	$46.21 \pm 16.97^{\dagger}$	52.87 ± 16.51	50.36 ± 15.95	59.08 ± 14.63
	Age (range)	0-120*	5-99	6-98	2-96
	Male	27,043 (50.01%)	6,658 (56.28%)	5,917 (51.86%)	4,153 (65.81%)
	Female	27,034 (49.99%)	5,172 (43.72%)	5,493 (48.14%)	2,156 (34.19%)
Imaging	Images / volumes	1.44 ± 0.84	5.18 ± 2.56	12.95 ± 7.58	2.00 ± 0.00
	Total slices	–	19,594,389	5,593,842	3,579,984
	Slices / volume	–	285.71 ± 223.87	35.96 ± 48.72	280.33 ± 42.36

*Age 0 indicates patients younger than one year.
†X-ray age statistics are computed over the 50,701 patients (of 54,077) with an available age.

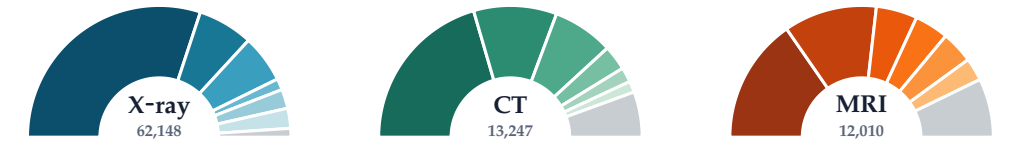
Medical VLMs and Their Evaluation. Serving many imaging domains with one model is an active line of work, from RadFM Wu et al. (2025) and MedVersa Zhou et al. (2025), which interleave 2D and 3D inputs, to Uni-Med Zhu et al. (2024), UniMedVL Ning et al. (2025) and Lingshu Xu et al. (2026), which unify eight or more modalities. All train on aggregated public English corpora, so modality, data source and reporting convention vary together and the cost of unification cannot be attributed to any one of them. Evaluation is equally unsettled: BLEU and ROUGE reward surface wording, motivating content-based metrics such as CheXbert similarity Smit et al. (2020), RadGraph-F1 Jain et al. (2021), RaTEScore Zhao et al. (2024) and GREEN Ostmeier et al. (2024), all built for English chest radiographs and dependent on English schemas, labelers or entity extractors. ViMed-Bench therefore fixes splits and protocol across four modalities and scores findings, locations, attributes, negation and certainty against radiologist-designed schemas.

Cross-Lingual Medical Report Generation. The cross-lingual behaviour of medical VLMs has received far less attention than their English performance. Med-UniC Wan et al. (2023) shows that language communities introduce a measurable bias into medical vision-language pretraining that degrades multimodal and vision-only tasks alike, and adapting a VLM to the target language outperforms general and domain-only baselines in European languages Salmè et al. (2025). Whether these observations extend to a low-resource language and to volumetric and functional imaging remains untested. We therefore evaluate five medical VLMs, RadFM Wu et al. (2025), LLaVA-Med Li et al. (2023), M3D Bai et al. (2024), MedVersa Zhou et al. (2025), and MedGemma 1.5 Sellergren et al. (2026), on public X-ray, CT, and MRI benchmarks against English references and their Vietnamese translations. As shown in Table 2, most models degrade in Vietnamese, most sharply in BLEU and ROUGE, although the magnitude varies across models and modalities. Current medical VLMs are therefore weak cross-lingual report generators, which motivates the native Vietnamese supervision provided by the proposed ViMed-Rad.

3 ViMed-Rad: A Vietnamese Multimodal Medical Image-Report Dataset

3.1 The Proposed Dataset

Dataset Overview. ViMed-Rad pairs 93,781 radiology studies with free-text reports authored by the radiologist who signed the examination, covering X-ray, CT, MRI, and PET/CT from two hospitals (Table 3). Three properties distinguish it from existing report resources. First, every study is native DICOM exported from the hospital archive with its acquisition metadata intact, rather than a screenshot, a re-captured figure, or a re-encoded image, which preserves the intensity scales, geometry, and sequence information that volumetric encoders depend on. Second, images and reports are linked through the request identifier shared by the radiology information and archiving systems, never by patient and date, so a report cannot be attached to a different examination of the same patient. Third, X-ray, CT, and MRI are drawn from the same institution and reporting convention, so



REGION	CASES (SHARE)
Chest	37,369 (60.1%)
Spine	8,442 (13.6%)
Lower extremity	7,298 (11.7%)
Upper extremity	1,721 (2.8%)
Abdomen and pelvis	2,928 (4.7%)
Bedside	3,058 (4.9%)
Others/Unknown	1,332 (2.1%)

REGION	CASES (SHARE)
Chest/lung	5,428 (41.0%)
Brain	2,707 (20.4%)
Abdomen and pelvis	1,979 (14.9%)
Urinary system	811 (6.1%)
Heart/coronary arteries	483 (3.6%)
Maxillofacial region	363 (2.7%)
Others/Unknown	1,476 (11.1%)

REGION	CASES (SHARE)
Lumbar spine	3,689 (30.7%)
Head	2,749 (22.9%)
Knee	1,216 (10.1%)
Brain	1,004 (8.4%)
Cervical spine	947 (7.9%)
Shoulder	672 (5.6%)
Others/Unknown	1,733 (14.4%)

Figure 1: Distribution of examinations by anatomical region across modalities. The six most frequent regions are shown. PET/CT is excluded as all examinations are whole-body.

comparisons among these three modalities isolate the effect of modality rather than of data source or writing style; PET/CT is contributed by a second national referral hospital, integrating the ViMed-PET collection Nguyen et al. (2025), under a comparable but distinct reporting convention. The resulting resource contains 28.8M slices across diverse anatomical regions and acquisition protocols, and is used under the applicable ethical and data-protection requirements, with ethical approval and consent in place.

Anatomical and Protocol Diversity. The dataset covers diverse anatomical regions and acquisition protocols across modalities. For X-ray and CT, anatomical regions are determined from the examination protocol, while MRI regions are assigned at the patient level by majority voting over `BodyPartExamined` across series. Figure 1 shows the six most frequent regions for each modality, with the remainder grouped as Others. X-ray is dominated by chest examinations, CT by chest/lung and brain, and MRI by lumbar spine and head. PET/CT is not included because all examinations are whole-body.

Data Pre-processing. Reports are standardized, parsed into findings and conclusions, and de-identified. X-ray DICOM images are converted to 8-bit grayscale using their intensity metadata. CT slices are grouped into series, filtered for non-diagnostic series, and reconstructed into 3D NIfTI volumes with `dicom2nifti`. MRI series are reconstructed, mapped to T1, T2, FLAIR or T1c, resampled to 1mm^3, skull-stripped with HD-BET Isensee et al. (2019) and split into 128^3 patches, with missing sequences zero-filled and flagged by an availability mask. PET/CT follows the protocol of Nguyen et al. (2025), partitioning whole-body volumes into head-neck, chest and abdomen-pelvis. Appendix gives the full statistics and processing details.

3.2 Benchmark Tasks

ViMed-Bench defines four tasks over ViMed-Rad, report generation and three report-grounded VQA tasks, together with a cross-lingual alignment set built from public data. Every modality is split at the patient level, and images from one study stay grouped and share a single reference report, so no study is divided across splits. Appendix gives the per-modality counts for every task and split.

Report Generation. The primary task pairs each study with its signed report, giving 54,608 cases across the four modalities. PET/CT reports are dictated region by region, so each whole-body scan and its report are divided into head and neck, chest, and abdomen and pelvis pairs, which keeps every target report within the span the radiologist actually described.

Report-grounded VQA. Report generation alone confounds visual understanding with fluency in long clinical prose, so three question-answering tasks probe the same content in shorter form. Questions are generated from the imaging indication, radiological description, and diagnostic conclusion of the paired report, and must preserve its negation and uncertainty, so a model cannot score well by ignoring hedged or denied findings. The tasks differ in how much scaffolding they leave the model:

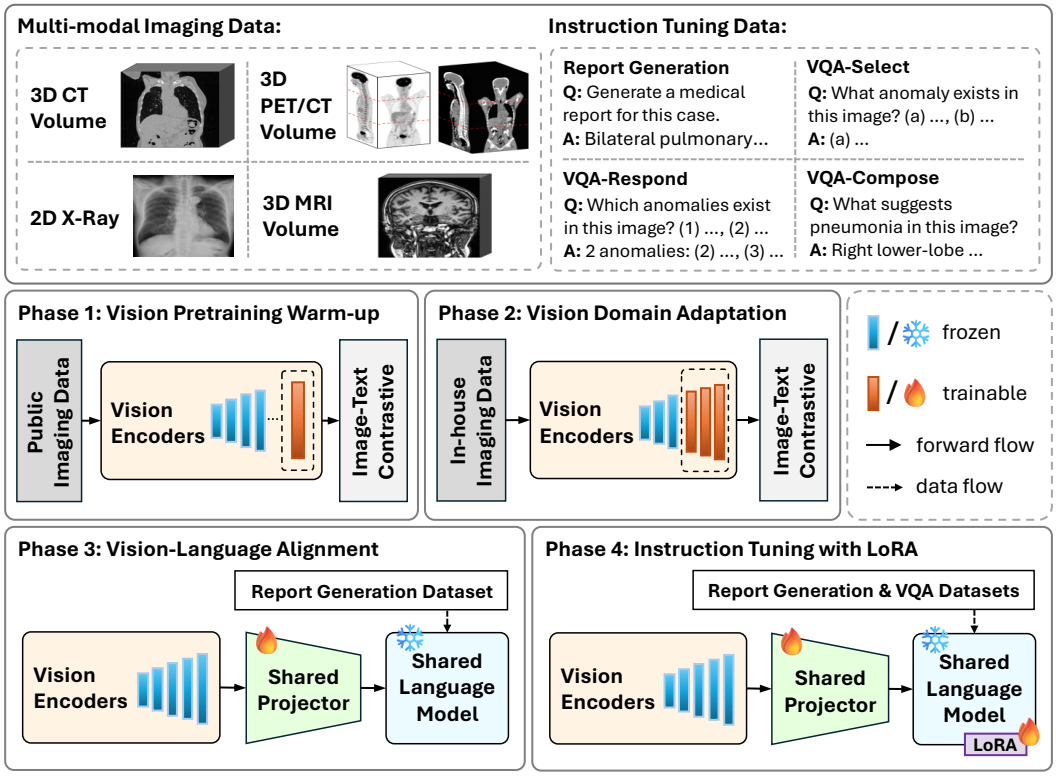


Figure 2: **Overview of the proposed ViMed-Bench. Top:** four imaging modalities paired with report generation and report-grounded VQA tasks. **Bottom:** the four training phases from contrastive encoder adaptation to instruction tuning with Low-Rank Adaptation (LoRA).

- **VQA-Select** (25,341 cases) asks modality-aware multiple-choice questions on acquisition, anatomy, and abnormal findings, with four randomly ordered candidates.
- **VQA-Respond** (23,876 cases) removes the candidates and requires a concise open-ended answer.
- **VQA-Compose** (20,290 cases) draws on structured findings extracted from the reference report, so one question can combine anatomy, presence, location, size, morphology, and certainty; each item carries a difficulty score from 1 to 4.

Cross-lingual Alignment. To test whether public English supervision transfers, we translate MIMIC-CXR Johnson et al. (2024) and CT-RATE Hamamci et al. (2026) reports into Vietnamese while preserving the original image-report pairing, giving 41,077 X-ray and 10,636 CT cases. The set supports contrastive alignment and report generation for the two modalities with comparable public resources, and MRI and PET/CT are excluded for lack of an equivalent.

