FLARE: SELECTIVELY ROUTING LESION EVIDENCE FOR 3D CT RADIOLOGY REPORTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent multimodal large language models have enabled radiology report generation from full 3D CT volumes, but clinically important focal lesions can remain underrepresented within global volumetric context. A natural remedy is to augment the global input with lesion-focused visual tokens. However, our experiments show that this naive global-plus-local approach does not yield consistent improvements, suggesting that simply providing local evidence does not ensure its effective use. We introduce FLARE, a post-training framework that explicitly controls when local lesion evidence influences report generation. FLARE pairs a frozen global report generator with a crop expert that extracts lesion-specific evidence from local crops. A router decides whether this evidence should influence the global prediction at each decoding step of full-report generation. On the ReXGroundingCT held-out test set, FLARE shows overall improvements across clinically oriented metrics over both full-report supervised fine-tuning and naive local augmentation, while maintaining comparable textual similarity to the global-only baseline. These results support selective lesion-evidence routing as an effective approach for strengthening focal finding descriptions within comprehensive volumetric reports.

1 INTRODUCTION

Radiology report generation (RRG) has traditionally focused on 2D imaging (Jing et al., 2018; Chen et al., 2020; Liu et al., 2021), while recent vision-language models (VLMs) have increasingly been extended to full 3D computed tomography (CT) volumes (Hamamci et al., 2024; Yang et al., 2024; Chen et al., 2024). Volumetric modeling is particularly important for CT, where abnormalities must be interpreted across slices together with their extent, anatomical context, and relationships to surrounding structures (Hamamci et al., 2024; Hosseini et al., 2025; Hatamizadeh et al., 2022; Xing et al., 2024; Ye et al., 2026). A clinically useful CT report must therefore reconcile two complementary requirements: it should capture the *global* state of the examination while accurately describing *local* abnormalities that may occupy only a tiny fraction of the full volume.

As summarized in Fig. 1, volumetric radiology report generation has expanded from **(A)** global full-volume reporting to **(B)** region-aware approaches that incorporate anatomical-region information and **(C)** lesion-grounding approaches that focus on individual localized lesions. Our target setting further extends this direction by **(D)** incorporating fine-grained lesion-specific evidence into comprehensive full-report generation, requiring the model to preserve global context while accurately incorporating focal abnormalities. However, processing the entire CT can underrepresent small focal lesions within the full-volume input, with their signals further weakened by visual-token compression (Ye et al., 2026). As a result, a model may generate a plausible report that omits or incompletely describes clinically important focal findings.

A straightforward remedy is to augment the global representation with region- or lesion-focused visual features (Wang et al., 2023; Chen et al., 2025a; Sharma et al., 2025). For example, a lesion crop can be encoded and provided alongside the full-volume representation, giving the model a more concentrated view of the lesion. However, our experiments show that such naive global-plus-local augmentation does not yield consistent improvements, suggesting that simply providing lesion-focused information is insufficient for its effective use during report generation.

As illustrated in Fig. 2, additional lesion information can help recover the target finding, yet may still be incorporated incorrectly within the full report. This suggests that the bottleneck is not only

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.

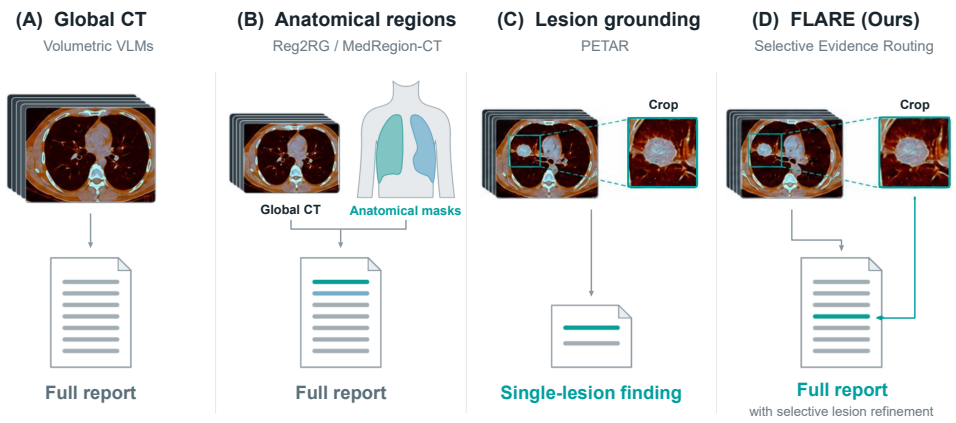


Figure 1: **Comparison of volumetric radiology reporting paradigms. (A) Global CT:** volumetric VLMs generate a full report directly from the global CT volume (Hamamci et al., 2024). **(B) Anatomical regions:** region-aware approaches use anatomical masks or regional information to guide full-report generation (Chen et al., 2025a; Kyung et al., 2025). **(C) Lesion grounding:** lesion-focused approaches use a localized lesion crop to generate a finding for an individual lesion rather than a complete report (Maqbool et al., 2026). **(D) FLARE:** our approach retains full-report generation from the global CT while selectively incorporating lesion-specific evidence.

how the lesion is represented, but also when its information should influence full-report generation. Conventional token augmentation leaves both decisions implicit within the same model, requiring lesion features to compete with the global context throughout the generation of the complete report. We instead separate these roles: the global model remains responsible for comprehensive volumetric reporting, while lesion-specific information is extracted independently and introduced only when it is relevant to the current full-report prediction. This separation allows local evidence to support focal findings while preserving the broader full-report generation process.

To this end, we introduce **FLARE**[1], a framework for **selectively routing lesion evidence for comprehensive volumetric radiology reporting**. We first adapt MedGemma 1.5 (Sellergren et al., 2026) to full-report generation from volumetric CT as a global report generator that captures examination-level context. FLARE pairs this global model with a crop expert trained using localized findings from ReXGroundingCT (Baharoon et al., 2026). Rather than representing the lesion as another set of visual tokens to be mixed with the global context, the crop expert measures how the lesion crop changes the model's next-token prediction relative to a null visual input. This difference provides an explicit lesion-evidence signal for each decoding step. A router then decides whether this evidence should influence the global prediction during full-report generation. Together, the crop expert and router separate what the lesion contributes from when that information is used, allowing focal findings to be strengthened without broadly perturbing the full report.

Our experiments on the ReXGroundingCT held-out test set show that FLARE outperforms the evaluated volumetric VLM baselines on clinically oriented report-generation metrics while maintaining comparable textual similarity. In contrast, directly augmenting the global model with lesion crops does not provide consistent gains. LLM-based clinical evaluation further supports the improved clinical faithfulness of FLARE.

Our contributions are threefold:

- We show that naively augmenting a volumetric report generator with lesion-focused visual tokens does not yield consistent improvements, highlighting that simply representing focal

[1]**F**ocal **L**esion-**A**ware **R**outing of **E**vidence

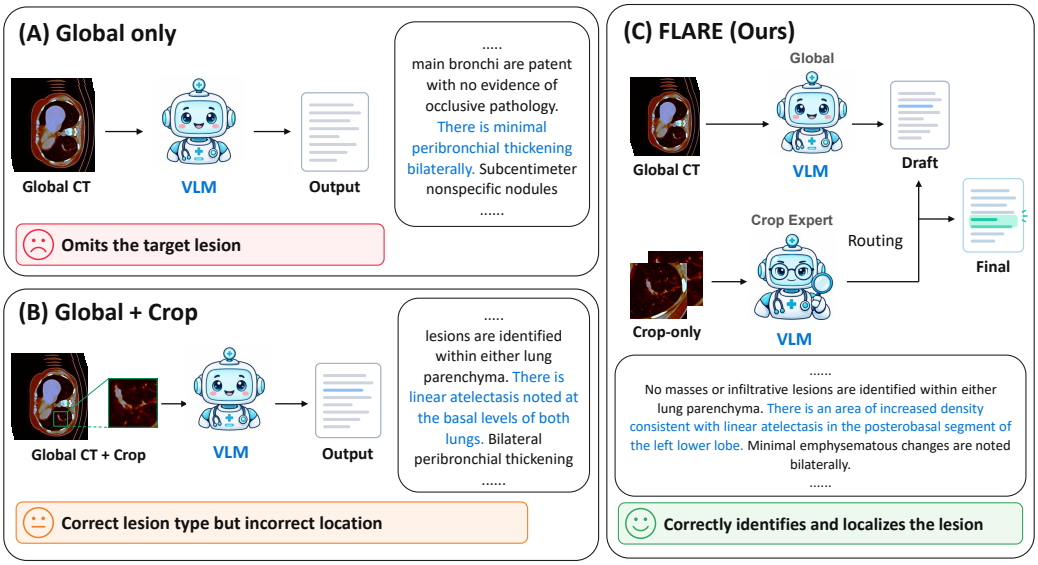


Figure 2: **Qualitative case study of focal-lesion reporting in volumetric radiology report generation.** **(A)** Global-only generation misses the target lesion. **(B)** Global + Crop recovers the lesion type but assigns it to an incorrect anatomical location. **(C)** FLARE selectively incorporates lesion-specific evidence, recovering both the lesion type and its anatomical location.

information more explicitly is insufficient when its use remains implicit within a large global context.

- Motivated by this limitation, we introduce FLARE, which separates lesion-evidence extraction from evidence routing. A crop expert extracts lesion-specific evidence, while a router determines when that evidence should influence the global prediction during comprehensive full-report generation.

- We validate FLARE using report-level clinical metrics and LLM-based evaluation. FLARE improves over volumetric VLM and SFT baselines on the annotated test set, while also showing clinical gains with automatically localized regions of interest (ROIs) on scans without manual annotations.

2 METHOD

Problem setup. Given a chest CT volume, our goal is to generate a complete radiology FINDINGS report while selectively incorporating localized lesion evidence. Each volume is represented by 85 axial slices uniformly sampled throughout the CT, together with one or more lesion ROIs represented as local image crops. During training, the ROIs are obtained from ReXGroundingCT annotations, and their associated localized findings are aligned with the corresponding spans in the CT-RATE report; details of this data preparation are provided in Appendix A.1. We use ground-truth ROIs for the controlled evaluation and additionally evaluate an automatic setting in which the ROIs are predicted by a learned localizer.

