EasyLens: A Training-Free Plug-and-Play Subtle-Lesion Representation Amplifier for Medical Vision-Language Models

Hao Wang[1,2,3*] **Qiwei Zeng**[1,3,*] **Jinghao Lin**[4,*] **Shuchang Ye**[2]
Yuezhe Yang[1,2] **Yige Peng**[1] **Haoyuan Che**[3,†] **Jinman Kim**[2,†] **Lei Bi**[1,†]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]The University of Sydney, Sydney, NSW, Australia
[3]Jilin University, Changchun, China
[4]Northeastern University, Shenyang, China
[*]Equal contribution.
[†]Corresponding authors
chy@jlu.edu.cn jinman.kim@sydney.edu.au lei.bi@sjtu.edu.cn

Abstract

Medical vision-language models (VLMs) have shown increasing potential for clinical image interpretation, including lesion detection and report generation. However, their practical utility remains limited by insufficient sensitivity to subtle lesions, whose visual evidence is often sparse, low-contrast, and embedded within complex anatomical context. As local visual tokens are aggregated, these weak lesion cues can become underrepresented in global image representations, making them difficult for medical VLMs to recognize. Existing efforts to improve lesion sensitivity mainly rely on medical-domain vision-encoder pre-training, clinical-term-guided alignment, or trainable pathological representation enhancement. Although effective, these approaches usually require additional training or model-specific adaptation and may overfit to particular disease morphologies, limiting their applicability to frozen medical VLMs. To address these limitations, we propose **EasyLens**, a training-free plug-and-play subtle-lesion representation amplifier for medical VLMs. EasyLens first constructs **EasyBank**, a pathology-anatomy prototype space that provides lesion-related prototypes and anatomy-aware normal references for comparing suspicious patches against both pathological and normal anatomical patterns. To avoid blindly amplifying normal tissues, **EasyTag** selects lesion-relevant patches through counterfactual prototype reasoning. To counteract the dilution of subtle lesion cues in global image representations, **EasyAmplifier** strengthens the selected lesion-relevant patch representations through morphology-guided residual enhancement, thereby increasing their contribution to the global image embedding. Experiments on multiple medical image datasets and frozen medical VLM backbones show that EasyLens consistently improves subtle-lesion detection and outperforms existing encoder-enhancement baselines without model fine-tuning. Code is available at: https://github.com/wanghao1118/EasyLens

1 Introduction

Medical vision-language models (VLMs) are increasingly explored for clinical image interpretation, including lesion detection and report generation [10, 20]. As diagnostic decision-support tools, their practical value depends on reliable recognition across a broad spectrum of lesion appearances, rather than only on highly visible abnormalities [5]. Here, salient lesions refer to abnormalities with large

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

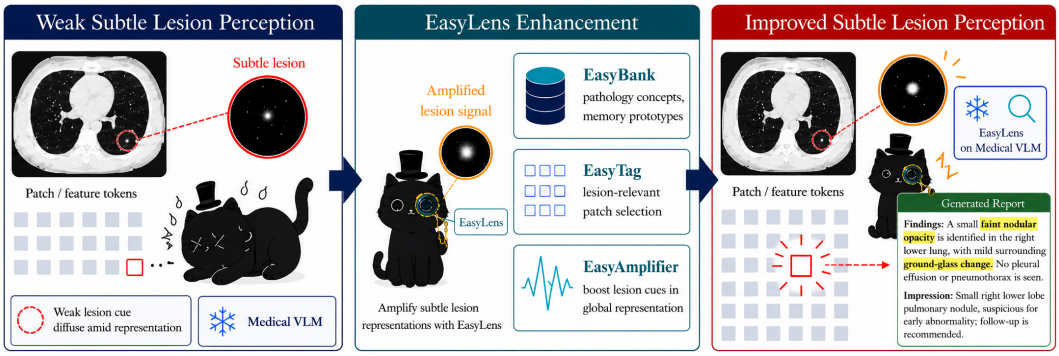


Figure 1: **EasyLens** brings weak subtle-lesion cues into focus in frozen medical VLMs.

spatial extent, high contrast, or pronounced morphological signatures, whereas subtle lesions exhibit sparse, low-contrast, or weakly distinguishable visual cues embedded within complex anatomical context [12, 18]. Although salient lesions can often be captured by strong visual patterns, subtle lesions provide weak local evidence that can be easily confused with normal anatomical variation. As local visual tokens are aggregated into global image representations, these weak lesion cues may become underrepresented, causing medical VLMs to miss or under-recognize subtle abnormalities [22]. Therefore, improving the sensitivity of medical VLMs to subtle lesions is critical for making them more reliable in clinical image interpretation [29].