4 ViMed-Bench: Benchmark for Vietnamese Multimodal Medical Vision-Language Learning

Unified Architecture. A single shared encoder is ruled out by the differences in dimensionality and structure across X-ray, CT, MRI, and PET/CT, while a separate model for each modality forfeits the clinical language they have in common. We therefore keep the encoder modality-specific and share everything above it, so that one decoder and one set of task instructions serve all four modalities. For X-ray we use MedSigLIP from MedGemma Sellergren et al. (2026), a medical image-text pretrained SigLIP encoder. For CT we use the ViT-3D of M3D Bai et al. (2024), which models volumetric context, and for MRI we use M2FTrans Shi et al. (2024), which fuses the T1, T2, FLAIR, and T1c sequences and tolerates missing ones. For PET/CT we adopt CT-ViT Hamamci et al. (2026), initialized from the checkpoint released by Nguyen et al. (2025), to encode tracer uptake alongside anatomy. All four feed a shared projector and Qwen3-4B-Instruct-2507 Yang et al. (2025).

Progressive Fine-tuning The four modalities differ by an order of magnitude in available cases, so training the encoders and the language model end-to-end would let the largest dominate the gradient, while full updates would erase the pretrained visual knowledge the smallest ones rely on. We therefore adapt the components in four phases (Figure 2), unfreezing at each phase only what that phase requires. In both contrastive phases we draw study-disjoint batches with at most one image-report pair per study, so two views sharing a report are never treated as negatives.

- **Phase 1: Vision Pretraining Warm-up.** Encoders pretrained on English clinical text may be weakly aligned with Vietnamese despite strong medical representations, so we run a lightweight contrastive warm-up on the Cross-Lingual Alignment set. We update only the last layer of each X-ray and CT encoder while keeping MRI and PET encoders frozen, as no comparable public data are available.
- **Phase 2: Visual Domain Adaptation.** Phase 1 leaves the visual distribution shift to in-house clinical data unaddressed, so we adapt the encoders contrastively on the Report Generation training split, using paired images and reports without task supervision. We unfreeze the last three to four layers of all four encoders, adapting each modality to the local acquisition distribution while retaining the pretrained features below.
- **Phase 3: Vision-Language Alignment.** Modality-specific adapters project the four frozen encoders into a shared 1024-dimensional space, followed by a Perceiver resampler that compresses them into 128 visual tokens and a projector that maps them into the language model space with learned modality embeddings and tags. We freeze the encoders and language model, training only adapters, resampler, projector, and tag embeddings on Report Generation data.
- **Phase 4: Instruction Tuning with LoRA.** We apply Low-rank Adaptation (LoRA) Hu et al. (2022) to every attention and MLP projection layer of the language model and optimize them jointly with the visual interface, at a lower learning rate for the latter, across Report Generation and three VQA tasks. X-ray contributes roughly twelve times as many report-generation cases as MRI, so we apply sampling quotas over modalities and tasks to keep any one from dominating.

Evaluation Metrics. We score reports on two axes. Text-level agreement uses BLEU-4 (B-4), ROUGE-1 (R-1), ROUGE-L (R-L), and BERTScore (BERT), covering n-gram precision, lexical overlap, sequence-level similarity, and contextual semantic similarity. These reward wording rather than clinical content, so we score each modality against the radiologist-designed finding schema of Appendix, extracting findings and their attributes from the reference and the prediction independently and reporting macro-averaged Precision (P), Recall (R), and F1 over the finding labels.

5 EXPERIMENTS

Model Variants. We develop three variants to study modality specialization, multimodal joint learning, and supervision scaling. **Single** trains a separate model for each modality on ViMed-Rad alone. **Joint-I** trains one unified model across all four ViMed-Rad modalities, isolating the effect of joint multimodal learning. **Joint-F** adds the Cross-Lingual Alignment set to evaluate the impact of larger and more diverse supervision. Together, these variants examine whether supervised adaptation improves Vietnamese report generation, whether joint training can match modality-specific models, and how data scaling affects generalization. All experimental settings are provided in the Appendix.

Comparison with Existing Medical VLMs. We compare the three variants with existing medical VLMs and GPT-4o on the proposed ViMed-Rad test set (Table 4). Existing models are evaluated in a few-shot setting, while the proposed variants are fine-tuned on Vietnamese clinical reports. All proposed models consistently outperform prior VLMs across four modalities, demonstrating the value of supervised adaptation for low-resource report generation. Compared with GPT-4o, the best variants achieve 1.42×, 1.28×, 1.17×, and 1.12× higher BERTScore on X-ray, CT, MRI, and PET/CT, respectively. On CT, Joint-F further achieves 24.50× higher BLEU-4 and 2.38× higher ROUGE-L, showing improved semantic alignment and report structure generation.

Effect of Joint Multimodal Training and Supervision Scaling. We next analyze the contribution of different training strategies. Single achieves strong modality-specific performance but requires maintaining separate models. Joint-I provides a competitive unified alternative, achieving comparable semantic quality while learning from all modalities jointly. Scaling from Joint-I to Joint-F further improves generation performance on several modalities when additional public supervision

Table 4: Report generation and clinical evaluation on the proposed ViMed-Rad. **Bold** and underline indicate the best and second-best results within each modality and metric. Dashes (–) denote results that are unavailable or below 0.01. R-1 results are reported in Appendix.

NLP-based Metrics on Report Generation												
	X-ray			**CT**			**MRI**			**PET/CT**		
Model	B-4	R-L	BERT	B-4	R-L	BERT	B-4	R-L	BERT	B-4	R-L	BERT
RadFM	0.02	0.18	12.21	0.43	5.83	56.73	0.11	1.69	57.26	0.06	28.33	69.49
LLaVA-Med	2.53	23.62	56.28	4.04	22.41	64.16	0.05	3.57	59.20	0.01	27.89	64.63
M3D	–	–	–	3.41	33.05	66.98	–	–	–	0.04	23.53	67.21
MedVersa	2.22	26.00	57.30	0.01	0.36	55.54	0.01	0.26	53.27	0.01	0.24	56.98
MedGemma 1.5	0.39	9.19	50.34	5.64	35.89	70.93	2.19	14.30	68.65	1.38	15.03	63.53
GPT-4o	2.26	23.50	55.85	1.88	26.69	66.29	29.66	39.00	76.88	31.12	52.76	81.09
Single (Ours)	14.50	**41.08**	78.72	43.75	63.02	84.89	**53.91**	65.49	89.55	52.24	**72.16**	**90.73**
Joint-I (Ours)	10.64	40.56	**79.21**	42.11	61.97	84.66	52.98	69.16	89.97	49.40	71.15	90.44
Joint-F (Ours)	**16.62**	39.89	78.18	**46.06**	**63.56**	**85.07**	52.96	**69.30**	**90.10**	**53.30**	70.96	90.14
Clinical Evaluation												
	X-ray			**CT**			**MRI**			**PET/CT**		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT-4o	44.81	45.17	43.65	45.95	49.27	46.77	43.55	43.97	42.74	10.08	5.90	6.97
DeepSeek-v4.1	46.03	42.41	40.81	48.21	49.95	47.83	45.36	44.85	43.84	26.83	18.20	20.16
Single (Ours)	**48.66**	**49.74**	**48.82**	**51.03**	**50.79**	**48.86**	**50.72**	**49.55**	**48.67**	68.80	**56.24**	58.63
Joint-I (Ours)	46.04	46.83	46.42	46.31	49.86	46.74	46.09	46.92	44.95	**79.47**	52.20	**58.95**
Joint-F (Ours)	48.05	48.45	48.13	46.65	50.01	47.69	48.87	47.44	44.98	67.99	55.46	57.59

Table 5: Generalization evaluation on translated public test sets. Non-Vietnamese datasets are translated into Vietnamese. PET/CT is unavailable due to no public counterpart.

Modalily	**Model**	**B-4**	**R-1**	**R-L**	**BERT**
X-ray	MedGemma 1.5	0.82	13.33	8.88	63.32
	Joint-F (Ours)	**8.11**	**31.93**	**24.45**	**74.69**
CT	LLaVA-Med	0.60	8.38	4.23	61.10
	Joint-F (Ours)	**28.23**	**58.09**	**34.73**	**77.97**
MRI	M3D	0.21	1.57	4.82	61.84
	Joint-F (Ours)	**10.83**	**39.79**	**27.96**	**73.33**

is introduced. On X-ray, Joint-F improves BLEU-4 from 10.64 to 16.62 (+56.2%), while CT BLEU-4 increases from 42.11 to 46.06 (+9.4%).

Clinical Evaluation. To assess clinical relevance beyond lexical similarity, we report finding-level precision, recall, and F1 against two general-purpose baselines, GPT-4o and DeepSeek. All fine-tuned variants exceed GPT-4o on every modality, with the best obtaining 1.12×, 1.04×, 1.14×, and 8.46× higher F1 on X-ray, CT, MRI, and PET/CT. DeepSeek is the stronger baseline, coming within 1.0 F1 point of Single on CT and exceeding both joint models there, which suggests that templated CT reporting is partly recoverable without volumetric supervision. It trails the best variant elsewhere, by 8.0 points on X-ray, 4.8 on MRI, and 38.8 on PET/CT, where lesion-level reporting demands supervision a general-purpose model lacks. Single achieves the strongest modality-specific performance, while Joint-I and Joint-F remain competitive using a single multimodal architecture.

Generalization to Vietnamese Public Benchmarks. We evaluate Joint-F on Vietnamese translations of public radiology benchmarks, covering X-ray, CT, and MRI modalities. As reported in Table 5, Joint-F consistently outperforms zero-shot baselines across all modalities. The largest gains are observed on CT and MRI, with BLEU-4 improvements of 47.1× and 51.6×, respectively, together with consistent improvements in ROUGE and BERTScore. These results demonstrate effective generalization beyond the in-house clinical distribution.

Visual Question Answering with Cumulative Training Data. To examine whether structured supervision improves multimodal reasoning, we progressively add VQA-Select, VQA-Respond, and

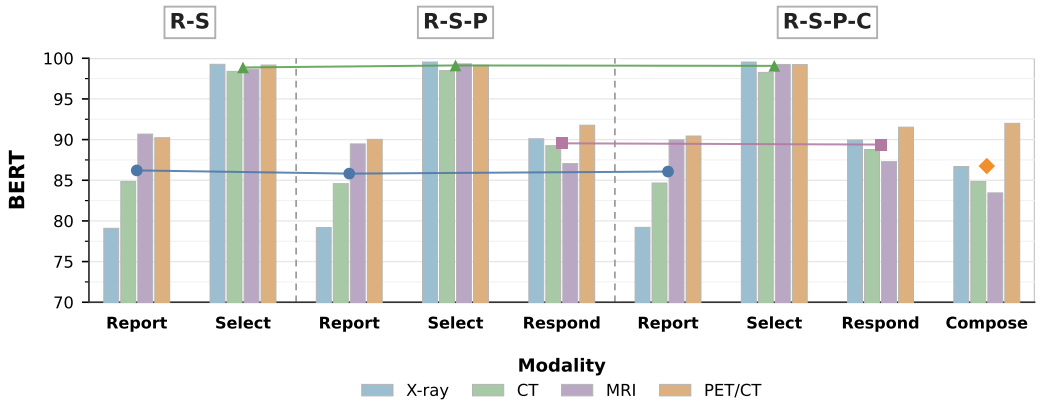


Figure 3: Cumulative training across Report Generation and VQA, measured by BERTScore. R, S, P, and C denote Report Generation, VQA-Select, VQA-Respond, and VQA-Compose; R-S, R-S-P, and R-S-P-C indicate progressively expanded supervision. Each line follows one task across the cumulative settings, averaged over the four imaging modalities. Per-modality scores for each metric and the baselines are reported in the Appendix.

VQA-Compose supervision on top of Report Generation. Each configuration is trained from the same initial checkpoint. We use BERTScore as the primary metric for semantic comparison across tasks with different output formats. As shown in Figure 3, the results reveal two insights. First, the consistent performance gap across tasks reflects increasing reasoning difficulty from report generation to selective, open-ended, and compositional VQA, supporting the proposed hierarchical design. Second, cumulative training improves complex reasoning while preserving earlier capabilities, demonstrating effective incremental learning.