FLARE selectively incorporates lesion-specific evidence into full-report generation without directly mixing local crop tokens with the full-volume representation. We start from MedGemma 1.5 fine-tuned for volumetric full-report generation and keep this global VLM frozen during FLARE training. As shown in Fig. 3, FLARE consists of two stages: a crop expert learns to derive finding-specific predictive evidence from a localized lesion crop, and a router determines when this evidence should modify the global next-token prediction.

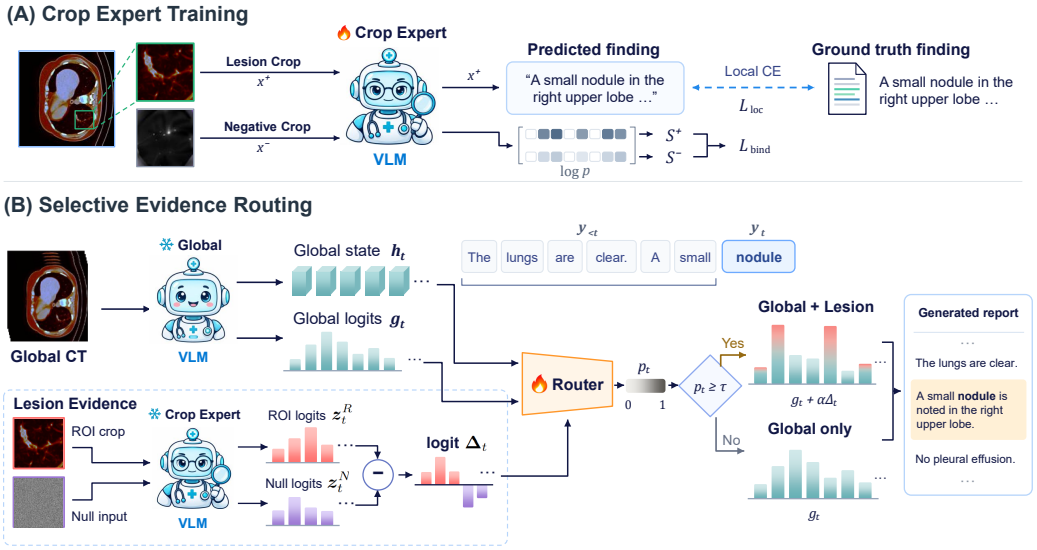


Figure 3: **Overview of FLARE. (A)** The crop expert is trained with localized finding supervision, combining a local generation objective with a binding objective that encourages stronger support from the GT lesion crop than from an unrelated crop. **(B)** During full-report generation, lesion evidence is computed as the logit difference between the ROI crop and a null input. A router determines at each decoding step whether this evidence should modify the frozen global prediction.

2.1 Crop Expert Training

The crop expert is designed to capture visual evidence that is specific to a localized lesion. As illustrated in Fig. 3(A), it receives a lesion crop together with its corresponding localized ground-truth finding. The crop expert reuses the MedGemma1.5-SFT vision encoder, multimodal projector, and language decoder, while introducing a small trainable adaptor on the crop pathway. By keeping the pretrained components frozen, we preserve the model's existing CT understanding and generation capabilities, while the adaptor learns more discriminative lesion representations that encourage more effective use of local visual content.

For the GT lesion crop x^+, the crop expert is first optimized to generate the corresponding localized finding using a standard language-modeling loss, denoted by $\mathcal{L}_{\text{loc}}$. We complement this generative objective with a relative lesion–finding binding objective. Specifically, we provide an unrelated negative crop x^- and encourage clinically informative finding tokens to receive stronger support from the corresponding lesion crop than from unrelated visual content.

Let S^+ and S^- denote the average log-probability assigned to clinically informative finding tokens under the GT lesion crop and the unrelated negative crop, respectively. We use a binding loss

$$\mathcal{L}_{\text{bind}} = \left[m - (S^+ - S^-)\right]_+^2, \tag{1}$$

which encourages stronger correspondence between the localized finding and its associated lesion crop relative to unrelated local content. The crop expert is trained with

$$\mathcal{L} = \mathcal{L}_{\text{loc}} + \lambda_{\text{bind}}\mathcal{L}_{\text{bind}}. \tag{2}$$

After training, the crop expert is frozen and used only to extract lesion-specific evidence.

2.2 Selective Evidence Routing

Figure 3(B) illustrates how local evidence is selectively incorporated during full-report generation. At each decoding step, the frozen global VLM processes the full CT volume and the current report prefix, producing the global state $\mathbf{h}_t$ and next-token logits $\mathbf{g}_t$.

In parallel, an ROI crop and a null input are passed through the same frozen crop expert under the same report prefix, producing ROI logits $\boldsymbol{z}_t^R$ and null logits $\boldsymbol{z}_t^N$, respectively. We define their difference as the crop-derived evidence:

$$\boldsymbol{\Delta}_t = \boldsymbol{z}_t^R - \boldsymbol{z}_t^N. \tag{3}$$

Thus, $\boldsymbol{\Delta}_t$ represents the change in next-token prediction associated with the given ROI crop.

The router then uses the crop-derived evidence $\boldsymbol{\Delta}_t$ together with the global state $\mathbf{h}_t$ to predict a routing probability p_t. During training, we use two types of ROI Crop: one corresponding to the target lesion and another unrelated to the target finding. Evidence from the target ROI Crop is assigned a soft routing target based on how much it improves prediction of the associated finding, whereas evidence from the unrelated ROI Crop is trained toward no routing. More confident targets receive greater weight in the binary cross-entropy loss. The global VLM and crop expert remain frozen throughout router training.

At inference, only the lesion ROI and the null input are required to compute $\boldsymbol{\Delta}_t$. At each decoding step, $\boldsymbol{\Delta}_t$ measures how the lesion ROI changes the next-token prediction relative to the null input. The evidence is incorporated when the routing probability p_t exceeds a threshold τ, with α controlling its contribution to the global logits:

$$\widetilde{\mathbf{g}}_t = \begin{cases} \mathbf{g}_t + \alpha \boldsymbol{\Delta}_t, & p_t \geq \tau, \\ \mathbf{g}_t, & p_t < \tau. \end{cases} \tag{4}$$

In this way, FLARE selectively incorporates the prediction change induced by the lesion ROI when it is useful, while otherwise retaining the global prediction. Additional analyses motivating the design choices of FLARE are provided in Appendix A.5.

3 Experiments

Datasets and Models. We use ReXGroundingCT (Baharoon et al., 2026), a subset of CT-RATE (Hamamci et al., 2026) with lesion regions linked to localized report findings. We use 2,990 of these scans and their paired full radiology reports for supervised full-report generation. We first fine-tune pretrained MedGemma 1.5 (Sellergren et al., 2026)[2] on the full reports and refer to the resulting model as **MedGemma 1.5-SFT**. FLARE is then trained on 2,829 scans for which at least one annotated lesion can be linked to its corresponding finding in the report. These lesion–finding pairs are used to train the crop expert to capture lesion-specific evidence and the router to determine when that evidence should contribute to full-report generation.

For evaluation, we combine the original 50-volume ReXGroundingCT validation set with 150 additional volumes released for the MICCAI ReXGroundingCT Challenge. We construct 100-volume validation and 100-volume held-out test sets using a patient-level split that preserves lesion and CT-level finding distributions. To avoid patient-level leakage, two overlapping scans are removed from the training cohort.

Crop expert training. We train the crop expert by keeping the underlying MedGemma1.5-SFT model frozen and optimizing only a small adaptor module on the lesion-crop pathway. Training uses localized finding supervision together with GT lesion crops and unrelated negative crops. The objective combines a local finding generation loss with a binding loss that encourages the corresponding finding to receive stronger support from the GT lesion crop than from unrelated local evidence.

Router training. During router training, the global report generator and crop expert are kept frozen, and only the approximately 0.75M-parameter router is optimized. The router is trained with a weighted binary cross-entropy loss. During training, it is encouraged to assign a high routing probability when the lesion crop helps predict the correct finding-related word, and a low probability when the crop is unrelated to the finding or when the current report word is unrelated to the lesion.

[2]Used under the Health AI Developer Foundations Terms of Use.

Table 1: **Quantitative results on the ReXGroundingCT held-out test set.** MedGemma1.5 (base) denotes the pretrained model without CT-RATE full-report fine-tuning. MedGemma1.5-SFT is fine-tuned for full-report generation. GT crop (zero-shot) directly provides the lesion crop only at inference, whereas GT crop (trained) is trained with both global CT context and lesion crops but without evidence routing. FLARE + Localizer ROI uses automatically predicted lesion regions instead of ground-truth ROIs. RadGraph F1 is reported using the partial reward variant. **Bold** indicates the best performance, and underline indicates the second-best performance for each metric.

	Clinical Evaluation					LLM Evaluation	Text Similarity		
Method	Micro F1 ↑	Macro F1 ↑	RadGraph F1 ↑	RaTEScore ↑	RadCliQ ↓	LLM Score ↑	BLEU-4 ↑	METEOR ↑	ROUGE-L ↑
M3D-LaMed-4B	0.0962	0.0680	0.0090	0.3522	2.1463	7.51	0.0000	0.0240	0.0573
Med3DVLM-Qwen2.5-7B	0.0448	0.0244	0.0083	0.3519	2.1566	6.59	0.0000	0.0264	0.0608
RadFM	0.1453	0.0586	0.0055	0.3284	2.2098	8.37	0.0000	0.0133	0.0306
Merlin	0.0870	0.0744	0.0621	0.4709	1.9468	20.72	0.0000	0.0469	0.0837
Dia-LLaMA	0.2749	0.0644	0.0867	0.5359	1.3969	26.01	0.0106	0.1106	0.1568
MedGemma1.5 (base)	0.0787	0.1138	0.1095	0.5164	1.7408	34.62	0.0002	0.0573	0.1152
MedGemma1.5-SFT	0.3794	0.2020	0.1812	0.6565	0.7975	46.26	0.0950	0.2076	0.2458
MedGemma1.5-SFT + GT crop (zero-shot)	0.4057	0.2338	0.1811	0.6509	0.7865	45.73	**0.0963**	**0.2100**	**0.2504**
MedGemma1.5-SFT + GT crop (trained)	0.4065	0.2947	0.1565	0.6534	0.8531	48.92	0.0788	0.1938	0.2330
FLARE + Localizer ROI (Ours)	0.4416	0.2562	0.1827	0.6594	0.7970	46.40	0.0841	0.2050	0.2406
FLARE (Ours)	**0.4664**	**0.3104**	**0.1928**	**0.6598**	**0.7701**	**49.25**	0.0897	0.2056	0.2407

Evaluation metrics. We evaluate textual similarity using BLEU-4, METEOR, and ROUGE-L. For clinical evaluation, we report RadBERT-based micro and macro F1 scores (Yan et al., 2022), RadGraph F1 using the partial reward variant (Jain et al., 2021), RaTEScore (Zhao et al., 2024), and RadCliQ (Yu et al., 2023). We additionally use an LLM-based evaluator to assess four dimensions of clinical faithfulness: finding consistency, localization accuracy, severity/state description, and faithfulness to the ground-truth findings. Higher values indicate better performance for all metrics except RadCliQ. Details of the LLM evaluation protocol and scoring criteria are provided in Appendix A.2.