To improve the image interpretation capabilities of medical VLMs, existing approaches have primarily focused on enhancing their visual encoders to extract richer semantic features from medical images and better align image representations with clinical text [30, 1]. Initially, many approaches pre-trained vision encoders to bridge the domain gap on large-scale paired medical images and reports [30], enabling them to capture pathological relevant semantic information when encoding medical images [24]. Although these pre-trained encoders effectively improve the interpretation capability of medical VLMs in the disease diagnosis, these coarse-grained building methods for pre-trained datasets make vision encoders lack the effective guidance to detect subtle lesions [18, 12].

To provide clinically relevant guidance and enhance the pathological semantics of medical image embeddings, subsequent studies introduced domain-specific modules into pretrained vision encoders in medical VLMs. To guide visual encoding with clinical semantics, several studies explicitly inject clinical semantic entities into the image encoding process. MedKLIP [25] extracts disease-related clinical entities and their spatial attributes from radiology reports to establish entity-patch correspondence, while KAD [27] constructs clinical entities and relations from reports and uses knowledge-guided disease queries to condition visual representation learning. Complementary to clinical-entity guidance, other studies enhance pathological semantics by refining patch-level visual representations. MLIP [15] improves local image-text alignment through masked local representation learning, while AdaMatch [2] uses adaptive patch matching to capture abnormalities with varying sizes and locations. PLACE [23] further enriches fine-grained visual details through pathological-level alignment and patch correlation modeling.

Although these encoder-enhancement methods improve the detection of subtle lesions, their general applicability to current medical VLMs remains limited. First, clinical-semantic guidance is insufficient to cover the complex pathological variations in real clinical scenarios. Since lesions often present with diverse morphological patterns and imaging appearances, supervision based on predefined clinical semantics provides coarse guidance and fails to capture subtle pathological details, leading to confusion among visually similar abnormalities. Specifically, subtle lesions occupy a limited portion of image patches, making fine-grained pathological detail preservation essential for accurate detection. Second, representation-level enhancement typically relies on extensive training or re-training, leading to considerable computational and time costs. In addition, adapting the encoder to specific pathological patterns reduce its sensitivity to other abnormalities, thereby limiting the applicability of medical VLMs across diverse clinical scenarios.

To address these limitations, we propose **EasyLens**, a training-free plug-and-play subtle-lesion representation amplifier for medical VLMs. EasyLens is designed to strengthen weak lesion cues that are preserved in frozen visual representations but become underrepresented during global image aggregation. It first constructs **EasyBank**, a pathology-anatomy prototype space that provides

lesion-related prototypes and anatomy-aware normal references for patch-level comparison. By contrasting suspicious patches with both pathological prototypes and normal anatomical references, EasyBank supports fine-grained discrimination between subtle lesion evidence and normal anatomical variation. Built upon EasyBank, **EasyTag** selects lesion-relevant patches through counterfactual prototype reasoning, thereby avoiding blind amplification of normal tissues. **EasyAmplifier** then strengthens the selected lesion-relevant patch representations through morphology-guided residual enhancement, increasing their contribution to the global image embedding while preserving the original visual context. Both modules operate without updating model parameters or requiring lesion annotations at inference time, making EasyLens applicable to frozen medical VLMs. We validate EasyLens on a unified subtle-lesion benchmark built from ReXGroundingCT, LIDC-IDRI, and AbdomenAtlas 3.0 Mini. Experiments across multiple frozen medical VLM backbones show that EasyLens consistently improves subtle-lesion detection and report generation, and outperforms existing encoder-enhancement baselines without model fine-tuning. Our contributions are summarized as follows:

(1) We propose **EasyLens**, a training-free plug-and-play amplifier that improves subtle-lesion recognition in frozen medical VLMs by exploiting latent pathological evidence through prototype-based reasoning.