6 Conclusion

We introduced **ViMed-Rad**, a Vietnamese radiology dataset of 93,781 native-DICOM studies paired with expert-authored reports across X-ray, CT, MRI, and PET/CT. Building on it, we established **ViMed-Bench**, which couples report generation and report-grounded VQA tasks of increasing difficulty with a unified framework over four modality-specific encoders and a shared language model. Benchmarking existing medical VLMs exposes a Vietnamese-English gap for most models, with ROUGE-L falling by more than 75% for several of them and BERTScore by up to about a quarter on identical images, whereas supervised adaptation on ViMed-Rad improves BERTScore by 1.12-1.42$\times$ over GPT-4o and raises clinical F1 in all four modalities. Comparing joint and modality-specific models on identical splits and evaluation protocols makes the cost of unification directly measurable: a single model retains most of the clinical accuracy of four modality-specific counterparts, indicating that planar, volumetric, and functional imaging can be served within a single framework rather than by four separate models.

Limitations and Future Work. ViMed-Rad draws on national-level tertiary hospitals with broad referral catchments, but its scanners and reporting conventions remain those of a few centres, and the clinical metrics rely on an LLM extractor that radiologists validated on a sample. Generated reports are not a substitute for radiologist review, and the Vietnamese-English gap reported here suggests that models validated in English should not be assumed safe in another clinical language. Three questions follow: whether modality-aware routing recovers the clinical precision that joint training gives up, how far cross-modal transfer substitutes for in-modality data when none is public, as for PET/CT in most languages, and how clinical evaluation can reach radiologist-level reliability beyond English radiography.

AI Use Statement

We comply with the ICLR 2027 policy regarding the use of generative AI in research and manuscript preparation.

Ethics Statement

This research follows the ICLR Code of Ethics. As our study involves retrospective medical imaging data, we recognize the importance of patient confidentiality, data protection, and responsible data use. We adhere to applicable ethical standards and data usage requirements. The proposed dataset and framework are intended solely for research purposes and are not designed for direct clinical decision-making. We acknowledge potential concerns regarding privacy, dataset bias, and unintended clinical applications, and emphasize the responsible use of our research outputs.

Reproducibility Statement

To support reproducibility, we provide comprehensive descriptions of dataset construction, preprocessing, model architecture, training procedures, and evaluation protocols in the main paper and appendix. We intend to make our implementation and instructions for reproducing the experiments publicly available upon publication. Further details on experimental settings, hyperparameters, and computational resources are documented to facilitate independent verification of our findings.

References

Fan Bai, Yuxin Du, Tiejun Huang, Max Q. H. Meng, and Bo Zhao. M3D: Advancing 3D medical image analysis with multi-modal large language models. *Computing Research Repository, arXiv Preprints,* arXiv:2404.00578, pp. 1–35, 2024.

Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, et al. Merlin: a computed tomography vision-language foundation model and dataset. *Nature*, 652(8112):1318–1328, 2026. doi: 10.1038/s41586-026-10181-8.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayò. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101797. URL `https://www.sciencedirect.com/science/article/pii/S1361841520301614`.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *Computing Research Repository, arXiv Preprints,* arXiv:2405.19538, pp. 1–13, 2024.

Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. Bimcv-r: A landmark dataset for 3D CT text-image retrieval. In *Proceedings of the 2024 International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 124–134, 2024.

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

Forithmus. MR-RATE: A vision-language foundation model and dataset for magnetic resonance imaging, 2026. URL `https://huggingface.co/datasets/Forithmus/MR-RATE`. Dataset; accompanying paper not yet released.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, 10:1610–1628, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LORA: Low-rank adaptation of large language models. *Proceedings of the 2022 International Conference on Learning Representations*, 1(2):1–13, 2022.

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

Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, Martin Bendszus, Klaus H. Maier-Hein, and Philipp Kickingereder. Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, 40(17):4952–4964, 2019.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. *Advances in Neural Information Processing Systems — Datasets and Benchmarks Track*, 34:1–12, 2021.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR Database. *PhysioNet*, 2024. URL https://doi.org/10.13026/4jqj-jw95. Version 2.1.0.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):1–10, 2018.

Jiayu Lei, Xiaoman Zhang, Chaoyi Wu, Lisong Dai, Ya Zhang, Yanyong Zhang, Yanfeng Wang, Weidi Xie, and Yuehua Li. Autorg-brain: Grounded report generation for brain mri. *Computing Research Repository, arXiv Preprints,* arXiv:2407.16684, 2024. URL https://arxiv.org/abs/2407.16684.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:1–17, 2023.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. In *Proceedings of the 2023 Medical Image Computing and Computer Assisted Intervention*, pp. 525–536, 2023.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *Proceedings of the 18th IEEE International Symposium on Biomedical Imaging*, pp. 1650–1654. IEEE, 2021.

National Electrical Manufacturers Association. Digital imaging and communications in medicine (DICOM) standard, 2025. URL https://www.dicomstandard.org/current. PS3.1–PS3.22.

Duy Vu Minh Nguyen, Chinh Thanh Truong, Phuc Hoang Tran, Hung Tuan Le, Nguyen Van-Thanh Dat, Trung Hieu Pham, and Kiet Van Nguyen. ViX-Ray: A vietnamese chest X-Ray dataset for vision-language models. In *Proceedings of the 15th Language Resources and Evaluation Conference*, pp. 3055–3074, 2026.

Huu Tien Nguyen, Dac Thai Nguyen, Trung Thanh Nguyen, Thao Nguyen Truong, Hieu Pham, Johan Barthelemy, Tran Minh Quan, Quoc Viet Hung Nguyen, Thanh Tam Nguyen, Mai Hong Son, et al. Toward a vision-language foundation model for medical data: Multimodal dataset and benchmarks for vietnamese pet/ct report generation. *Advances in Neural Information Processing Systems — Datasets and Benchmarks Track*, 38:1–32, 2025.

Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, Shujian Gao, Yuandong Pu, Huihui Xu, Chenhui Gou, Ziyan Huang, Yi Xin, Qi Qin, Diping Song, Bin Fu, Guang Yang, Yuanfeng Ji, Tianbin Li, Yanzhou Su, Jin Ye, Shixiang Tang, Zhongying Deng, Lihao Liu, Ming Hu, and Junjun He. UniMedVL: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. *Computing Research Repository, arXiv Preprints,* arXiv:2510.15710, pp. 1–46, 2025.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing 2024*, pp. 374–390, 2024.

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

Marco Salmè, Rosa Sicilia, Paolo Soda, and Valerio Guarrasi. Evaluating vision language model adaptations for radiology report generation in low-resource languages. *Computing Research Repository, arXiv Preprints,* arXiv:2505.01096, pp. 1–8, 2025.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *Computing Research Repository, arXiv Preprints,* arXiv:2604.05081, pp. 1–23, 2026.

Junjie Shi, Li Yu, Qimin Cheng, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan. M^2FTrans: Modality-masked fusion transformer for incomplete multi-modality brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(1):379–390, 2024.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117/.

Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. Med-UniC: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36:1–19, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications*, 16(7866):1–22, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Can Hu, Siwei Pan, Jie Tan, Tingyang Xu, Deli Zhao, Hao Zhang, and Yu Rong. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2026. Early Access.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *Computing Research Repository, arXiv Preprints,* arXiv:2505.09388, pp. 1–35, 2025.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):1–17, 2018.

Yichi Zhang, Wenbo Zhang, Zehui Ling, Gang Feng, Sisi Peng, Deshu Chen, Yuchen Liu, Hongwei Zhang, Shuqi Wang, Lanlan Li, Limei Han, Yuan Cheng, Zixin Hu, Yuan Qi, and Le Xue. PET2Rep: Towards vision-language model-drived automated radiology report generation for positron emission tomography. *Proceedings of the 2026 AAAI Conference on Artificial Intelligence*, 40(15):12897–12906, 2026.

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

Hong-Yu Zhou, Juliòn Nicolòs Acosta, Subathra Adithan, Suvrankar Datta, Eric J. Topol, and Pranav Rajpurkar. Medversa: A generalist foundation model for medical image interpretation. *Computing Research Repository, arXiv Preprints,* arXiv:2405.07988, pp. 1–52, 2025.

Xun Zhu, Ying Hu, Fanbin Mo, Miao Li, and Ji Wu. Uni-Med: A unified medical generalist foundation model for multi-task learning via connector-MoE. *Advances in Neural Information Processing Systems*, 37:1–25, 2024.

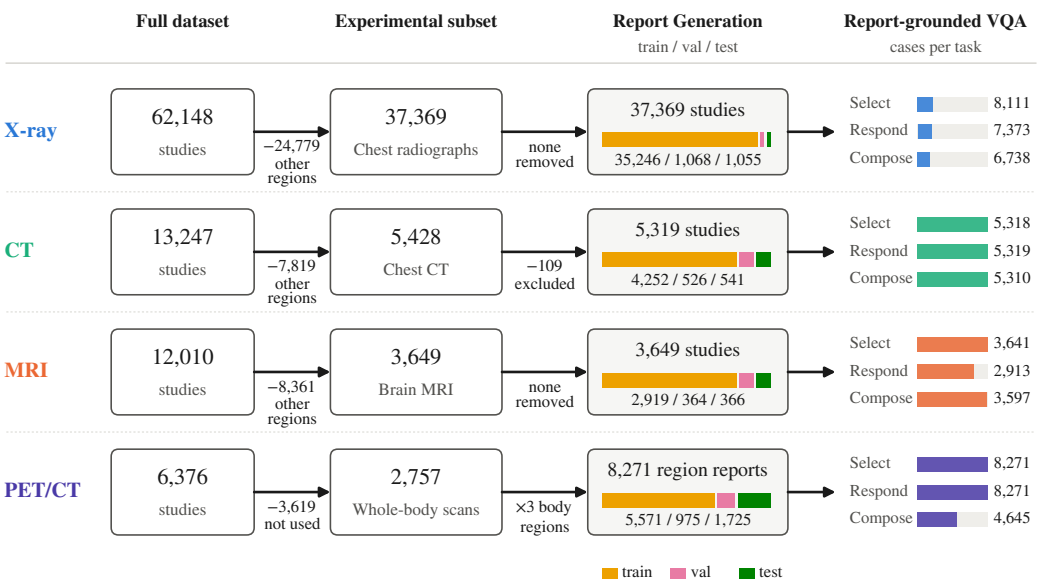


Figure 4: From the full dataset to the benchmark splits. Negative numbers under the arrows count studies that are not used in the experiments; they remain in the released dataset. VQA bars are relative to Report Generation.

A Additional Dataset Details

A.1 Data Sources and Cohort Construction

Experimental subset. ViMed-Rad is released with all anatomical regions (Figure 1). For the experiments, we restrict each modality to one region, which reduces anatomical variability for a given amount of data and allows each modality to be scored with a region-specific clinical schema (Appendix F). The chest X-ray and chest CT subsets contain the requests whose ordered procedure covers the chest (37,369 and 5,428 studies). The brain MRI subset contains the requests for brain MRI with at least one series mapped to T1, T2, FLAIR or T1c (3,649 of 3,657 requests); 99.3% of them carry a `HEAD` or `BRAIN` `BodyPartExamined` tag. For PET/CT, we use the 2,757 of 6,376 whole-body scans. PET/CT reports describe the findings region by region, so each scan and its report are divided into head and neck, chest, and abdomen and pelvis pairs. The PET/CT counts of the benchmark tasks (Section 3.2) are therefore region reports; for example, the 5,571 training reports come from 1,857 scans. Of the 5,428 chest CT studies, 5,319 enter the benchmark splits. Figure 4 and Table 6 give the resulting counts.

Splits. The DICOM `PatientID` is replaced by a different pseudonymous code in every study, so patients are grouped by the hashed RIS patient identifier, and every modality is split at the patient level. Under this key, 10.3%, 9.8% and 4.6% of the X-ray, CT and MRI patients have more than one study in the same modality.

Public data. The Cross-Lingual Alignment set contains 41,077 MIMIC-CXR studies (36,000 / 1,808 / 3,269 for training, validation and test) and 10,636 CT-RATE volumes (7,597 / 1,501 / 1,538, from 3,154 / 629 / 675 patients). CT-RATE does not release test reports, so its official validation set (1,304 patients) is split by patient into our validation and test sets. Joint-F additionally uses 800 / 100 / 100 MR-RATE studies with translated reports; they are not part of the Cross-Lingual Alignment set, and Table 5 reports their test results.