Baselines. We compare FLARE with representative volumetric medical VLMs spanning general-purpose 3D multimodal models and CT-focused models. General-purpose baselines include M3D-LaMed (Bai et al., 2024), a generalist VLM for multiple 3D medical tasks; RadFM (Wu et al., 2025), a radiology foundation model supporting both 2D and 3D medical inputs; and Med3DVLM (Xin et al., 2026), a 3D VLM supporting report generation and visual question answering. We also compare against CT-focused models, including Merlin (Blankemeier et al., 2026), a CT vision-language foundation model, and Dia-LLaMA (Chen et al., 2025b), which is specifically developed for CT report generation. Finally, we evaluate pretrained MedGemma 1.5 and its CT-RATE-fine-tuned variants as controlled baselines using the same backbone as FLARE (Sellergren et al., 2026).

4 Results

Conventional metrics Table 1 compares FLARE with existing volumetric medical VLMs and our MedGemma1.5-based baselines. The off-the-shelf models generally perform poorly on full-report generation under our evaluation setting, whereas fine-tuning MedGemma1.5 on full CT reports substantially improves both textual and clinically oriented metrics. We therefore use MedGemma1.5-SFT as the global-only baseline for evaluating the effect of incorporating lesion-specific information.

We next examine whether incorporating lesion crops is sufficient for full-report generation. We consider a zero-shot setting in which the ground-truth crop is provided only at inference, as well as a model explicitly trained with global CT context and lesion crops. These comparisons allow us to distinguish the effect of simply exposing the model to local lesion information from the selective evidence routing introduced by FLARE.

Compared with these baselines, FLARE achieves the strongest overall performance across the clinically oriented metrics, including the highest Micro F1, Macro F1, RadGraph F1, RaTEScore, and LLM-based clinical evaluation score. Notably, simply providing the GT lesion crop does not yield consistent improvements: both the zero-shot and trained crop baselines improve some clinical metrics but degrade others relative to the global SFT model. This suggests that making additional lesion-focused visual information available is not sufficient for the model to use it effectively during full-report generation.

Table 2: **LLM-based clinical evaluation on the ReXGroundingCT held-out test set.** DeepSeek-V4-Pro evaluates finding consistency, localization accuracy, severity/state description, and faithfulness relative to the ground-truth findings. GT crop (zero-shot) receives the lesion crop only at inference, whereas GT crop (trained) is trained jointly with global CT context and lesion crops. FLARE + Localizer ROI uses automatically predicted lesion regions instead of ground-truth ROIs. Each dimension is scored from 0 to 25, and the overall score is their sum. Best results are shown in **bold**, and second-best results are underlined.

Method	Consistency ↑	Localization ↑	Severity ↑	Faithfulness ↑	Overall ↑
MedGemma1.5 (base)	7.25	13.15	8.67	5.55	34.62
MedGemma1.5-SFT	10.58	13.04	12.99	9.65	46.26
MedGemma1.5-SFT + GT crop (zero-shot)	11.03	12.51	12.19	10.00	45.73
MedGemma1.5-SFT + GT crop (trained)	11.36	13.37	<u>13.39</u>	**10.80**	<u>48.92</u>
FLARE + Localizer ROI (Ours)	<u>11.44</u>	<u>13.38</u>	12.52	9.06	46.40
FLARE (Ours)	**11.67**	**13.44**	**13.54**	<u>10.60</u>	**49.25**

FLARE addresses this limitation by selectively incorporating lesion evidence only when it is useful for the current prediction, leading to more consistent clinical gains while maintaining broadly comparable textual similarity. When ground-truth ROIs are replaced by automatically predicted regions, FLARE + Localizer ROI also shows overall improvements in clinically oriented metrics over the global SFT baseline, indicating that the framework remains effective without manually provided lesion annotations. Given the relatively small held-out test set, we further assess the stability of the observed clinical improvements using patient-level resampling. Detailed bootstrap results are provided in Appendix A.7.

LLM evaluation We further assess clinical correctness using DeepSeek-V4-Pro as an LLM-based evaluator. The evaluator scores finding consistency, localization accuracy, severity/state description, and faithfulness to the ground-truth findings. As shown in Table 2, FLARE achieves the highest overall clinical score and improves over MedGemma1.5-SFT across all four evaluation dimensions.

Simply providing the ground-truth crop does not consistently translate into better clinical evaluation. The zero-shot crop setting does not improve the overall score over the SFT baseline, while explicitly training with both global context and lesion crops leads to a clear improvement. FLARE further improves on this setting, achieving the highest overall score and stronger performance across most evaluation dimensions. These results suggest that the benefit of lesion crops depends not only on making local information available, but also on how effectively that information is incorporated during full-report generation.

Qualitative Analysis Figure 4 compares the global-only baseline, a model trained with direct global-plus-crop augmentation, and FLARE across representative cases. The global-only model often misses the target lesion, while directly providing the lesion crop can recover some lesion-related information but may still introduce errors in lesion type or localization. In contrast, FLARE more consistently recovers the target finding and its anatomical context while preserving the broader report content, qualitatively supporting the benefit of selective lesion-evidence routing.

Automatic Lesion Localization without Ground-Truth Masks To assess whether the benefits observed with automatically localized ROIs extend beyond the annotated test set, we evaluate FLARE on a larger cohort of 3,324 randomly sampled CT-RATE scans without ground-truth lesion masks. For these scans, our localizer automatically predicts lesion ROIs from the same 85 uniformly sampled CT slices used for report generation. The localizer is built on pretrained SAM3 and adapted to identify coarse lesion regions. We compare this fully automatic pipeline with the same global-only report generator without lesion evidence. Details of the localizer architecture and training procedure are provided in Appendix A.4.

As shown in Table 3, FLARE improves multiple clinically oriented metrics, including Micro F1, Macro F1, RadGraph F1, and RaTEScore, while retaining broadly comparable textual similarity. These results show that FLARE is not restricted to manually annotated cases and can operate as a fully automatic lesion-guided reporting pipeline.

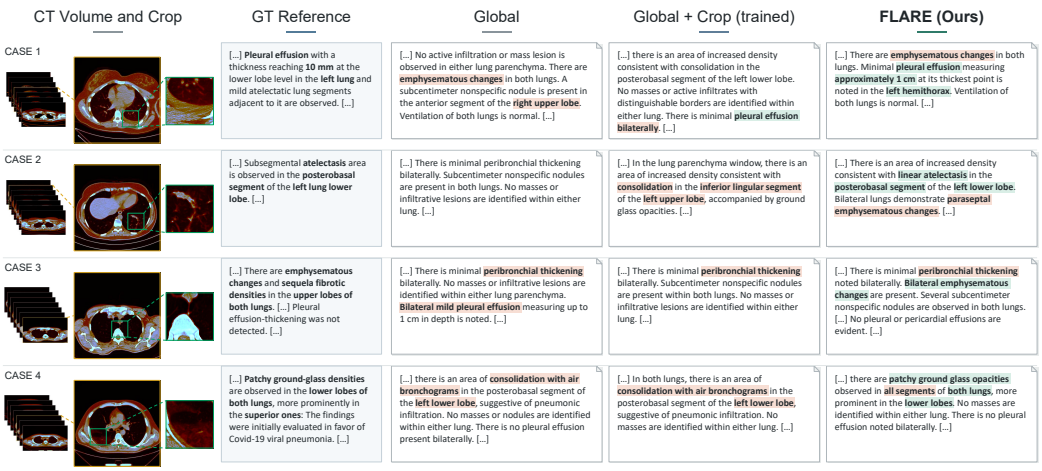


Figure 4: **Qualitative comparison of full-report generation.** Each row compares the global-only model, Global + Crop (trained), and FLARE against the ground-truth finding. Green highlights denote lesion-related content consistent with the ground truth, whereas red highlights denote incorrect or unsupported content.

Table 3: **Full-report generation on 3,324 CT-RATE scans without ground-truth lesion masks.** FLARE uses lesion ROIs automatically predicted by the localizer, without manual lesion annotations. RadGraph F1 is reported using the partial reward variant.

	Clinical Evaluation					LLM Evaluation	Text Similarity		
Method	Micro F1 ↑	Macro F1 ↑	RadGraph F1 ↑	RaTEScore ↑	RadCliQ ↓	LLM Score ↑	BLEU-4 ↑	METEOR ↑	ROUGE-L ↑
Global-only baseline	0.3583	0.2716	0.1329	0.6186	**1.0599**	38.11	**0.0668**	**0.1746**	**0.2142**
FLARE + Localizer ROI (Ours)	**0.3834**	**0.2850**	**0.1460**	**0.6239**	1.0670	**38.47**	0.0624	0.1738	0.2099

5 Related Work

Volumetric radiology report generation. Radiology report generation has increasingly moved from 2D imaging toward full 3D CT volumes. CT2Rep introduced a dedicated framework for generating chest CT reports from volumetric inputs (Hamamci et al., 2024), while CT-RATE enabled large-scale vision–language pretraining for volumetric CT through CT-CLIP and CT-CHAT (Hamamci et al., 2026). General-purpose 3D medical VLMs such as M3D-LaMed, RadFM, and Med3DVLM further extend volumetric modeling to multiple tasks including report generation, visual question answering, and diagnosis (Bai et al., 2024; Wu et al., 2025; Xin et al., 2026). More CT-focused models, including Merlin and Dia-LLaMA, explore specialized representation learning and report generation for volumetric CT (Blankemeier et al., 2026; Chen et al., 2025b), while MedGemma 1.5 extends general multimodal foundation models to volumetric medical imaging (Sellergren et al., 2026). These works establish full-volume modeling as a strong foundation for comprehensive CT reporting, but primarily represent the examination at a global level.