(2) We construct **EasyBank**, a pathology-anatomy prototype space that organizes lesion-related prototypes and anatomy-aware normal references, and design **EasyTag**, a counterfactual prototype-guided patch selector that identifies lesion-relevant regions through fine-grained pathological comparison.

(3) We introduce **EasyAmplifier**, a morphology-guided residual semantic amplifier that enhances disease-related morphological semantics in selected patch representations without model fine-tuning or inference-time lesion annotations.

2 Related Work

2.1 Medical VLMs in Radiology

Recent medical VLMs in radiology have evolved from narrow image-to-report pipelines into more general systems for multi-task interpretation, interactive querying, visual grounding, and structured reasoning. On 2D chest radiographs, recent studies have improved clinical usability by introducing agentic tool use, anatomy-centric reasoning, fine-grained vision-language alignment, and pixel-grounded interaction. For example, MedRAX integrates multimodal tools and large models for complex chest X-ray interpretation [6], AOR performs anatomy-centric region-level reasoning [13], and RadZero strengthens fine-grained alignment for zero-shot classification, grounding, and segmentation [19]. MIMO further extends medical VLMs beyond text-only responses by supporting visual referring inputs and pixel-grounded outputs [3]. These works reflect a shift from holistic report generation toward more interactive and evidence-grounded radiological interpretation.

Another major trend extends radiology VLMs from 2D radiographs to volumetric understanding. Argus studies large-scale 3D CT report generation and highlights the importance of vision encoder pretraining, visual token compression, and model/data scaling for high-resolution 3D radiology VLMs [14]. BTB3D further shows that effective volumetric tokenization is more critical than simply enlarging language backbones for scalable 3D medical VLMs [9]. Recent public systems continue this direction through organ-separated CT-language modeling, variable-length 3D visual tokens, and native CT/MRI interpretation [26, 8, 21]. In parallel, clinically grounded radiology VLMs incorporate step-by-step verification, reinforcement learning, workflow-level context, or radiologist gaze to align model reasoning with expert diagnostic procedures [7, 28, 16, 11]. Despite these advances, most radiology VLMs still primarily optimize global image-report alignment, report generation, VQA, or coarse region-level grounding. Sparse subtle-lesion cues can therefore be weakened by holistic diagnostic semantics or compressed visual tokens, leaving a gap between strong global radiology understanding and robust perception of subtle lesion-carrying patches.

2.2 Subtle Lesion Detection in Medical VLMs

Subtle lesions remain challenging for medical VLMs because their diagnostic cues are weak, spatially sparse, and often overwhelmed by surrounding anatomical structures. To improve the perception of such lesions, existing studies have enhanced medical visual representations by injecting clinically