Table 6: Cases per task and split (train / val / test). A case is a study for X-ray, CT and MRI, and a region report for PET/CT.

Modality	Full dataset	Subset	Report Gen.	VQA-Select	VQA-Respond	VQA-Compose
X-ray	62,148	37,369	35,246 / 1,068 / 1,055	5,991 / 1,066 / 1,054	5,250 / 1,068 / 1,055	4,769 / 962 / 1,007
CT	13,247	5,428	4,252 / 526 / 541	4,251 / 526 / 541	4,252 / 526 / 541	4,243 / 526 / 541
MRI	12,010	3,649	2,919 / 364 / 366	2,911 / 364 / 366	2,329 / 291 / 293	2,871 / 363 / 363
PET/CT	6,376	2,757[†]	5,571 / 975 / 1,725	5,571 / 975 / 1,725	5,571 / 975 / 1,725	2,754 / 166 / 1,725

[†]Scans; each scan contributes three region reports.

Table 7: Cohort and acquisition statistics not reported in Table 3. Acquisition parameters come from one DICOM header per series.

	X-ray	CT	MRI
Patients with >1 study (%)	10.3	9.8	4.6
Age, median [IQR] (years)	46 [33, 59]	55 [42, 65]	51 [39, 62]
Age missing (%)	6.2	0.0	0.0
In-plane matrix (median)	3048×2439	512×512	320×312
Pixel spacing, mm (median)	0.125	0.646	0.729
Slice thickness, mm (median [IQR])	N/A	1.50 [1.00, 5.00]	4.0 [3.0, 5.0]
Tube voltage, kVp (median)	85	120	N/A
Field strength (share of studies)	N/A	N/A	1.5 T 65%, 3 T 35%
Vendor (share of series)	Canon 96%, Agfa 4%	Siemens 60%, GE 40%	Siemens 100%

A.2 Dataset Statistics

Table 7 complements Table 3 with statistics computed on the full cohorts. Two properties matter for the benchmark. First, the scanners are few: nearly all radiographs come from one vendor and all MRI from another, so the benchmark does not measure robustness to acquisition hardware. Second, CT slice thickness is bimodal, with 43.6% of series at 1.25 mm or less and most of the rest at 3 to 5 mm.

Reports. Table 8 summarizes the reports of the full cohorts. X-ray reports are mostly impression-only: 24.1% have a findings section, and 53.1% have a normal impression. They are also highly repetitive: 71.2% of X-ray studies share their report verbatim with another study, and the ten most frequent texts cover 31.3%. This repetition may inflate lexical scores on X-ray when models generate common report templates, which is why we report clinical metrics alongside BLEU and ROUGE. Hedged statements occur in about a quarter of CT and MRI reports, mostly *theo dûi* (to be monitored, used for "suspected"), *ngh n* (suggestive of) and *cha loi tr* (cannot be excluded).

Report length at evaluation. About 29% of the CT benchmark reports exceed 1,024 characters; we truncate neither references nor predictions, since truncation would remove the impression. In the MRI benchmark files, 8.2% of the reports are cut at exactly 1,024 characters. We keep these references and truncate MRI predictions to 1,024 characters before scoring. The cut comes from the export of the benchmark files; the RIS reports themselves reach 768 tokens.

MRI sequences. In the brain MRI subset, T2 and FLAIR are available for almost every study (99.7% and 99.8%), T1 for 92.1% and T1c for 33.8%. In the training split, the T1c mask agrees with the contrast statement of the report's protocol sentence for 92.5% of the studies where that sentence is informative (2,333 of 2,522), but 99.8% of the reports also name sequences that are not encoder inputs (DWI, ADC, GRE, SWI, TOF). We therefore do not list the available sequences in the prompt, since the list would contradict the reference answer to questions about the protocol.

VQA. VQA-Select covers acquisition, anatomy and abnormal findings. Every X-ray study has five VQA-Select questions, one per template: view, body region, anatomical structure, presence of an abnormality and its location (Figure 5a). CT VQA-Select covers nine categories of findings, the largest (abnormality) with 19% of the questions (Figure 5b). Every VQA-Compose question carries a difficulty level from 1 to 4 (Appendix B.1). CT VQA-Compose contains only levels 3 and 4 by construction, whereas most X-ray questions are at level 2 and most MRI questions at level

Table 8: Report statistics of the full cohorts. Length in tokens of the Qwen3 tokenizer. Cues are matched with a Vietnamese lexicon; "normal impression" is a rule-based estimate.

	X-ray	CT	MRI
Findings section present (%)	24.1	98.9	100.0
Report length, median [IQR]	48 [32, 125]	282 [204, 373]	392 [302, 465]
Report length, 95th percentile / max	168 / 368	486 / 1,008	604 / 768
Mostly upper-case text (%)	74.1	0.5	0.0
Negation cue (%)	68.3	97.0	82.1
Uncertainty cue (%)	7.1	26.8	25.3
Normal impression (%)	53.1	10.6	6.7
Report shared verbatim with another study (%)	71.2	14.4	16.6
Studies covered by the 10 most frequent reports (%)	31.3	9.5	8.6

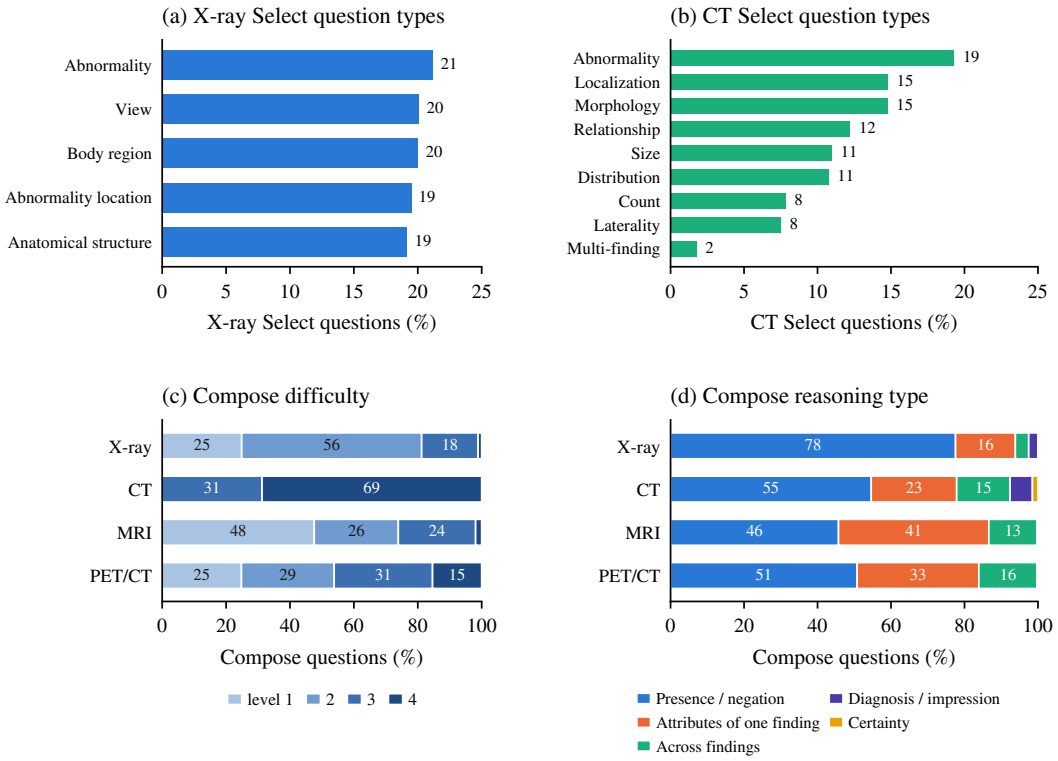


Figure 5: Composition of the VQA sets over all splits; MRI VQA-Compose is counted on its test split, the only one whose questions keep their labels. (a) X-ray VQA-Select question types; every X-ray study has one question per template, and the type is read from the template. (b) CT VQA-Select categories. (c) Difficulty of VQA-Compose, assigned by the question generator with one rubric for all modalities: level 1 asks for a single fact such as the presence of a finding; level 2 for one attribute of a finding or a finding with its location; level 3 for two or more attributes of one finding or for two findings combined; level 4 for three or more attributes, a synthesis across findings, or diagnostic reasoning under uncertainty. CT questions are restricted to levels 3 and 4. (d) Reasoning types of VQA-Compose, grouped from the generator labels.

1 (Figure 5c). In all modalities, presence and negation questions form the largest VQA-Compose group (46% to 78%); for X-ray and CT, most of them ask about a finding that the report explicitly rules out.

A.3 Image Processing

X-ray. Radiographs are converted from DICOM to 8-bit grayscale JPEG using the DICOM intensity metadata: rescale and VOI transformations, then min-max normalization. All series that record a

Table 9: Model inputs. Tokens are the encoder outputs passed to the shared resampler, which maps each study to 128 visual tokens, or 32 per available MRI sequence.

Modality	Encoder	Input	Intensity	Tokens
X-ray	MedSigLIP	up to 4 images, 448×448	min-max	256 per image
CT	ViT-3D	one series, $32 \times 256 \times 256$	percentile window	256
MRI	M2FTrans	4 sequences, 128^3 patches	z-score in brain mask	4,096 per sequence
PET/CT	CT-ViT	one region, $140 \times 480 \times 480$	scaled to $[0, 1]$	1,008

photometric interpretation are `MONOCHROME2`. The model receives up to four radiographs per study (Table 9).

CT. Series are converted to NIfTI with `dicom2nifti` and kept only if they form a geometrically consistent volume, which holds for 58,127 of 61,871 series (93.9%). Most excluded series are bolus-tracking scans (1,818) or ECG-gated cardiac series (851); the rest are conversion failures, derived renderings, scouts and dose reports. Voxels are kept in Hounsfield units.

MRI. Series are assigned by name to the four encoder channels (T1, T2, FLAIR and T1c), at most one per channel. Each volume is reoriented to RAS+, resampled to 1 mm isotropic spacing, skull-stripped with HD-BET and cropped to the brain mask.

Report normalization. In-house CT reports contain spelling errors, spacing issues, and fragmented formatting. GPT-4o-mini (temperature 0) was used to rewrite reports into fluent medical prose while preserving all findings, numbers, technical terms, and uncertainty expressions without adding or removing clinical content. Reports with altered content were reverted to the original text (7.1% of 5,314 cases). The released dataset provides the normalized reports.

A.4 De-identification

DICOM headers. DICOM headers contain patient identifiers, study information, acquisition parameters, and scanner metadata. We removed direct identifiers such as `PatientName` and `PatientID`, replacing them with study-specific codes. Remaining metadata used for model development include acquisition dates, series information, and imaging-related attributes.

Reports. A pattern audit of all reports finds no phone number, e-mail address, URL or digit string of seven or more digits. Date strings occur in a small fraction of the reports, almost always in comparisons with a prior examination, and person titles followed by a name are rare.

B VQA Dataset Construction

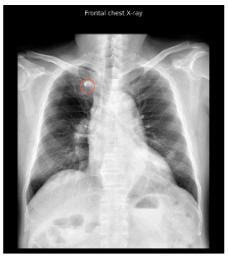
B.1 Generation Procedure

All VQA items are generated from the paired report, so every answer is supported by the report. The three tasks differ in the scaffolding they leave the model. VQA-Select gives four candidates, of which one is correct. VQA-Respond removes the candidates and asks for a short free-text answer. VQA-Compose asks questions that combine several fields of the structured report, and each of them carries a difficulty level.

VQA-Compose and its difficulty levels. VQA-Compose questions are generated by GPT-4o-mini from the structured findings that the clinical schemas of Appendix F.1 extract from the reference report, together with the report text. For each question, the generator returns the answer, the reasoning type, the schema fields needed to answer it and a difficulty level, following a rubric shared by all modalities:

- Level 1: a single fact, such as the presence or absence of one finding.
- Level 2: one attribute of a finding, or a finding together with its location.

X-ray



Report.