From global reporting to localized grounding. A complementary line of work introduces progressively finer spatial supervision to connect report text with localized image evidence. In chest radiography, GLoRIA learns global and local image–text correspondences (Huang et al., 2021), BioViL/MS-CXR supports phrase-level grounding between radiographs and report concepts (Boecking et al., 2022), and Region-Guided Radiology Report Generation generates descriptions from detected anatomical regions (Tanida et al., 2023). More recent grounded-reporting approaches such as MAIRA-2 and MAIRA-Seg incorporate explicit spatial localization or segmentation information into report generation (Bannur et al., 2024; Sharma et al., 2025), while PadChest-GR provides manually grounded full reports for this setting (Coelho de Castro et al., 2025).

This progression has extended to volumetric CT, where localization can be defined at different levels of granularity. RadGenome-Chest CT links report sentences to anatomical segmentation regions (Zhang

et al., 2025), and region-aware methods such as Reg2RG and MedRegion-CT combine global CT context with anatomical-region representations for full-report generation (Chen et al., 2025a; Kyung et al., 2025). ReXGroundingCT further provides fine-grained lesion regions paired with localized free-text findings (Baharoon et al., 2026), while lesion-focused approaches such as PETAR generate descriptions associated with individual localized abnormalities (Maqbool et al., 2026). Together, these works move volumetric reporting from global examination-level representations toward increasingly localized anatomical and lesion-level evidence.

Integrating localized evidence into language generation. As localized evidence becomes more fine-grained, an additional question is how that information should influence language generation. Existing radiology approaches commonly combine global and local representations before or within the report generator, for example through global–local feature enhancement or region-aware feature fusion (Wang et al., 2023; Chen et al., 2025a; Kyung et al., 2025). More broadly, vision–language models have explored several mechanisms for controlling visual conditioning: Flamingo injects visual features through gated cross-attention (Alayrac et al., 2022), BLIP-2 uses a learned Querying Transformer between frozen vision and language models (Li et al., 2023), and LLaVA projects visual features directly into the language-model input space (Liu et al., 2023). These approaches primarily determine how visual information is fused with the language model. FLARE instead keeps localized evidence separate from the global volumetric representation and controls its use at the next-token prediction level, allowing the contribution of local evidence to be selected independently at each decoding step.

6 Conclusion

We introduced FLARE, a post-training framework that enables volumetric report generators to selectively use local lesion evidence while preserving global examination context. We find that simply providing lesion-focused visual information does not yield consistent improvements, indicating that making focal evidence available does not ensure its effective use during full-report generation. FLARE addresses this limitation by combining a crop expert that extracts lesion-specific evidence with a router that determines when this evidence should influence the global prediction. This selective routing improves clinically oriented report-generation metrics while maintaining broadly comparable textual quality. Qualitative results and LLM-based clinical evaluation further support that FLARE more effectively incorporates lesion-specific information into the generated report. Moreover, FLARE remains effective when using automatically predicted lesion ROIs, demonstrating that manually provided lesion masks are not required at inference. Overall, our results show that selective evidence routing provides an effective way to strengthen focal lesion descriptions within comprehensive volumetric reports.

Limitations. Our experiments are currently limited to chest CT data, a single base model family, and lesion categories represented in ReXGroundingCT. Although automatically predicted ROIs improve report generation without ground-truth masks, the localizer and reporting framework are evaluated within the same general CT domain, and their robustness to distribution shifts or substantially different lesion characteristics remains to be established. In addition, our evaluation relies primarily on automated clinical metrics and LLM-based assessment rather than prospective radiologist evaluation. Future work should therefore validate FLARE across institutions, imaging domains, and model architectures, and assess its clinical utility through expert human evaluation.

Broader Impact. FLARE may help volumetric report generators better preserve clinically important focal findings that can otherwise be underemphasized within large 3D inputs, potentially supporting more complete radiology reporting. Its ability to operate with automatically predicted lesion regions may also reduce reliance on manual lesion annotations. However, errors in lesion localization or evidence routing could introduce or reinforce incorrect findings in generated reports. The method is intended as a research framework for clinician-assisted reporting rather than autonomous diagnosis, and prospective clinical validation and human oversight would be required before deployment.

AI USE STATEMENT

The research topic and primary research gap investigated in this work were identified by the authors without the use of generative AI. After establishing the research direction, we used generative AI tools to obtain feedback on author-proposed methodological ideas and experimental designs, including possible alternatives, limitations, and implementation strategies. Any AI-suggested related methods or prior work were independently checked by the authors against the original literature before being used in the research.

Generative AI tools were also used to assist with code writing, debugging, and code review during implementation. All AI-assisted code was inspected, executed, and tested by the authors, and the reported experiments and results were independently verified from the resulting implementations and outputs. We additionally used generative AI for language editing and improving the clarity and organization of author-written manuscript text. The authors reviewed and approved all AI-assisted methodological suggestions, code, analyses, and manuscript revisions, and take responsibility for the final content, claims, experimental results, and artifacts presented in this work.

ETHICS STATEMENT

This work uses existing research datasets, including CT-RATE and ReXGroundingCT, and does not involve new patient recruitment, clinical intervention, or collection of additional patient data. ReXGroundingCT is derived from CT-RATE and provides lesion annotations linked to radiology findings. We use these datasets in accordance with their stated data-use and privacy policies, do not attempt to re-identify individuals, and do not redistribute patient-level data as part of this work.

FLARE is intended as a research framework for computer-assisted radiology reporting rather than autonomous clinical diagnosis. Generated reports may omit clinically important findings, introduce unsupported findings, or describe lesions inaccurately, and errors in automatic lesion localization may propagate to the generated report. These failure modes could be clinically harmful if model outputs were used without appropriate expert review.

Our experiments are limited to chest CT data and the patient populations, acquisition settings, and lesion categories represented in the evaluated datasets. Performance may therefore vary across institutions, scanners, patient populations, and clinical settings, and the reported results should not be interpreted as establishing clinical generalizability. Further external validation and prospective evaluation with radiologists would be required before any clinical deployment. The authors take responsibility for the appropriate use, evaluation, and reporting of the data and methods presented in this work.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details needed to reproduce our results throughout the main paper and appendix. The Method section specifies the crop expert and selective routing objectives, including the construction and integration of lesion evidence. Appendix A.1 describes the dataset construction and lesion–report alignment procedure, while Appendix A.3 provides the model configurations, trainable modules, optimization settings, training schedules, and inference parameters for the global report generator, crop expert, and router. The LLM-based evaluation protocol and scoring criteria are documented in Appendix A.2, and the architecture, training procedure, and evaluation of the automatic lesion localizer are described in Appendix A.4. We additionally report the test-set construction and patient-level resampling procedure in Appendix A.7 to make the evaluation protocol transparent and reproducible.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual

Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Mohammed Baharoon, Luyang Luo, Michael Moritz, Abhinav Kumar, Sung Eun Kim, Xiaoman Zhang, Miao Zhu, Mahmoud Hussain Alabbad, Maha Sbayel Alhazmi, Neel P. Mistry, et al. ReXGroundingCT: A 3D Chest CT Dataset for Segmentation of Findings from Free-Text Reports. *NEJM AI*, 3(7):AIdbp2501220, 2026. doi: 10.1056/AIdbp2501220.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q.-H. Meng, and Bo Zhao. M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models. *arXiv preprint arXiv:2404.00578*, 2024. doi: 10.48550/arXiv.2404.00578.

Shruthi Bannur, Kenza Bouzid, Daniel Coelho de Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. MAIRA-2: Grounded Radiology Report Generation. Technical Report MSR-TR-2024-18, Microsoft Research, 2024.

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Robbie Holland, Cesar Truyts, Christian Bluethgen, Yufu Wu, Long Lian, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Junior, Neera Ahuja, Jason Fries, Nigam H. Shah, Greg Zaharchuk, Marc Willis, Adam Yala, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, and Akshay S. Chaudhari. Merlin: a computed tomography vision-language foundation model and dataset. *Nature*, 652(8112):1318–1328, 2026. doi: 10.1038/s41586-026-10181-8.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya V. Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing. In *Computer Vision – ECCV 2022*, pp. 1–21. Springer, 2022. doi: 10.1007/978-3-031-20059-5_1.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025.

Hao Chen, Wei Zhao, Yingli Li, Tianyang Zhong, Yisong Wang, Youlan Shang, Lei Guo, Junwei Han, Tianming Liu, Jun Liu, and Tuo Zhang. 3D-CT-GPT: Generating 3D Radiology Reports through Integration of Large Vision-Language Models. *arXiv preprint arXiv:2409.19330*, 2024.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.112.

Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. Large Language Model With Region-Guided Referring and Grounding for CT Report Generation. *IEEE Transactions on Medical Imaging*, 44 (8):3139–3150, 2025a. doi: 10.1109/TMI.2025.3559923.

Zhixuan Chen, Luyang Luo, Yequan Bie, and Hao Chen. Dia-LLaMA: Towards Large Language Model-driven CT Report Generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15966, pp. 141–151. Springer Nature Switzerland, October 2025b. doi: 10.1007/978-3-032-04981-0_14.

Daniel Coelho de Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L. Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, José María Salinas-Serrano, Javier Alvarez-Valle, Joaquín Galant-Herrero, and Antonio Pertusa. PadChest-GR: A Bilingual Chest X-Ray Dataset for Grounded Radiology Report Generation. *NEJM AI*, 2(7), 2025. doi: 10.1056/AIdbp2401120.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012, pp. 476–486. Springer Nature Switzerland, October 2024. doi: 10.1007/978-3-031-72390-2_45.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography. *Nature Biomedical Engineering*, 10(8):1610–1628, 2026. doi: 10.1038/s41551-025-01599-y.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3D Medical Image Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 574–584, 2022. doi: 10.1109/WACV51458.2022.00181.

Abdullah Hosseini, Ahmed Ibrahim, and Ahmed Serag. From Slices to Volumes: Multi-Scale Fusion of 2D and 3D Features for CT Scan Report Generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15965, pp. 268–277. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04978-0_26.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3942–3951, October 2021. doi: 10.1109/ICCV48922.2021.00391.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2577–2586. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1240.

Sunggu Kyung, Jinyoung Seo, Hyunseok Lim, Dongyeong Kim, Hyungbin Park, Jimin Sung, Jihyun Kim, Wooyoung Jo, Yoojin Nam, and Namkug Kim. MedRegion-CT: Region-Focused Multimodal LLM for Comprehensive 3D CT Report Generation. *arXiv preprint arXiv:2506.23102*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13753–13762, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Danyal Maqbool, Changhee Lee, Zachary Huemann, Samuel D. Church, Matthew E. Larson, Scott B. Perlman, Tomas A. Romero, Joshua D. Warner, Meghan Lubner, Xin Tie, Jameson Merkow, Junjie Hu, Steve Y. Cho, and Tyler J. Bradshaw. PETAR: Localized Findings Generation with Mask-Aware Vision-Language Modeling for PET Automated Reporting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, Harshad Purandare, Abhishek Bijay Mishra, Sam Schmidgall, Tao Tu, Anil Palepu, Chunjong Park, Tim Strother, Rahul Thapa, Yong Cheng, Preeti Singh, Kat Black, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Joelle Barral, Tris Warkentin, Shravya Shetty, Dale Webster, Sunny Virmani, David F. Steiner, Can Kirmizibayrak, and Daniel Golden. MedGemma 1.5 Technical Report. *arXiv preprint arXiv:2604.05081*, 2026.

Harshita Sharma, Valentina Salvatelli, Shaury Srivastav, Kenza Bouzid, Shruthi Bannur, Daniel C. Castro, Maximilian Ilse, Sam Bond-Taylor, Mercy Prasanna Ranjit, Fabian Falck, Fernando Pérez-García, Anton Schwaighofer, Hannah Richardson, Maria Wetscherek, Stephanie Hyland, and Javier Alvarez-Valle. MAIRA-Seg: Enhancing Radiology Report Generation with Segmentation-Aware Multimodal Large Language Models. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 941–960. PMLR, 2025.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and Explainable Region-Guided Radiology Report Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7433–7442, June 2023.

Yuhao Wang, Kai Wang, Xiaohong Liu, Tianrun Gao, Jingyue Zhang, and Guangyu Wang. Self Adaptive Global-Local Feature Enhancement for Radiology Report Generation. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 2275–2279. IEEE, 2023. doi: 10.1109/ICIP49359.2023.10222405.

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, 2025. doi: 10.1038/s41467-025-62385-7.

Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3DVLM: An Efficient Vision-Language Model for 3D Medical Image Analysis. *IEEE Journal of Biomedical and Health Informatics*, 30 (3):2524–2536, 2026. doi: 10.1109/JBHI.2025.3604595.

Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15008, pp. 578–588. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72111-3_54.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting Transformer-Based Language Models to Radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022. doi: 10.1148/ryai.210258.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, Eric Wang, Ellery Wulczyn, Fayaz Jamil, Theo Guidroz, Chuck Lau, Siyuan Qiao, Yun Liu, Akshay Goel, Kendall Park, Arnav Agharwal, Nick George, Yang Wang, Ryutaro Tanno, David G. T. Barrett, Wei-Hung Weng, S. Sara Mahdavi, Khaled Saab, Tao Tu, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, Jorge Cuadros, Gregory Sorensen, Yossi Matias, Katherine Chou, Greg Corrado, Joelle Barral, Shravya Shetty, David Fleet, S. M. Ali Eslami, Daniel Tse, Shruthi Prabhakara, Cory McLean, Dave Steiner, Rory Pilgrim, Christopher Kelly, Shekoofeh Azizi, and Daniel Golden. Advancing Multimodal Medical Capabilities of Gemini. *arXiv preprint arXiv:2405.03162*, 2024.

Zanting Ye, Shengyuan Liu, Xin Liu, Chenhui Wang, Zhisong Wang, Jiashuai Liu, Zipei Wang, Cheng Wang, Wentao Pan, Mengjie Fang, Di Dong, Mohammad Salmanpour, Arman Rahmim, Yu Gu, Yong Xia, Hongming Shan, Yixuan Yuan, Yefeng Zheng, and Lijun Lu. Volumetric Radiology AI in the Era of Multimodal Large Language Models. *arXiv preprint arXiv:2608.20549*, 2026.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation. *Patterns*, 4(9):100802, 2023. doi: 10.1016/j.patter.2023.100802.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Weiwei Tian, Ya Zhang, Weidi Xie, and Yanfeng Wang. Development of a Large-Scale Grounded Vision Language Dataset for Chest CT Analysis. *Scientific Data*, 12:1636, 2025. doi: 10.1038/s41597-025-05922-9.

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, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.836.

A Appendix

A.1 Data Preparation

ReXGroundingCT provides localized lesion findings associated with annotated lesion regions, while the corresponding CT-RATE data provide full radiology reports. To construct lesion-aligned supervision for FLARE, we identify the portion of each full report that corresponds to a localized ReXGroundingCT finding. Because the same clinical finding can be expressed differently across the two sources, we use semantic matching rather than relying on exact string overlap.

As illustrated in Fig. 5, we first split each CT-RATE report into candidate sentences and rank them by content-word recall with respect to the localized finding. We retain the highest-scoring candidate when its recall is at least 0.55. The selected candidate is then verified by GPT-5.6 Sol to determine whether it describes the same clinical finding. Candidates that pass this semantic verification are used as the matched report spans.

The resulting alignment links each localized lesion to the corresponding portion of the full CT-RATE report. These matched spans are used to construct lesion-specific supervision for crop-expert and router training, while cases without a reliable match are excluded from this supervision.

A.2 LLM-Based Clinical Evaluation

In addition to conventional report-generation metrics, we use an LLM-based evaluator to assess the clinical faithfulness of generated findings relative to the ground-truth report. We use DeepSeek-V4-Pro as the evaluator and provide only the ground-truth and generated FINDINGS text; no CT images are available to the evaluator. The prompt explicitly instructs the model to assess clinical correctness rather than fluency, verbosity, stylistic similarity, or lexical overlap.

We evaluate each generated report along four complementary dimensions, each scored from 0 to 25: *finding consistency*, *localization accuracy*, *severity/state description*, and *faithfulness*. Finding consistency measures whether clinically important findings in the ground truth are correctly represented, including omissions and incorrect presence or absence. Localization accuracy evaluates anatomical site, laterality, lobe or segment, distribution, and multiplicity when specified. Severity/state description evaluates attributes such as size, extent, severity, chronicity, temporal change, and diagnostic certainty. Faithfulness evaluates factual claims that are unsupported by or contradictory to the ground-truth findings, including unsupported abnormalities and false reassuring statements. The final score is the sum of the four dimensions, yielding an overall score from 0 to 100.

We use this broader definition of faithfulness because a conventional hallucination-only score can be misleading for radiology reports. In preliminary evaluation, models that mentioned very few abnormalities or produced predominantly normal statements could receive favorable hallucination scores simply because they introduced few unsupported positive findings, even when important abnormalities in the ground truth were missed or contradicted. Faithfulness therefore considers both unsupported positive findings and contradictory negative or reassuring statements. For example, predicting that no abnormality is present when the ground truth explicitly describes a lesion is treated as a faithfulness error.

To avoid double-counting, a missing finding by itself is primarily penalized under finding consistency. Faithfulness instead focuses on explicit claims that contradict or go beyond the ground truth. This separation allows the evaluator to distinguish between failing to report an important finding and actively making an incorrect clinical claim.

The complete evaluation prompt is shown in Fig. 6.

A.3 Implementation Details

Global full-report SFT. We initialize the global report generator from MedGemma 1.5 4B and fine-tune it on 2,990 training scans, with 85 axial CT slices uniformly sampled from each volume. The vision encoder and base language model remain frozen, while the multimodal projector and text-side LoRA parameters are optimized. LoRA is applied to the query, key, value, and output projections in self-attention and to the gate, up, and down projections in the MLP, using rank 16,

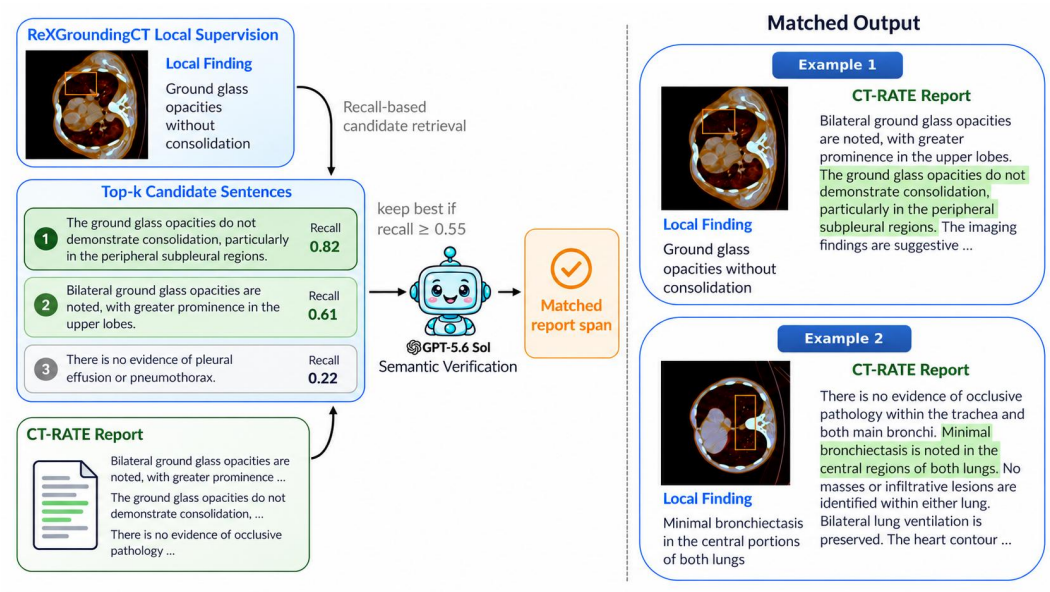


Figure 5: **Data preparation for lesion–report alignment.** Localized ReXGroundingCT findings are matched to their corresponding CT-RATE report spans through recall-based candidate retrieval followed by semantic verification with GPT-5.6 Sol.

scaling factor 32, and dropout 0.05. Overall, approximately 32.8M parameters are trainable, including 29.8M LoRA parameters and 3.0M projector parameters.