VI

- Hai phế trường: nốt vôi 1/3 trên phổi phải.
- Bóng tim to. Động mạch chủ vôi hóa.
- Màng phổi, xương lồng ngực, vòm hoành: không thấy bất thường.

EN

- Lung fields: calcified nodule in the upper third of the right lung.
- Enlarged cardiac silhouette. Calcified aorta.
- Pleura, bony thorax, diaphragm: no abnormality.

VQA-Select.

VI

Câu hỏi: Nếu có bất thường trên phim X-quang, thì vị trí nào có thể gặp phải?
A. Vùng tim B. Vùng xương C. Vùng bụng D. Vùng phổi

Đáp án: D. Vùng phổi.

EN

Question: If an abnormality is present, where is it located?
A. Cardiac B. Skeletal C. Abdominal D. Pulmonary

Answer: D. Pulmonary region.

VQA-Respond.

VI

Câu hỏi: Kết luận của ca chụp này có mô tả bất thường nào không?

Đáp án: Kết luận mô tả bất thường là nốt vôi 1/3 trên phổi phải và bóng tim to.

EN

Question: Does the conclusion describe any abnormality?

Answer: A calcified nodule in the upper third of the right lung and an enlarged cardiac silhouette.

VQA-Compose.

VI

Câu hỏi: Tóm tắt các dấu hiệu bất thường liên quan đến tim mạch trên phim này.

Đáp án: Bóng tim to và quai động mạch chủ có vôi hóa.

EN

Question: Summarize the cardiovascular abnormalities on this radiograph.

Answer: Enlarged cardiac silhouette and a calcified aortic arch.

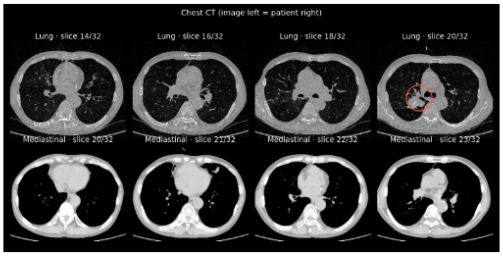
Figure 6: Report and VQA examples for chest X-ray.

- Level 3: two or more attributes of the same finding, or two findings that must be combined.
- Level 4: a rich characterization with three or more attributes, a synthesis across findings, or diagnostic reasoning that accounts for uncertainty.

B.2 Representative Examples

Figures 6 to 9 show one test case per modality. English translations are for readability; models are trained and evaluated on the Vietnamese text.

CT



Report.

VI

- Khối đậm độ mô mềm cạnh rốn phổi phải 43×30 mm, bờ tua gai, gây xẹp nhu mô lân cận.
- Vôi hóa sau phẫu thuật thùy trên phổi phải; khí phế thũng hai phổi.
- Kết luận: khối cạnh rốn phổi phải và hạch rốn phổi phải, nghĩ tổn thương thứ phát.

EN

- Right perihilar soft-tissue mass, 43×30 mm, spiculated, with adjacent atelectasis.
- Post-surgical calcification in the right upper lobe; bilateral emphysema.
- Impression: right perihilar mass with right hilar nodes, suggestive of secondary lesions.

VQA-Select.

VI

Câu hỏi: Trên CT, có khối tổn thương nào được ghi nhận không?
A. Không có khối tổn thương nào B. Có khối tổn thương đậm độ mô mềm
C. Có khối tổn thương dạng nốt D. Có khối tổn thương dạng u mỡ

Đáp án: B. Có khối tổn thương đậm độ mô mềm.

EN

Question: Is any mass identified on this CT?
A. No mass B. Soft-tissue density mass
C. Nodular mass D. Lipomatous mass

Answer: B. Soft-tissue density mass.

VQA-Respond.

VI

Câu hỏi: Kết luận chính của báo cáo là gì?

Đáp án: Khối tổn thương đậm độ mô mềm vùng cạnh rốn phổi phải và hạch rốn phổi phải, nghĩ tổn thương thứ phát.

EN

Question: What is the main conclusion of the report?

Answer: A right perihilar soft-tissue mass with right hilar lymph nodes, suggestive of secondary lesions.

VQA-Compose.

VI

Câu hỏi: Khối tổn thương đậm độ mô mềm ở vị trí nào và kích thước bao nhiêu?

Đáp án: Ở vùng cạnh rốn phổi phải, kích thước dmax 43×30 mm.

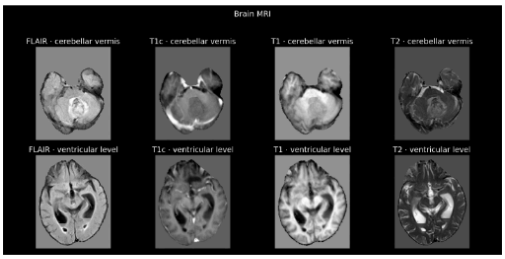
EN

Question: Where is the soft-tissue mass and what is its size?

Answer: Adjacent to the right hilum, 43×30 mm at its largest dimension.

Figure 7: Report and VQA examples for chest CT.

MRI



Report.

VI

- Mảng xuất huyết thùy nhộng tiểu não 40×30 mm, giai đoạn cấp, phù quanh tổn thương.
- Đè ép não thất IV gây giãn hệ thống não thất; xuất huyết não thất.
- Nốt nhồi máu mới vành tia trái; nhiều nốt vi xuất huyết.

EN

- Acute cerebellar vermis hemorrhage, 40×30 mm, with perilesional edema.
- Fourth-ventricle compression causing hydrocephalus; intraventricular hemorrhage.
- Acute infarct in the left corona radiata; multiple microbleeds.

VQA-Select.

VI

Câu hỏi: Trong ca chụp MRI này, có ghi nhận bất thường nào không?
A. Có khối u B. Có viêm não C. Có mảng xuất huyết D. Không ghi nhận bất thường

Đáp án: C. Có mảng xuất huyết.

EN

Question: Is any abnormality observed in this MRI?
A. Tumor B. Encephalitis C. Hemorrhagic lesion D. No abnormality

Answer: C. Hemorrhagic lesion.

VQA-Respond.

VI

Câu hỏi: Các tổn thương trên ca chụp MRI này nằm ở vị trí nào?

Đáp án: Thùy nhộng tiểu não, nhân bèo hai bên và chất trắng sâu rải rác hai bán cầu.

EN

Question: Where are the lesions located?

Answer: Cerebellar vermis, both lentiform nuclei, and scattered deep white matter of both hemispheres.

VQA-Compose.

VI

Câu hỏi: Có nhồi máu não và xuất huyết não không?

Đáp án: Có cả nhồi máu não và xuất huyết não.

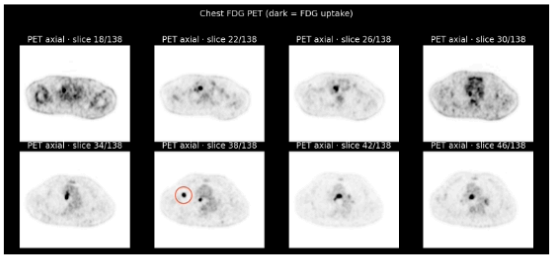
EN

Question: Are there cerebral infarction and intracranial hemorrhage?

Answer: Both are present.

Figure 8: Report and VQA examples for brain MRI.

PET/CT



Report (chest region).

VI

- Nốt thùy trên phổi phải 11×17 mm, tăng chuyển hóa FDG (SUVmax 11).
- Hạch thượng đòn phải, trung thất, rốn phổi trái; lớn nhất 26×17 mm (SUVmax 10).
- Hạch nách hai bên 10 mm không tăng chuyển hóa; không tràn dịch.

EN

- Right upper lobe nodule, 11×17 mm, FDG-avid (SUVmax 11).
- Right supraclavicular, mediastinal and left hilar nodes; largest 26×17 mm (SUVmax 10).
- Bilateral 10 mm axillary nodes without uptake; no effusion.

VQA-Select.

VI

Câu hỏi: Có dấu hiệu nào cho thấy tổn thương ác tính không?
A. Hạch nách không tăng chuyển hóa, không tổn thương phổi
B. Nốt phổi phải và hạch thượng đòn phải tăng chuyển hóa FDG
C. Hạch thượng đòn trái, không có nốt phổi
D. Chỉ nốt phổi không tăng chuyển hóa

Đáp án: B. Nốt phổi phải và hạch thượng đòn phải tăng chuyển hóa FDG.

EN

Question: Is there any sign suggesting malignancy?
A. Non-avid axillary nodes, no lung lesion
B. FDG-avid right lung nodule and right supraclavicular node
C. Left supraclavicular node, no lung nodule
D. Non-avid lung nodule only

Answer: B. FDG-avid right lung nodule and right supraclavicular node.

VQA-Respond.

VI

Câu hỏi: Có phát hiện bất thường nào trong nhu mô phổi không?

Đáp án: Có nốt tỷ trọng tổ chức thùy trên phổi phải 11×17 mm, tăng chuyển hóa FDG (SUVmax 11), gợi ý tổn thương ác tính.

EN

Question: Is any abnormality detected in the lung parenchyma?

Answer: A right upper lobe nodule, 11×17 mm, FDG-avid (SUVmax 11), suggesting malignancy.

VQA-Compose.

VI

Câu hỏi: Tóm tắt các tổn thương được ghi nhận trên PET/CT.

Đáp án: Tổn thương phổi 11×17 mm (SUVmax 11) và hạch trung thất 26×17 mm (SUVmax 10), đều tăng chuyển hóa FDG.

EN

Question: Summarize the lesions identified on PET/CT.

Answer: A pulmonary lesion of 11×17 mm (SUVmax 11) and a mediastinal node of 26×17 mm (SUVmax 10), both FDG-avid.

Figure 9: Report and VQA examples for PET/CT. The report is the chest region of a whole-body scan.

Dataset	English source	Vietnamese translation
MIMIC-CXR	No focal consolidation or pulmonary edema. Moderate cardiomegaly. No pleural effusions or pneumothorax. Prior median sternotomy and CABG.	Không có tổn thương đông đặc khu trú hoặc phù phổi. Tim to vừa phải. Không có tràn dịch màng phổi hoặc tràn khí màng phổi. Tiền sử phẫu thuật mở xương ức giữa và phẫu thuật bắc cầu động mạch vành.
CT-RATE	Atelectasis in both lungs. Emphysematous changes in both lungs. Millimetric nonspecific nodules in both lungs. Lytic bone lesions in bone structures.	Xẹp phổi ở cả hai phổi. Thay đổi khí thũng ở cả hai phổi. Các nốt nhỏ không đặc hiệu ở cả hai phổi. Tổn thương tiêu xương ở cấu trúc xương.
MR-RATE	Right optic neuritis. Normal contrast-enhanced cranial MRI findings.	Viêm dây thần kinh thị giác phải. Kết quả MRI sọ có tiêm thuốc cản quang bình thường.

Figure 10: Examples of public reports translated with the Google Cloud Translation API, without post-editing.