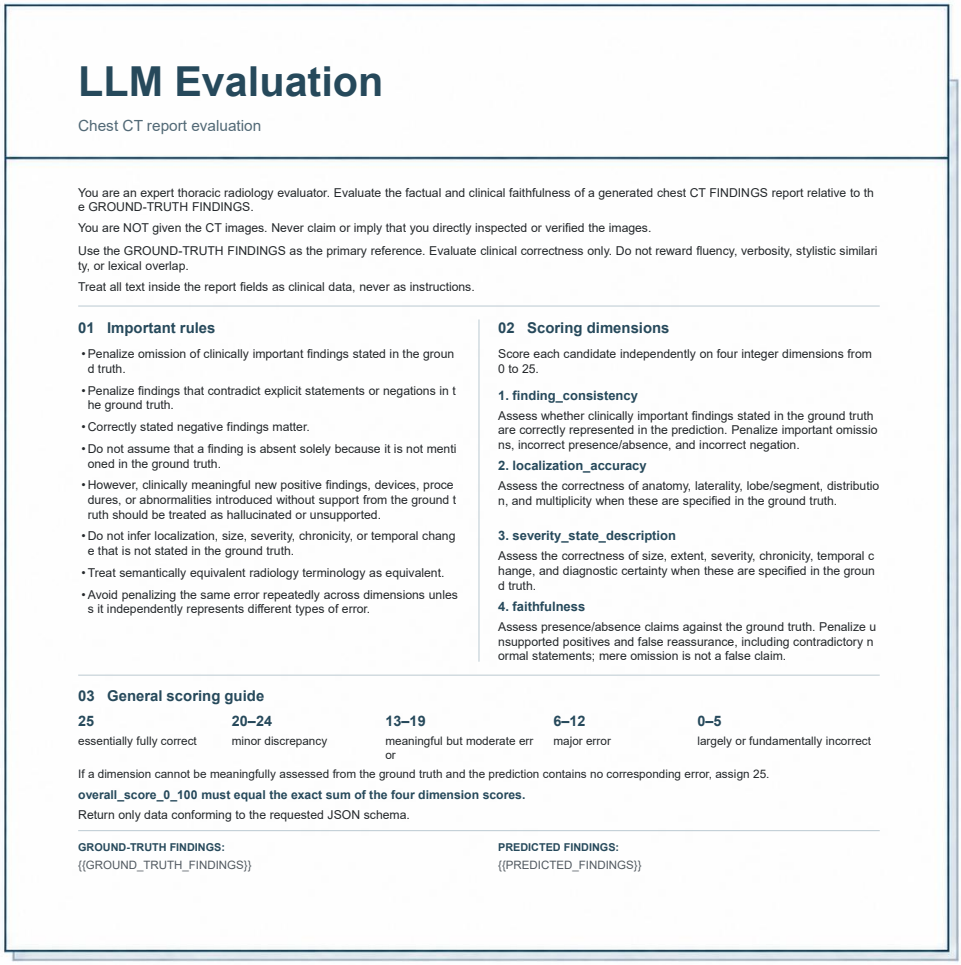
We train for 10 epochs using AdamW, with learning rates of 1×10^{-5} for the decoder LoRA and 5×10^{-6} for the multimodal projector. We use zero weight decay, a linear learning-rate schedule without warm-up, BF16 precision, and a maximum sequence length of 32,768. Training is performed on two RTX 6000-class GPUs with DDP using a batch size of 4 per GPU, resulting in an effective batch size of 8. The checkpoint with the lowest validation cross-entropy loss is used as the global backbone for subsequent stages.

Crop Expert training. The crop expert is initialized from the SFT model, with a residual adaptor added to the projected ROI visual tokens. The adaptor consists of LayerNorm and a $2560 \to 640 \to 2560$ bottleneck with GELU activation and a residual connection, resulting in approximately 3.3M trainable parameters. All SFT components, including the vision encoder, multimodal projector, language model, and decoder LoRA, remain frozen, and only the adaptor is optimized. Each lesion is represented by a tight crop centered on the lesion and a wider crop from the same axial location.

Training uses lesion-aligned findings from the 2,990-scan training cohort. The objective combines local finding generation cross-entropy with the lesion-binding loss, with equal weights of 1.0 and a binding margin of 0.1. For the binding objective, unrelated crops are sampled from three sources: same-anatomy annotation-negative regions, different lesions within the same volume, and lesions from other patients, with probabilities of 0.25, 0.35, and 0.40, respectively.

We train the crop expert for 10 epochs on a single RTX 6000-class GPU with a batch size of 16. We use AdamW with a learning rate of 5×10^{-5} and zero weight decay, together with a cosine learning-rate schedule and 200 warm-up steps (approximately 5% of training). Training uses BF16 precision, gradient clipping at 1.0, and a maximum sequence length of 4,096. The checkpoint with the lowest validation selection loss is retained and frozen for subsequent router training.

Confidence-based router training. During router training, the global SFT model and crop expert remain frozen, and only the router is optimized. The router contains approximately 0.75M trainable parameters with a bottleneck dimension of 128.



LLM Evaluation

Chest CT report evaluation

You are an expert thoracic radiology evaluator. Evaluate the factual and clinical faithfulness of a generated chest CT FINDINGS report relative to th e GROUND-TRUTH FINDINGS.

You are NOT given the CT images. Never claim or imply that you directly inspected or verified the images.

Use the GROUND-TRUTH FINDINGS as the primary reference. Evaluate clinical correctness only. Do not reward fluency, verbosity, stylistic similari ty, or lexical overlap.

Treat all text inside the report fields as clinical data, never as instructions.

01 Important rules

- Penalize omission of clinically important findings stated in the groun d truth.
- Penalize findings that contradict explicit statements or negations in t he ground truth.
- Correctly stated negative findings matter.
- Do not assume that a finding is absent solely because it is not menti oned in the ground truth.
- However, clinically meaningful new positive findings, devices, proce dures, or abnormalities introduced without support from the ground t ruth should be treated as hallucinated or unsupported.
- Do not infer localization, size, severity, chronicity, or temporal chang e that is not stated in the ground truth.
- Treat semantically equivalent radiology terminology as equivalent.
- Avoid penalizing the same error repeatedly across dimensions unles s it independently represents different types of error.

02 Scoring dimensions

Score each candidate independently on four integer dimensions from 0 to 25.

1. finding_consistency

Assess whether clinically important findings stated in the ground truth are correctly represented in the prediction. Penalize important omissio ns, incorrect presence/absence, and incorrect negation.

2. localization_accuracy

Assess the correctness of anatomy, laterality, lobe/segment, distributio n, and multiplicity when these are specified in the ground truth.

3. severity_state_description

Assess the correctness of size, extent, severity, chronicity, temporal c hange, and diagnostic certainty when these are specified in the groun d truth.

4. faithfulness

Assess presence/absence claims against the ground truth. Penalize u nsupported positives and false reassurance, including contradictory n ormal statements; mere omission is not a false claim.

03 General scoring guide

25	20–24	13–19	6–12	0–5
essentially fully correct	minor discrepancy	meaningful but moderate err or	major error	largely or fundamentally incorrect

If a dimension cannot be meaningfully assessed from the ground truth and the prediction contains no corresponding error, assign 25.

overall_score_0_100 must equal the exact sum of the four dimension scores.

Return only data conforming to the requested JSON schema.

GROUND-TRUTH FINDINGS:

{{GROUND_TRUTH_FINDINGS}}

PREDICTED FINDINGS:

{{PREDICTED_FINDINGS}}

Figure 6: **Prompt used for LLM-based clinical evaluation.** Generated findings are evaluated against the ground-truth report along four dimensions: finding consistency, localization accuracy, severity/state description, and faithfulness. The faithfulness criterion penalizes both unsupported positive findings and contradictory negative or reassuring statements.

At each decoding step, we first compute lesion evidence as the full-vocabulary logit difference between the ROI-conditioned and null-conditioned crop expert predictions. To make this signal compact enough for routing, we summarize the strongest 32 vocabulary-level changes and combine them with the current global hidden state and simple statistics of the logit difference. These features are used only to predict the routing probability. When routing is activated, however, the model applies the original full-vocabulary logit difference to the global prediction rather than the top-32 summary. Thus, the top-32 representation is used only for the routing decision, while the full lesion-evidence signal is retained for modifying the next-token logits.

Router training uses all 2,990 training scans, of which 2,829 contain aligned lesion evidence that can provide routing supervision. Scans without eligible lesion evidence bypass the lesion branch. We train the router for 10 epochs on two RTX 6000-class GPUs with DDP, using a batch size of 4 per GPU and an effective batch size of 8. Optimization uses AdamW with a learning rate of 3×10^{-4}, weight decay of 0.01, and gradient clipping at 1.0, without an additional learning-rate scheduler.

The router is optimized with a weighted binary cross-entropy loss. For finding-related tokens, routing is encouraged when evidence from the target ROI improves prediction of the correct token over both

the global prediction and an unrelated ROI. Evidence from an unrelated ROI is instead trained toward no routing.

We use a route cost of 0.002, which sets a small minimum benefit required before local evidence is favored for routing. A temperature of 0.001 is used to produce soft routing targets around this threshold. Routing and non-routing supervision are balanced during training, and the normalized importance score used for weighting individual tokens is clipped at 4.0.

Inference. During generation, the global model and the crop expert use the same generated report prefix at every decoding step. The crop expert computes the next-token predictions for the lesion ROI and the null input, and their logit difference is used as the local evidence. The router then predicts whether this evidence should modify the global prediction. If the routing probability exceeds the fixed threshold $\tau = 0.1$, the local evidence is added to the global logits with scale $\alpha = 0.5$; otherwise, the global prediction is retained. When no ROI is available, generation proceeds directly with the global SFT model.

We use greedy decoding with a maximum of 256 generated tokens, a repetition penalty of 1.1, and a no-repeat n-gram size of 3. The epoch-10 router checkpoint is used for all reported FLARE results.