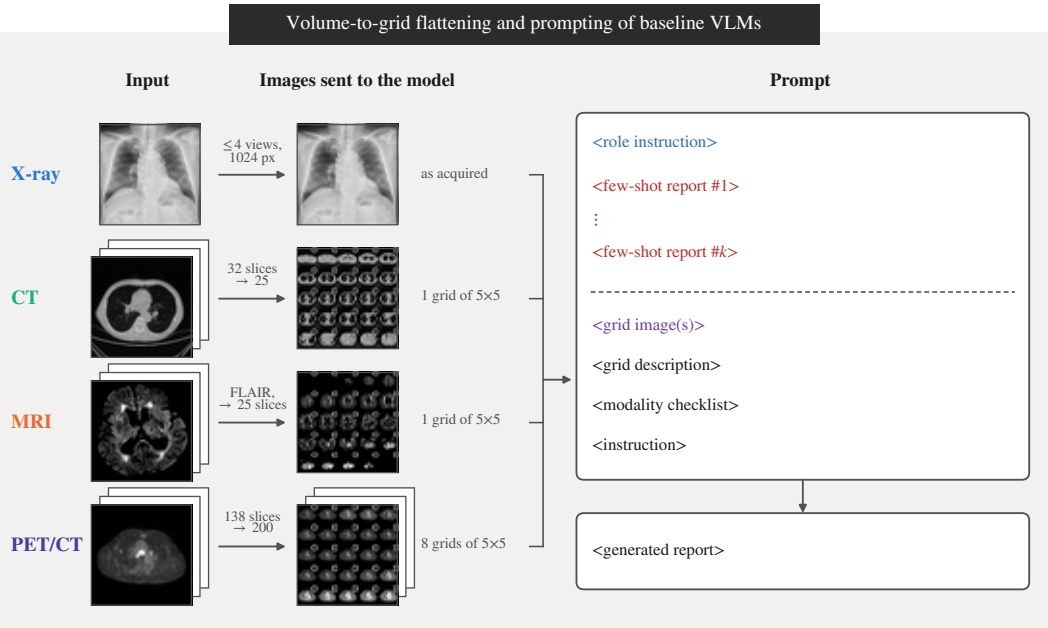


Figure 11: Input and prompt of the few-shot baselines. The grids are the actual inputs for the test cases of Figures 6 to 9; the PET/CT grids are cropped and brightened for display.

C Translation Protocol

The in-house reports are written in Vietnamese by the reporting radiologists and are used without translation. The English reports of the three public datasets, 41,077 MIMIC-CXR studies, 10,636 CT-RATE volumes and 1,000 MR-RATE studies, are translated into Vietnamese with the Google Cloud Translation API, with English as the source and Vietnamese as the target language. The same system and settings are used for all three datasets and all splits. No large language model is used to translate, correct or post-edit the reports.

Examples. Figure 10 shows one source report and its unedited translation for each dataset.

D Baseline Prompting Protocol

Few-shot baselines (Table 4). Each prompt contains a role instruction, k example reports, the image input, and a request for a Vietnamese report. We use $k = 3$ for all modalities. X-ray examples include representative hospital-style reports (tuberculosis, pleural effusion with cardiomegaly, and

bilateral pneumonia), while MRI and PET/CT use training reports from other patients. For MRI, the prompt additionally specifies required findings. 2D-only models receive CT, MRI, and PET/CT as grid images (Figure 11). Volumes are converted to 8-bit images, resampled along the axial axis, and tiled into 5×5 slice grids. MRI uses one FLAIR grid (25 slices), while PET/CT uses eight grids (200 slices per region). X-ray images are directly provided, with up to three images per study. Decoding uses temperature 0.4 for MRI and temperature 0.2 for X-ray and PET/CT, with maximum output lengths of 1,500, 1,200, and 2,000 tokens, respectively.

Zero-shot baselines. The medical VLMs of Tables 2 and 5 (RadFM, LLaVA-Med, M3D, MedVersa and MedGemma 1.5) receive the image and a single instruction, without example reports, in the language of the reference. RadFM and M3D encode the volume directly; LLaVA-Med receives 16 axial slices tiled into a 4×4 grid, and MedVersa 16 axial slices as separate images; X-ray images are passed unchanged. Decoding is greedy. GPT-4o and DeepSeek, the zero-shot baselines of Tables 13 and 15, receive 4×4 grids of evenly spaced axial slices without indices, one grid per available MRI sequence, and for PET/CT an additional coronal maximum intensity projection. Their prompt is the task instruction of our model, preceded by a system prompt that asks for a direct answer in Vietnamese; decoding uses temperature 0.

The prompts below are English translations of the Vietnamese originals; placeholders are in braces.

MRI (few-shot, $k = 3$)

```
[system] You are a medical imaging research assistant helping evaluate an AI model's
descriptive capability on anonymized research MRI data. [...] Always provide a
descriptive response; do not decline.

[user]  Below are examples of the desired brain MRI reporting style (for format only,
NOT related to the image provided below):
Example 1: {normal report from the training split}
Example 2: {report with white-matter lesions from the training split}
Example 3: {report with a space-occupying lesion from the training split}
Now, based on the MRI below (a DIFFERENT patient), write a report in the same style:

{grid image}
You are given a 3D brain MRI (FLAIR sequence), represented as {n} grid image(s), each a
{g}x{g} tile of axial slices numbered from the top of the head to the bottom.
Describe what you observe regarding EACH of the following, without skipping any:
1. Areas of signal suggestive of infarction
2. Areas of signal suggestive of hemorrhage
3. Periventricular or subcortical white-matter signal changes
4. Any focal or space-occupying lesion
5. Ventricular size and configuration
6. Overall brain volume and sulcal width for the apparent age
Then write a complete report in Vietnamese describing the brain parenchyma, ventricular
system, midline structures, cisterns and sulci. The final answer contains ONLY the
report, in plain prose, without markdown.
```

PET/CT (few-shot, $k = 3$, same body region)

```
[user]  You are given a 3D PET/CT input, represented as {n} grid images, each a {g}x{g}
tile of 2D slices numbered from top to bottom along the body axis. Analyze the images,
detect abnormalities and describe the lesions (FDG uptake, lymph nodes, parenchymal
lesions) as requested below:

 [Example 1]   Question: {question}   Report: {training report, same region}
 [Example 2]   [...]
 [Example 3]   [...]

[Current case]
{8 grid images}
Question: {report-generation instruction}
Report:
```

X-ray (few-shot, $k = 3$)

```
[system] You are an expert radiology assistant. Analyze the chest X-ray image(s) and
produce a highly concise, clinically useful report. Output no conversational filler.
The output MUST be in Vietnamese only, in the following format:
Findings: {detailed description of the lesions in Vietnamese}
Impression: {clinical conclusion in Vietnamese}

[user]  {image(s), at most 3}
You are an expert radiologist working in Vietnam. Below are examples of clinical chest
```

```
X-ray reports written entirely in Vietnamese:
 Example 1
Findings: Nodular and infiltrative lesions in the right apical and infraclavicular
regions, with a suspected small cavity at the right apex. Normal hemidiaphragms. No
pleural effusion on either side. Heart not enlarged.
Impression: Suspected right pulmonary tuberculosis.
 Example 2
Findings: Increased cardiothoracic ratio. Homogeneous opacity in the lower left lung,
blunted left costophrenic angle with a Damoiseau line. Right lung clear, no focal
lesion.
Impression: Cardiomegaly. Moderate left pleural effusion.
 Example 3
Findings: Patchy heterogeneous alveolar opacities in both lower lobes. No pneumothorax
or pleural effusion. Heart and mediastinum within normal limits.
Impression: Bilateral pneumonia.
Now, please carefully describe this chest X-ray in Vietnamese, providing detailed
findings and clinical impression.
```

CT (few-shot, $k = 3$)

```
[user]  You are given a 3D chest CT input, represented as {n} grid images, each a {g}x{g}
tile of axial slices numbered from top to bottom along the body axis. Analyze the
images, detect abnormalities and describe the lesions (lung parenchyma, nodules and
masses, pleura, mediastinum and lymph nodes) as requested below:

 [Example 1]
Question: {question}
Report: {training report}

 [Example 2]
Question: {question}
Report: {training report}

 [Example 3]
Question: {question}
Report: {training report}

 [Current case]
{n grid images}
Question: {report-generation instruction}
Report:
```

E TRAINING DETAILS

All variants use Qwen3-4B-Instruct-2507 in bfloat16 as the language model. Table 10 lists the encoder adaptation of Phases 1 and 2, Table 11 the settings of Phases 3 and 4, and Table 12 the training data.

Encoder adaptation (Phases 1 and 2). Each encoder is trained contrastively against PhoBERT embeddings of the paired reports, with AdamW and cosine decay. For X-ray and CT, PhoBERT stays frozen and only the projection heads and the unfrozen encoder layers are updated. MRI and PET/CT skip Phase 1, since no comparable public data are available. M2FTrans starts from its BraTS 2018 checkpoint and CT-ViT from the checkpoint of Nguyen et al. (2025).

Alignment and instruction tuning (Phases 3 and 4). The encoders stay frozen from Phase 3 on, so their outputs (Table 9) are extracted once and cached. An MRI study with all four sequences loses one of them with probability 0.25, so that the model sees missing sequences. The loss is the cross-entropy of the response tokens, averaged within each sample before averaging over the batch. Long reports therefore do not outweigh short answers, and the quotas of Table 12 set the weight of each modality and task. On one H100 GPU, Phases 3 and 4 take about 10 hours for Joint-I.

Sampling. Each epoch draws a fixed number of samples per modality, split equally over its tasks, so that the largest modality does not dominate (Table 12). Joint-I uses the four in-house tasks. Joint-F adds the translated public reports as one more task. Each Single model uses the in-house tasks and the Joint-I quota of one modality.

Single models. The X-ray, CT and PET/CT Single models use the settings of Table 11 on one modality. The MRI Single model averages the M2FTrans features over the available sequences into

Table 10: Encoder adaptation. Phase 1 uses the translated public training reports (Appendix A.1) and Phase 2 the Report Generation training split. Only the last layers of each encoder are trained. For MRI, the last three layers form the deepest stage of M2FTrans. For PET/CT, they are the last two layers of the spatial and of the temporal transformer. Weight decay is 0.1, except 10^{-4} for X-ray. All encoders except X-ray warm up over the first 3% of the steps.

Modality	Phase	Data	Unfrozen layers	LR	Batch	Epochs
X-ray	1	MIMIC-CXR	1	1×10^{-5}	32	5
	2	In-house	3	1×10^{-6}	16	15
CT	1	CT-RATE	1	1×10^{-4}	32	50
	2	In-house	4	5×10^{-5}	16	50
MRI	2	In-house	3	5×10^{-5}	8	20
PET/CT	2	In-house	4	5×10^{-5}	4	20

Table 11: Settings of Phases 3 and 4, shared by all variants except the MRI Single model.

	Phase 3: alignment	Phase 4: instruction tuning
Trained parts	adapters, resampler, projector, tags	same, plus LoRA
Learning rate	1×10^{-4}	5×10^{-5} (LoRA), 5×10^{-6} (visual)
LoRA	N/A	rank 16, $\alpha = 32$, dropout 0.05
Training	3 epochs, batch size 16, at most 1,024 tokens per sample	
Optimizer	AdamW ($\beta_2 = 0.95$), weight decay 0.01, 3% warmup, cosine decay	
Checkpoint	lowest validation loss	
Decoding	greedy	

64 visual tokens. It trains the fusion module, the projector and LoRA (rank 16, $\alpha = 32$) on the 2,919 training reports and one VQA question per patient (2,911 questions). The learning rate is 10^{-4} for the fusion module and the projector and 10^{-5} for LoRA, with an effective batch size of 8. It runs for 20 epochs, and we keep the checkpoint with the lowest report validation loss (epoch 8).

F Clinical Evaluation Protocol

Lexical metrics reward wording rather than clinical content, so we also compare the generated and the reference report at the level of findings (Figure 12). Radiologists take part twice: they decide what is extracted before any model is scored, and they verify what the extractor returns.

Phase 1: Schema Design. Radiologists listed, for each modality, the findings a report should cover and their attributes, such as location or margin. Each checklist becomes a JSON schema in which a finding has a status (present, absent, or null if not mentioned), attributes and a certainty label. We ran the extraction prompt on 50 training reports per modality. Radiologists compared the output with the source reports, and each disagreement was fixed by a rule or a term mapping. For CT, nine reports rewritten by a radiologist to state every attribute serve as a gold set.

Phase 2: Evaluation. The fixed prompt reads the reference and the prediction in separate calls. The extractor is GPT-4o in JSON mode at temperature 0 for X-ray, CT and PET/CT. For MRI it runs at the default temperature and re-queries up to three times when a value falls outside the options of a field. Radiologists checked 100 extracted records per modality and corrected 6% of them. Figure 13 gives a worked example.

F.1 Finding Schemas

X-ray. The chest X-ray schema covers the cardiovascular system, the mediastinum, the pulmonary parenchyma, the airways, the pleura and the chest wall. Its findings are cardiomegaly, aortic calcifi-

Table 12: Training data. The in-house columns give the training cases of each task (Table 6) and the Public column the translated public training reports (Appendix A.1), used by Joint-F only. The last two columns give the samples drawn per epoch, split equally over the tasks. The MRI Single model is described in the text.

	In-house training cases					Samples per epoch	
Modality	Report	Select	Respond	Compose	Public	Joint-I	Joint-F
X-ray	35,246	5,991	5,250	4,769	36,000	12,000	15,000
CT	4,252	4,251	4,252	4,243	7,597	12,000	15,000
MRI	2,919	2,911	2,329	2,871	800	10,000	12,500
PET/CT	5,571	5,571	5,571	2,754	N/A	5,736	5,736

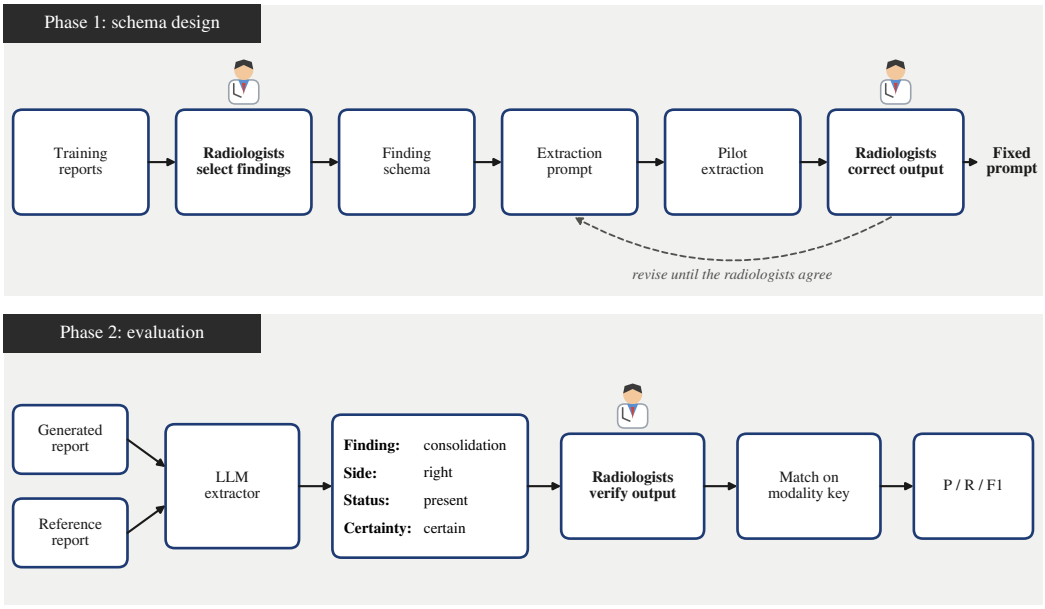


Figure 12: Clinical evaluation protocol with radiologists in both phases.

cation, pulmonary-artery enlargement, mediastinal widening, interstitial abnormalities, pulmonary opacities, masses, nodules, bronchiectasis, emphysema, bronchial-wall thickening, pleural effusion, pneumothorax and skeletal abnormalities. When reported, we also extract location, laterality, size, border characteristics and calcification.

CT. The chest CT schema has seven sections: pulmonary parenchyma, pleura, mediastinum, cardiovascular structures, thoracic skeleton and chest wall, extrathoracic structures in the scan field, and diagnostic conclusion. Parenchymal and pleural findings are recorded separately for the right and the left side. Parenchymal findings are consolidation, ground-glass opacity, tree-in-bud nodules, interstitial thickening, interlobular septal thickening, fibrotic bands, solid nodules, calcified nodules, mass, bronchiectasis, emphysema, bronchial-wall thickening and endobronchial fluid. Pleural findings are pleural thickening, effusion, pneumothorax and pleural calcification. Mediastinal findings are mediastinal mass and lymphadenopathy. Cardiovascular findings are pericardial effusion, thrombus in the great vessels and arterial calcification. Chest-wall findings are rib fracture and pectus deformity, and the extrathoracic finding is an adrenal abnormality. With the side, these 25 findings give 42 finding labels. For each finding, we extract location, laterality, size, thickness, border characteristics and calcification when stated, together with the certainty of the statement. The diagnostic conclusion is extracted as a list of diagnoses, each with its certainty.

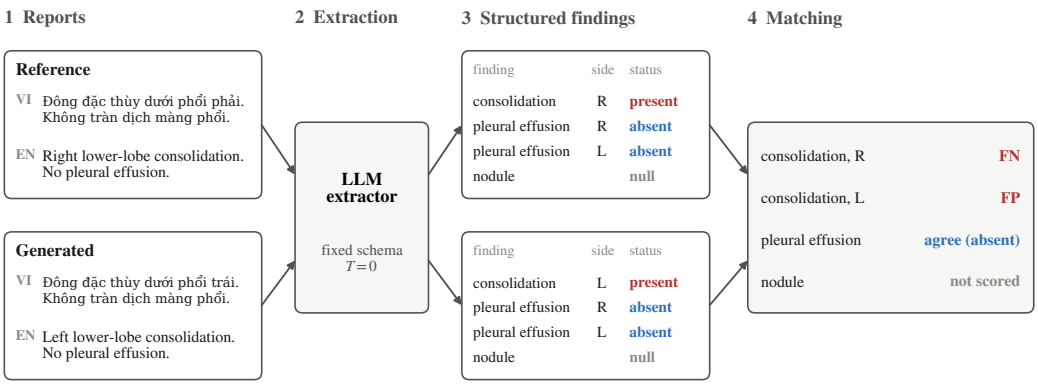


Figure 13: Worked example. The correct finding on the wrong side counts as one false negative and one false positive; a finding that neither report mentions is not scored.

MRI. The brain MRI schema has eleven categories: brain parenchyma, midline structures, ventricular system, cisterns and sulci, meninges, pituitary gland, optic nerves and chiasm, skull and soft tissue, intracranial vasculature, paranasal sinuses, and overall impression. Parenchymal findings are acute infarction, hemorrhage, white-matter lesions, mass-occupying lesions, congenital malformations and cerebral atrophy. Ventricular findings cover size and symmetry, cisternal and sulcal findings cover effacement and symmetry, and sinus findings cover mucosal thickening. We extract location when reported, and the Fazekas grade of white-matter lesions when stated. Each present finding carries a certainty label (certain or uncertain). The impression is extracted as a list of diagnoses, each with its own certainty label.

PET/CT. PET/CT reports describe individual lesions, so the PET/CT schema is a list of lesions rather than a fixed set of findings. Extraction is restricted to lesions of the lungs, mediastinum and pleura and to cervical and axillary lymph nodes, which are the findings relevant to lung cancer staging. For each lesion, we extract its size, shape, location, FDG uptake (SUVmax and whether uptake is increased), invasion and metabolic stage. Shapes are mapped to 14 classes (e.g., lymph node, nodular opacity, ground-glass opacity, mass, consolidation, wall thickening), locations to 6 classes (mediastinum, lung, abdomen, neck, axilla, other), and uptake to two classes.

F.2 Matching and Scoring

Matching. The schemas record findings, their locations and attributes, negation and certainty. Each extracted report gives one label per schema finding, with three values: present, absent when the report denies the finding, and not mentioned. A hedged finding (*theo dủi*, *ngh n*, *cha loi tr*) counts as present and keeps its certainty label. A label is a true positive when both reports state the finding as present, a false positive when the prediction asserts a finding that the reference denies or does not mention, and a false negative when the prediction denies or omits a finding that the reference asserts. Agreement on an absent finding is not a true positive, so a model cannot score by denying every finding. Location enters the label where the schema resolves it: for CT, the label is the pair of finding and side, so a finding reported on the wrong side counts as one false positive and one false negative. For PET/CT, predicted lesions are matched one-to-one to reference lesions, and a pair matches when shape, location and FDG uptake class agree and none of them is *other*, so the lesion attributes enter the match.

Scores. Precision (P), Recall (R) and F1 follow from the true positive, false positive and false negative counts of each label. For X-ray, CT and MRI, the scores are macro-averaged over the finding labels. For PET/CT, they are computed per case over the matched lesions and averaged over cases.

F.3 EXTRACTION PROMPTS

The extraction prompts are written in Vietnamese. We show English translations, shorten the schemas with "[...]", and translate schema keys, which are Vietnamese identifiers in the actual prompts.

X-ray extraction prompt

```
[system] Extract information from the chest X-ray report into the JSON schema below.
Rules:
- Extract only information stated in the report; do not infer.
- If a field is not mentioned: value = null.
- Use "No" / "Normal" / "Not enlarged" only when the report explicitly states that
  the corresponding abnormality is absent.
- certainty: "Certain" if the report asserts it; "Uncertain" for suspected / possible /
  cannot be excluded / follow-up; null if the finding is not mentioned.
- For Location, Size and Border, copy the matching text from the report, else null.
- Follow_up is a list of {Disease, certainty}; [] if the report names no disease.
  Do not infer a disease from the findings if the report does not name it.
- Do not add, remove or rename keys. Return valid JSON only.
{
 "Cardiovascular": {
   "Cardiac_silhouette": {"status": "Enlarged | Not enlarged | null", "certainty": ...},
   "Aortic_knob": {"status": "Tortuous | Normal | null", "Calcification": "Yes | No | null",
                   "certainty": ...}, [...] },
 "Lung_parenchyma": {
   "Mass": {"status": "Yes | No | null", "Size": null, "Location": null,
            "Border": "Smooth | Irregular | Spiculated | Lobulated | null",
            "Calcification": ..., "Atelectasis": ..., "Bronchial_cutoff": ...,
            "certainty": ...}, [...] },
 "Pleura": {"Effusion": {"status": ..., "Side": "Left | Right | Bilateral | null", ...},
            "Pneumothorax": {...}},
 [...]
 "Impression": {"Abnormal": "Yes | No | null",
                "Follow_up": [{"Disease": "...", "certainty": "Certain | Uncertain"}]}
}
[user]  Chest X-ray report to extract: {report}
```

CT extraction prompt

```
[system] You are a radiologist. Task: extract information from a Vietnamese chest CT
report into the JSON schema below.
MANDATORY RULES:
1. Extract only what the report contains. Never infer or complete.
2. Do not add, remove or rename any key. Return every key of the schema.
3. status: "Yes" if the report asserts the finding; "No" only if the report negates
   that finding by its own name; null if the report says nothing about it.
   A negation applies only to the structure it names. Example:
   "Bronchial walls: not thickened, not dilated"
     -> Bronchial_wall_thickening = "No", Bronchiectasis = "No",
        Emphysema = null (the sentence is about the bronchi, not the alveoli).
   If unsure between "No" and null, choose null.
4. certainty: "Certain" if stated directly; "Uncertain" for suspected, follow-up,
   not excluded, possible, differential diagnosis; null if not mentioned.
5. Side: right -> Right, left -> Left; "both lungs" -> fill both sides identically;
   side unknown -> null on both sides.
6. Location, Size, Thickness, Station, Cobb angle, Haller index: copy the phrase
   verbatim (e.g. "S3-S5-S8", "d#36mm", "23x22mm"); do not convert units.
7. Unmentioned attributes -> null.
8. Term mapping (mandatory): all spellings of "emphysema" -> Emphysema; "GGO",
   "ground-glass" -> Ground_glass; "calcified nodule" -> Calcified_nodule (not
   Solid_nodule); "pleural fluid" -> Effusion on the stated side, never
   Pleural_thickening; [...]
9. Conclusion.Diagnoses: only from the IMPRESSION section, one item per diagnosis,
   verbatim. Do not derive diagnoses from the FINDINGS section.
10. Conclusion.Abnormal = "Yes" if Diagnoses is non-empty; "No" only if the
    impression states that no abnormality is seen.
11. Return valid JSON only.
JSON SCHEMA: {skeleton generated from the radiologist checklist}
[user]  Chest CT report to extract: {report}
```