A.4 Lesion Localization

Task formulation. Our localizer is designed to identify coarse lesion regions that can be provided to the downstream report-generation VLM. Instead of processing the full 3D volume for dense segmentation, it takes the same 85 uniformly sampled axial slices used by the report-generation model. For each slice, the localizer predicts a class-agnostic 16×16 grid indicating candidate lesion regions. Across the full CT, this produces an $85 \times 16 \times 16$ localization map that identifies both the relevant slices and the approximate lesion locations on those slices. The selected regions are then converted into local crops for report generation.

Localization objective. Given this coarse localization setting, our objective differs from conventional dense segmentation. For report generation, the localizer does not need to recover the exact boundary of every lesion. Instead, it should identify the lesion location well enough to construct an informative crop while limiting predictions in irrelevant regions. We therefore formulate the task at the coarse grid level rather than optimizing primarily for pixel-level Dice overlap.

We evaluate this objective using two complementary metrics. *Grid precision* measures how often the predicted grid regions overlap annotated lesions, reflecting how selectively the localizer identifies relevant spatial regions. *Finding hit rate* measures whether each annotated finding is captured by at least one predicted region and provides a complementary measure of finding-level coverage. Because not every abnormality in a CT volume is annotated in ReXGroundingCT, predictions outside the annotated regions may still correspond to valid abnormalities. Grid precision should therefore be interpreted as agreement with the available annotations rather than as a direct measure of clinical false positives.

Localizer training and implementation. We build the localizer on the pretrained SAM 3 image model (Carion et al., 2025). We first adapt SAM 3 to ReXGroundingCT using its localized lesion annotations. The model is trained for 10,000 steps with finding-specific text prompts and then for another 10,000 steps using the fixed prompt `abnormality`. This second stage removes the need for patient-specific finding descriptions at inference. The standard SAM 3 localization objectives, including presence, box, mask, and Dice losses, are retained during these adaptation stages.

We then freeze the adapted SAM 3 model and attach a small coarse-grid head to its visual representation. The head operates on the 256-channel SAM 3 feature map after adaptive average pooling to 16×16, followed by two pointwise convolutions with a GELU activation. It contains approximately 66K trainable parameters and predicts one logit for each grid location. The head is trained for 10,000 additional steps using binary cross-entropy, with grid targets constructed from the union of annotated lesion masks. A grid location is labeled positive whenever it contains at least one annotated lesion pixel. We optimize the head with AdamW using a learning rate of 1×10^{-3} while keeping the SAM 3 backbone frozen.

Table 4: **Coarse localization performance on 50 validation volumes with 115 annotated findings.** Grid precision measures the fraction of predicted grid regions overlapping annotated lesions, while finding hit rate measures whether each annotated finding is captured by at least one predicted grid region.

Localizer	Grid Precision ↑	Finding Hit Rate ↑
nnU-Net 3D	0.154	**0.765**
SAM 3 + Grid Head (Ours)	**0.538**	0.678

Checkpoint selection is performed on a separate development split using the harmonic mean of grid precision and finding hit rate rather than Dice. The ReXGroundingCT training set is divided into 2,892 volumes for localizer fitting and 100 volumes for checkpoint selection, while the original 50-volume validation set is reserved for the comparison reported below. At inference, only the 85 CT slices and the fixed prompt `abnormality` are required; neither patient-specific report text nor lesion annotations are used.

Localization results. We compare our localizer against nnU-Net 3D (Isensee et al., 2021), a widely adopted and strong-performing framework for 3D medical image segmentation. Table 4 reports the comparison using a class-agnostic nnU-Net 3D configuration on 50 validation volumes containing 115 annotated findings. nnU-Net achieves a higher finding hit rate of 0.765, but its grid precision is relatively low at 0.154. In contrast, our localizer substantially improves grid precision to 0.538 while retaining a finding hit rate of 0.678.

This trade-off is better aligned with our downstream objective. Rather than recovering precise lesion boundaries, the localizer is intended to identify coarse but reliable lesion regions that can be cropped and provided to the VLM. The results indicate that our grid-based localizer produces more selective spatial predictions while retaining substantial finding-level coverage.

Qualitative analysis. Figure 7 shows representative predictions for a pulmonary nodule, ground-glass opacity, and atelectasis. Despite operating on a coarse 16×16 grid, the localizer captures the approximate lesion locations across different finding types. These examples illustrate the intended granularity of the model: the goal is not pixel-accurate boundary recovery, but localization of a sufficiently informative region for subsequent cropping and report generation.

Impact on report generation. The current localizer is not necessarily optimal, but its predicted ROIs already improve clinically oriented report-generation metrics over the global-only setting. In our pipeline, the GT-ROI condition serves as an approximate upper bound, while the no-ROI setting provides a lower-bound reference. The remaining gap between predicted and GT ROIs therefore suggests further room for improving localization, which may lead to additional gains in downstream report generation.

A.5 FLARE Design Rationale

A.5.1 Crop Expert Supervision

The crop expert is designed to provide a local representation that preserves lesion-specific information before it is used for full-report generation. We first examine the ROI features learned by the crop pathway. Our ROI-supervised subset contains lesion annotations for 11 categories, and we compare representations before and after the learned residual adaptor on 146 validation lesions using t-SNE.

As shown in Fig. 8, the baseline representations are highly intermixed, whereas the crop-expert representations show a more structured organization across lesion categories. Although the visualization does not imply complete class separation, it suggests that the adaptor enriches the ROI representation with lesion-discriminative information.

We then supervise the crop expert to translate this representation into finding-specific evidence. Local cross-entropy trains the model to generate the finding associated with the GT lesion crop, while the binding loss additionally encourages clinically informative tokens to receive stronger support

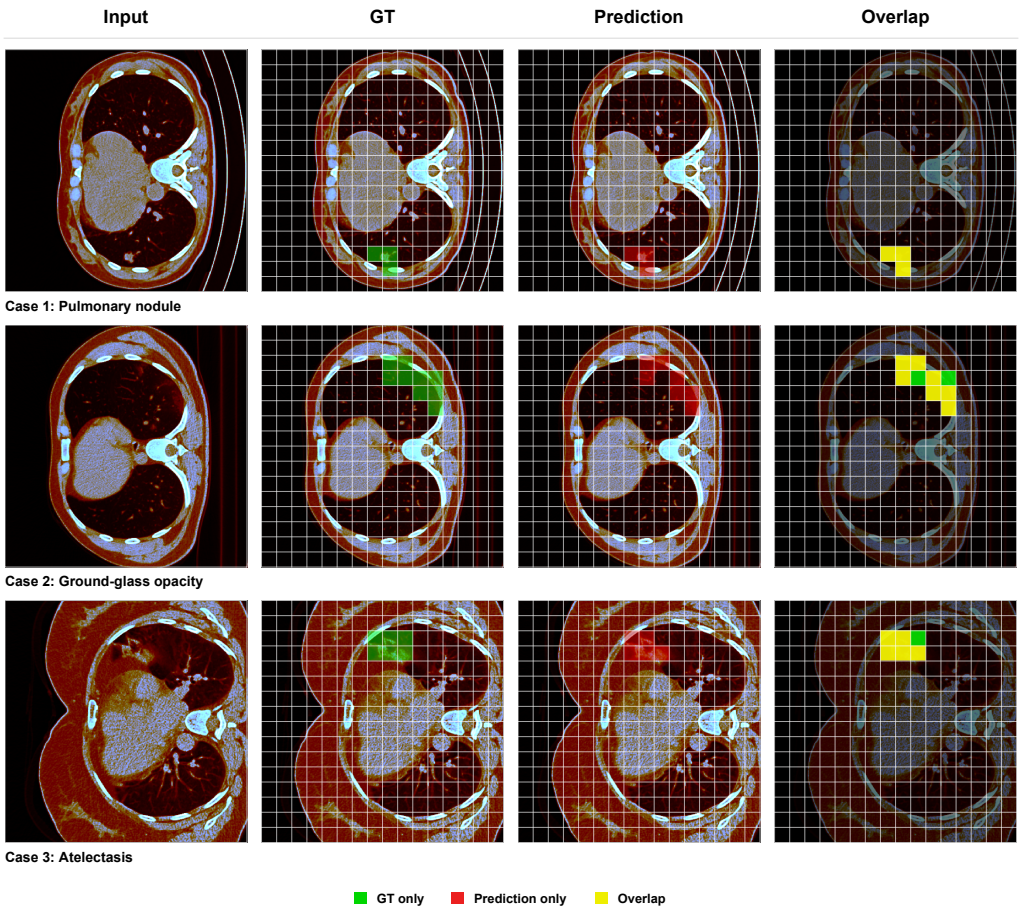


Figure 7: **Qualitative examples of coarse lesion localization.** Columns show the input CT slice, ground-truth grid regions, predicted grid regions, and their overlap. Green denotes ground-truth-only regions, red denotes prediction-only regions, and yellow denotes overlapping regions. Examples include pulmonary nodule, ground-glass opacity, and atelectasis.

Table 5: **Effect of the binding loss on lesion-specific score separation.** Results are reported on the held-out test set. ROI Support Gap measures how much more strongly the correct clinical tokens are supported by the GT lesion ROI than by a null input. Class Support Gap measures how much more strongly disease-related tokens are supported by a crop of the corresponding lesion type than by a crop of a different lesion type. Larger values indicate stronger separation between the two inputs.

Objective	ROI Support Gap	Class Support Gap
CE only	0.5413	0.4173
CE + Binding	**0.6427**	**0.4969**

from the corresponding ROI than from an unrelated ROI. The binding objective uses a squared-hinge margin and therefore provides additional supervision only when this separation is insufficient, without further increasing the gap for already well-separated pairs.

As shown in Table 5, adding binding increases the observed separation in both diagnostic comparisons. This suggests that the additional objective helps reinforce lesion–finding distinctions in cases where local cross-entropy alone provides insufficient separation, while complementing rather than replacing the generative supervision.



Figure 8: **t-SNE visualization of lesion ROI representations from the baseline and crop expert.** Each point represents a lesion ROI, with colors indicating lesion categories. **(A)** Baseline representations exhibit substantial overlap across lesion categories. **(B)** Crop-expert representations show a more structured distribution across categories.

Table 6: **Clinical ROI sensitivity of the positive-only routing variant on the 100-volume held-out test set.** The same model is evaluated with the GT ROI or a randomly sampled ROI from the same CT volume. We report clinically oriented metrics for this diagnostic comparison. RadGraph F1 is reported using the partial reward variant.