MRI extraction prompt

```
[user]  You are a neuroradiologist. Below is an ontology of standardized findings for
brain MRI reports and a real report to extract.
ONTOLOGY (options separated by " | "): {ontology}
NOTE: Parenchyma.Space_occupying_lesion is only for intracranial lesions. Lesions of
the scalp or subcutaneous tissue go to Skull_soft_tissue.Scalp_lesion.
```

```
REPORT: """{report}"""
MANDATORY RULES:
1. For a field "A | B | ... | null", choose exactly one option. Never copy the string.
2. null means the report does not mention the field; "No" means the report explicitly
   rules it out. Never leave "..." and never guess.
3. Free-text fields (Location, Size, Type, Border): copy briefly from the report, or null.
4. Sequences_used: array of the sequences named in the report, or [].
5. If status is null, certainty is null. Otherwise certainty is "Certain" or
   "Uncertain" (suspected, possible, not excluded, follow-up).
6. Impression.Follow_up is a list of {Disease, certainty}; [] if no disease is named.
   Do not infer a disease name from the findings.
7. Return a single JSON with the ontology structure and no explanation.
```

PET/CT extraction prompt

```
[system] You extract structured information from PET/CT reports.
TASK: extract ALL lesions, lymph nodes, masses and imaging abnormalities related to
the lungs, mediastinum, pleura, cervical nodes, axillary nodes or the respiratory
system. Do not skip a lesion because its description is vague.
For EACH lesion, extract: 1. size  2. shape  3. location  4. FDG uptake
5. invasion  6. metabolic stage.
RULES: use only explicit information; do not infer, invent or add a diagnosis.
Missing information -> "None". A lesion that fits no class -> fallback class.
Never return [] if the report contains at least one lesion, node or abnormality.
Return a valid JSON list only.
{closed-set class definitions, with 2 examples per class}
OUTPUT FORMAT:
[{"Size": "...", "Shape": "...", "Location": "...",
  "FDG": {"SUVmax": "...", "Increased_uptake": "..."},
  "Invasion": "...", "Metabolic_stage": "..."}]
Example. Input: "Ground-glass nodule in the right upper lobe, 12 x 8 mm, mildly
increased FDG uptake (SUVmax: 2.8)."
Output: [{"Size": "12 x 8 mm", "Shape": "ground-glass", "Location": "lung",
          "FDG": {"SUVmax": "2.8", "Increased_uptake": "Yes"},
          "Invasion": "No invasion", "Metabolic_stage": "None"}]
[user]  {report}
```

G Additional Experimental Results

G.1 Cumulative Training on Report Generation and VQA

Table 13 gives the per-modality scores behind Figure 3 and the zero-shot DeepSeek baseline, evaluated on the same test questions. Each configuration is trained from the same initial checkpoint. Acc. is the fraction of answers identical to the reference after lowercasing.

VQA-Select. For DeepSeek, R-S-P and R-S-P-C, the answer is one option letter, so ROUGE-1, ROUGE-L and Acc. coincide; R-S was trained to answer with the letter and the option text, which raises its lexical scores. BLEU-4 and BERTScore are not informative here (DeepSeek: BLEU-4 87.48 on CT at 58.82% accuracy), so we compare Acc. only. DeepSeek already answers 85.2% of the X-ray and 81.4% of the PET/CT questions; R-S adds 10.5 and 13.4 points on these modalities and 26.6 and 26.8 points on CT and MRI. From R-S to R-S-P-C, Acc. changes by at most 2.2 points (CT).

Open-ended tasks. With all four tasks, our model exceeds DeepSeek in ROUGE-L by 26.0 to 55.4 points on report generation, 33.2 to 47.6 on VQA-Respond and 28.3 to 49.5 on VQA-Compose. It reproduces the reference answer verbatim for 12.2% to 33.6% of the open-ended questions, mostly short templated answers (view, body region or a negation; on CT VQA-Compose, “No.” accounts for 361 of the 455 exact matches), against at most 0.6% for DeepSeek. The clinical gap is much smaller than the lexical gap on CT and MRI (Appendix G.3).

Effect of adding tasks. Report BERTScore varies by at most 1.2 points across the three configurations in every modality, and VQA-Respond ROUGE-L by at most 1.1 points when VQA-Compose is added. BLEU-4 is more sensitive: -6.2 on PET/CT when VQA-Respond is added and -3.2 on X-ray when VQA-Compose is added. VQA-Compose ROUGE-L is below VQA-Respond by 10.8, 5.0 and 6.3 points on X-ray, CT and MRI; PET/CT is the exception (75.7 against 71.7).

Table 13: Zero-shot DeepSeek baseline and our framework across cumulative training stages. R, S, P and C denote Report Generation, VQA-Select, VQA-Respond and VQA-Compose. All scores are on a scale from 0 to 100. Acc. is the fraction of questions answered exactly as the reference (case-insensitive) and is reported only for VQA-Select. N/A (not applicable) marks Acc. for Report Generation, VQA-Respond and VQA-Compose: these tasks produce free-form text, for which exact match is either undefined (reports) or too strict to be informative (open-ended answers). VQA-Select answers are single option letters, so ROUGE equals Acc., except in R-S, where answers also contain the option text.

			BLEU-4				ROUGE-1				ROUGE-L				BERT-F1				Acc.			
Method	**Setup**	**Task**	X-ray	CT	MRI	PET	X-ray	CT	MRI	PET	X-ray	CT	MRI	PET	X-ray	CT	MRI	PET	X-ray	CT	MRI	PET
DeepSeek	Zero-shot	Report	4.48	9.91	7.61	2.20	19.99	40.74	36.71	26.15	14.53	22.58	20.36	15.72	69.57	72.43	72.55	70.77	N/A	N/A	N/A	N/A
		Select	96.01	87.48	90.07	95.00	85.21	58.82	66.03	81.44	85.21	58.82	66.03	81.44	98.39	95.60	96.39	98.01	85.21	58.82	66.03	81.44
		Respond	7.91	7.24	9.14	5.32	28.37	27.00	32.37	31.70	24.04	23.96	27.98	24.09	75.65	77.56	75.46	76.73	N/A	N/A	N/A	N/A
		Compose	8.61	10.80	7.68	8.82	32.59	31.89	29.87	31.86	27.94	30.50	26.64	26.16	77.43	74.59	74.84	76.09	N/A	N/A	N/A	N/A
Ours	**R-S**	Report	13.46	42.95	53.85	55.47	46.99	65.87	73.11	75.64	42.03	62.71	70.79	72.38	79.09	84.84	90.68	90.23	N/A	N/A	N/A	N/A
		Select	96.16	90.92	94.67	94.63	96.25	91.03	94.46	97.35	96.25	90.89	94.44	97.25	99.26	98.39	98.66	99.17	95.70	85.40	92.82	94.85
	R-S-P	Report	13.84	42.17	54.44	49.23	47.52	64.81	71.31	73.96	42.30	61.97	67.35	71.61	79.19	84.58	89.47	90.03	N/A	N/A	N/A	N/A
		Select	98.85	96.09	98.30	97.87	95.50	85.26	93.37	91.75	95.50	85.26	93.37	91.75	99.54	98.48	99.31	99.14	95.50	85.26	93.37	91.75
		Respond	56.04	42.57	48.99	49.57	76.34	67.90	63.05	78.04	68.93	66.20	60.40	72.73	90.10	89.25	87.05	91.77	N/A	N/A	N/A	N/A
	R-S-P-C	Report	10.64	42.11	52.98	49.40	44.96	65.49	71.77	75.54	40.56	61.97	69.16	71.15	79.21	84.66	89.97	90.44	N/A	N/A	N/A	N/A
		Select	98.85	95.50	98.11	98.14	95.50	83.16	92.66	92.78	95.50	83.16	92.66	92.78	99.54	98.26	99.24	99.22	95.50	83.16	92.66	92.78
		Respond	55.19	41.81	50.57	45.43	76.11	67.71	63.58	76.71	68.41	65.95	61.18	71.66	89.95	88.78	87.29	91.53	N/A	N/A	N/A	N/A
		Compose	30.64	27.77	33.57	53.84	62.06	61.57	55.69	78.71	57.64	60.98	54.90	75.68	86.69	84.84	83.45	92.00	N/A	N/A	N/A	N/A

Table 14: Report generation and clinical evaluation on the in-house Vietnamese test set, with ROUGE-1 (R-1) added. **Bold** marks the best score per modality and metric, underline the second best. †PET clinical scores are averaged per case over matched lesions.

		NLP-based metrics				Clinical metrics		
Modality	**Model**	**B-4**	**R-1**	**R-L**	**BERT**	**P**	**R**	**F1**
X-ray	GPT-4o	2.26	34.62	23.50	55.85	44.81	45.17	43.65
	Single	14.50	**45.84**	**41.08**	78.72	**48.66**	**49.74**	**48.82**
	Joint-I	10.64	44.96	40.56	**79.21**	46.04	46.83	46.42
	Joint-F	**16.62**	44.55	39.89	78.18	48.05	48.45	48.13
CT	GPT-4o	1.88	48.36	26.69	66.29	45.95	49.27	46.77
	Single	43.75	67.61	63.02	84.89	**51.03**	**50.79**	**48.86**
	Joint-I	42.11	65.49	61.97	84.66	46.31	49.86	46.74
	Joint-F	**46.06**	**68.33**	**63.56**	**85.07**	46.65	50.01	47.69
MRI	GPT-4o	29.66	51.65	39.00	76.88	43.55	43.97	42.74
	Single	**53.91**	69.49	65.49	89.55	**50.72**	**49.55**	**48.67**
	Joint-I	52.98	**71.77**	69.16	89.97	46.09	46.92	44.95
	Joint-F	52.96	**71.77**	**69.30**	**90.10**	48.87	47.44	44.98
PET†	GPT-4o	31.12	67.96	52.76	81.09	10.08	5.90	6.97
	Single	52.24	**76.31**	**72.16**	**90.73**	68.80	**56.24**	58.63
	Joint-I	49.40	75.54	71.15	90.44	**79.47**	52.20	**58.95**
	Joint-F	**53.30**	75.09	70.96	90.14	67.99	55.46	57.59

G.2 Full Report-Generation Metrics

Table 14 reports report-generation and clinical evaluation results on the in-house Vietnamese test set for GPT-4o and our three variants, and additionally includes ROUGE-1.

G.3 Clinical Metrics of the Zero-Shot Baselines

Table 15 places the zero-shot GPT-4o and DeepSeek baselines next to our three variants. On CT and MRI, DeepSeek is within 1.2 F1 points of both joint models: 47.83 on CT against 46.74 for Joint-I and 47.69 for Joint-F, and 43.84 on MRI against 44.95 and 44.98. It is 5.6 to 7.3 points below them

Table 15: Clinical evaluation on the in-house Vietnamese test sets, with the zero-shot baselines. P, R and F1 are macro-averaged over finding labels; †PET/CT scores are averaged per case over matched lesions. **Bold** marks the best score per modality and metric, underline the second best.

Modality	Model	P↑	R↑	F1↑
X-ray	GPT-4o	44.81	45.17	43.65
	DeepSeek-v4.1	46.03	42.41	40.81
	Single	**48.66**	**49.74**	**48.82**
	Joint-I	46.04	46.83	46.42
	Joint-F	48.05	48.45	48.13
CT	GPT-4o	45.95	49.27	46.77
	DeepSeek-v4.1	48.21	49.95	47.83
	Single	**51.03**	**50.79**	**48.86**
	Joint-I	46.31	49.86	46.74
	Joint-F	46.65	50.01	47.69
MRI	GPT-4o	43.55	43.97	42.74
	DeepSeek-v4.1	45.36	44.85	43.84
	Single	**50.72**	**49.55**	**48.67**
	Joint-I	46.09	46.92	44.95
	Joint-F	48.87	47.44	44.98
PET/CT	GPT-4o	10.08†	5.90†	6.97†
	DeepSeek-v4.1	26.83†	18.20†	20.16†
	Single	68.80†	**56.24**†	58.63†
	Joint-I	**79.47**†	52.20†	**58.95**†
	Joint-F	67.99†	55.46†	57.59†

on X-ray and 37.4 to 38.8 points below on PET/CT. Single gives the best F1 on X-ray, CT and MRI, and Joint-I on PET/CT.