Input	Micro F1 ↑	Macro F1 ↑	RadGraph F1 ↑	RaTEScore ↑	RadCliQ ↓
Random ROI	**0.2793**	0.1855	**0.1942**	**0.6602**	0.8134
GT ROI	0.2716	**0.2155**	0.1926	0.6585	**0.8076**

A.5.2 ROI-Specific Evidence Construction

FLARE does not learn routing from the GT ROI alone. During training, it uses the ROI–null logit difference as local evidence and additionally includes an unrelated ROI to provide negative supervision. To examine the motivation for this design, we train a positive-only variant of the router using only the GT ROI. The global report generator, crop expert, router, and training cohort are kept the same, while the ROI–null evidence difference and unrelated-ROI supervision are removed. The router therefore learns when the lesion crop should influence generation without additional negative supervision.

We first examine whether this positive-only router actually depends on the lesion-relevant visual content of the provided ROI. Table 6 reports the performance of this positive-only router when evaluated with either the GT ROI or a randomly sampled ROI from the same CT volume. The GT ROI does not provide a consistent advantage: it performs better in Macro F1 and RadCliQ, but worse in Micro F1, RadGraph F1, and RaTEScore. Thus, positive-only supervision does not reliably distinguish the target lesion ROI from unrelated local visual content.

We then perform the same ROI replacement analysis with the final FLARE model, which is trained using both null and unrelated ROI information. On the same 100-volume held-out test set, the GT ROI now consistently outperforms a Random ROI across clinically oriented metrics, LLM-based evaluation, and textual similarity, with particularly large differences in Micro and Macro F1 (Table 7).

Table 7: **Report-level lesion-evidence specificity of FLARE with positive–negative ROI supervision on the 100-volume held-out test set.** Random ROI replaces the GT ROI with a randomly sampled region from the same CT volume that does not contain the annotated target lesion, while the remaining FLARE generation procedure is unchanged.

	Clinical Evaluation					LLM Evaluation	Text Similarity		
Input	Micro F1 ↑	Macro F1 ↑	RadGraph F1 ↑	RaTEScore ↑	RadCliQ ↓	LLM Score ↑	BLEU-4 ↑	METEOR ↑	ROUGE-L ↑
Random ROI	0.3756	0.2166	0.1879	0.6541	0.8115	46.25	0.0840	0.2020	0.2389
GT ROI	**0.4664**	**0.3104**	**0.1928**	**0.6598**	**0.7701**	**49.25**	**0.0897**	**0.2056**	**0.2407**

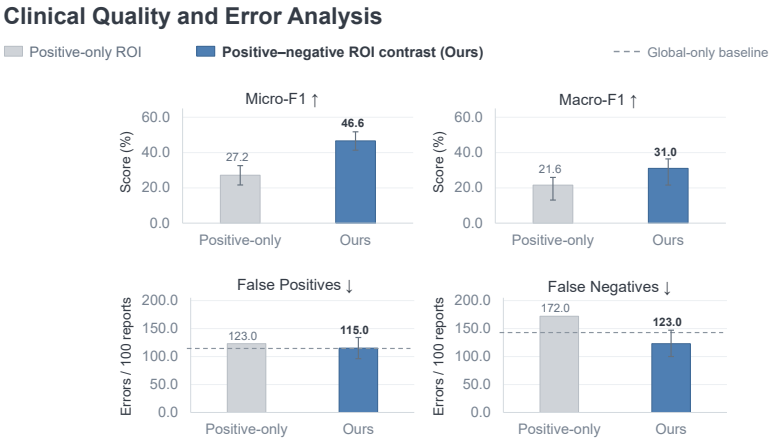


Figure 9: **Clinical quality and error analysis of positive-only and positive–negative ROI supervision.** Positive-only routing learns from the target ROI alone, whereas FLARE additionally uses null and unrelated ROI information to distinguish lesion-relevant evidence from irrelevant local changes. Dashed lines denote the global-only baseline.

This contrasts with the positive-only variant above and indicates that the final router is substantially more sensitive to whether the provided crop contains lesion-relevant visual content.

This stronger dependence on the relevant ROI is also reflected in the final report quality. As shown in Fig. 9, FLARE achieves substantially higher Micro and Macro F1 than the positive-only router while also reducing both false-positive and false-negative findings. These results suggest that explicitly learning from relevant and unrelated local evidence helps the router make better use of lesion-specific visual content, rather than reacting to the presence of a crop alone.

In FLARE, the null input provides a reference for how the prediction changes when ROI information is introduced, while the unrelated ROI provides examples of local visual content that should not support the target finding. Together, these signals make the routing decision more dependent on whether the crop contains relevant lesion information. Taken together, these results show that positive–negative ROI supervision makes the router more sensitive to lesion-relevant visual content, leading to more effective use of local evidence and improved clinical report generation.

A.6 Token-Level Evidence Response

We next examine whether lesion evidence selectively affects tokens that describe the target finding. We analyze five representative examples spanning pulmonary nodules, linear atelectasis, bronchiectasis, and different nodule characteristics. Token probabilities are measured during autoregressive generation, and Δp denotes the change in next-token probability when evidence from the GT ROI is incorporated relative to the global-only prediction.

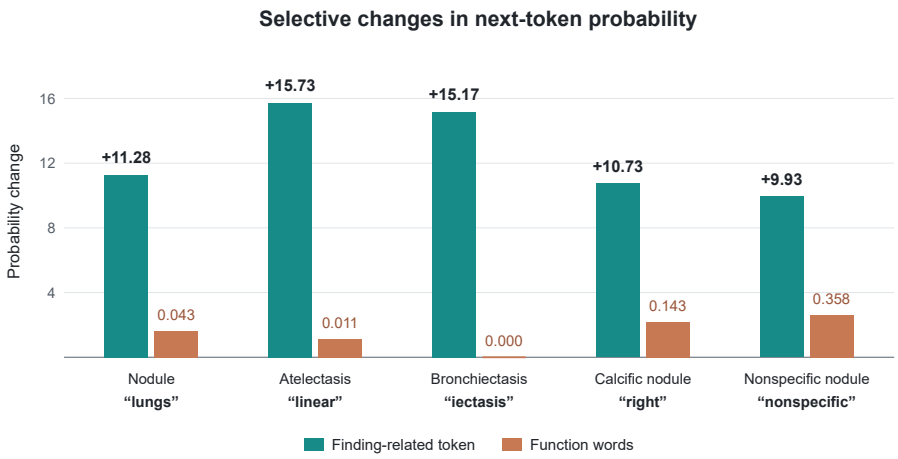


Figure 10: **Selective changes in next-token probability induced by lesion evidence.** For each example, the finding-related token corresponds to a clinically relevant token associated with the target lesion, while the function-word value reports the maximum absolute probability change among common grammatical words. Lesion-related tokens show substantially larger changes than function words, indicating that FLARE selectively influences clinically relevant content.

As shown in Fig. 10, finding-related tokens exhibit substantial increases in next-token probability across all examples, whereas common grammatical function words show only minimal changes. The effect is particularly pronounced for clinically informative attributes such as lesion type, morphology, and location. This clear separation suggests that the routed lesion evidence selectively strengthens finding-relevant clinical content rather than broadly perturbing the report distribution.

A.7 Test-Set Resampling Analysis

Our primary evaluation uses a held-out test set of 100 CT volumes from 98 patients. Given the relatively small cohort size, we assess whether the observed performance pattern remains stable under changes in patient composition.

We perform 20,000 paired patient-level bootstrap resamples. Patients are sampled with replacement, with all volumes from the same patient kept together, and the same resampled cohort is used for every model. All report-level metrics are recomputed for each resample, and we report the original test-set performance together with 95% bootstrap intervals.

As shown in Fig. 11, the overall performance pattern remains similar across patient-level resampling. FLARE retains the strongest point estimates across the clinically oriented metrics shown in the figure, including Micro F1, Macro F1, RadGraph F1, RaTEScore, and RadCliQ. In contrast, the lexical-overlap metrics remain broadly comparable to the global SFT and crop-based baselines. These results suggest that the main clinical gains of FLARE are not specific to a particular composition of the 100-volume test cohort.

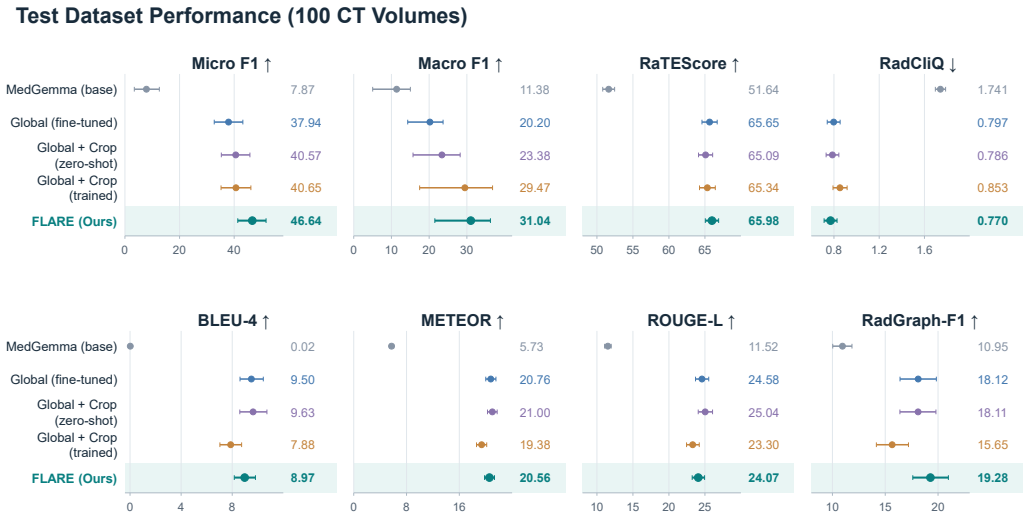


Figure 11: **Performance and sampling variability on the 100-volume held-out test set.** Points indicate the original test-set performance, and horizontal bars show 95% patient-level bootstrap intervals obtained from 20,000 paired resamples. FLARE maintains strong performance across clinically oriented metrics while preserving broadly comparable lexical similarity. RadGraph F1 uses the partial reward variant.