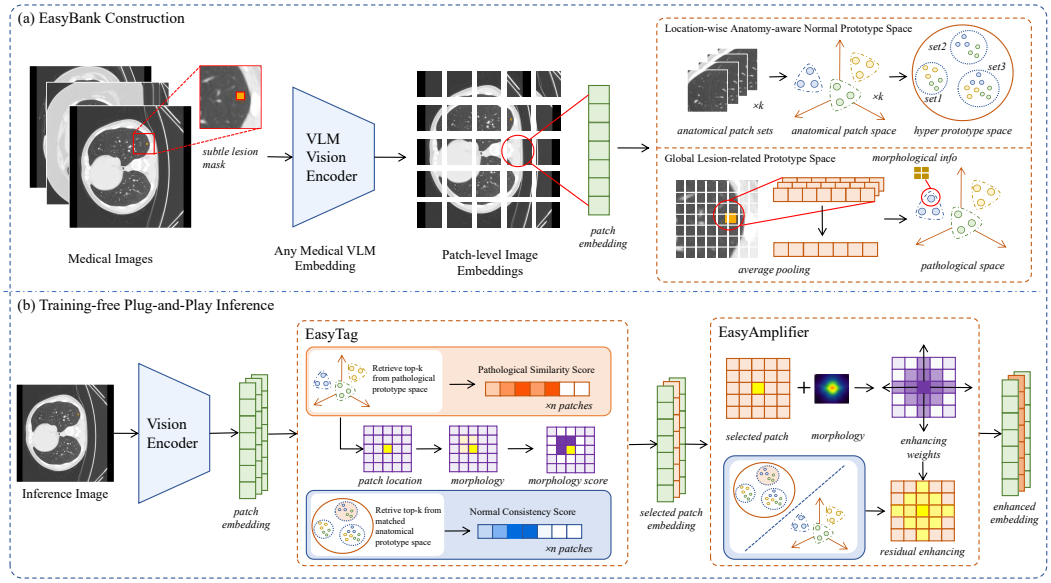


Figure 2: Overview of **EasyLens**. **(a) EasyBank Construction** builds an offline prototype space from CT images and lesion masks. **(b) Medical VLM Inference with EasyLens** selects lesion-relevant patches and amplifies their visual representations before feeding them into a frozen medical VLM for lesion-aware outputs.

relevant guidance into pretrained vision encoders. MedKLIP [25] extracts disease-related clinical entities and their spatial attributes from radiology reports to establish entity-patch correspondence, while KAD [27] constructs clinical entities and relations from reports and uses knowledge-guided disease queries to condition visual representation learning. Beyond clinical-entity guidance, other studies refine patch-level pathological semantics to better capture local abnormalities. MLIP [15] improves local image-text alignment through masked local representation learning, AdaMatch [2] uses adaptive patch matching to capture abnormalities with varying sizes and locations, and PLACE [23] further enriches fine-grained visual details through pathological-level alignment and patch correlation modeling.

Another related direction strengthens subtle lesion perception through region-aware grounding and abnormality-sensitive tuning. VividMed [17] and MIMO [3] extend medical VLMs with segmentation, referring, and pixel-grounded outputs, enabling pathological findings to be associated with specific image regions. Reg2RG [4] incorporates region-guided referring and grounding into CT report generation, while UMed-LVLM [31] and MMedPO[32] improve abnormal-region sensitivity through abnormal-aware fine-tuning or clinical-aware preference optimization. These studies push medical VLMs from coarse image-level diagnosis toward more localized subtle lesion understanding. However, most existing methods rely on predefined clinical semantics, explicit grounding modules, region-level supervision, or additional fine-tuning, making them costly and less directly applicable to frozen advanced medical VLMs. In contrast, our work identifies lesion-carrying patch representations and amplifies their pathological semantics at inference time, improving subtle lesion sensitivity without introducing a new detector or retraining the model.

3 Methodology

3.1 Overview

As shown in Appendix Fig. 4, our preliminary analysis suggests that subtle-lesion cues are not entirely absent from frozen medical VLMs. Instead, they can be partially preserved in patch-level visual representations but become underrepresented when local visual tokens are aggregated into global image representations. Since subtle lesions often occupy small regions, exhibit low contrast, or appear as weakly distinguishable cues within complex anatomical context, their representations can be easily mixed with normal anatomical patterns. This observation motivates us to strengthen lesion-relevant

patch representations directly in the frozen visual embedding space, rather than updating the vision encoder through additional training.

To this end, we propose **EasyLens**, a training-free plug-and-play subtle-lesion representation amplifier for medical VLMs. As illustrated in Fig. 2, EasyLens contains three components. First, **EasyBank** constructs a pathology-anatomy prototype space that stores lesion-related prototypes and anatomy-aware normal references. Second, **EasyTag** performs counterfactual prototype reasoning to select lesion-relevant patches by contrasting pathological similarity with location-matched normal consistency. Third, **EasyAmplifier** applies morphology-guided residual enhancement to the selected patch representations, increasing their contribution to the global image embedding while preserving the original visual context. The enhanced visual sequence is then passed to the subsequent components of the frozen medical VLM for downstream lesion-related tasks.

3.2 EasyBank: Pathology-Anatomy Prototype Space

EasyBank provides non-parametric visual references for distinguishing subtle lesion evidence from normal anatomical variation. Given a prototype construction set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{m}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ denotes a medical image and $\mathbf{m}_i$ denotes its lesion mask, we extract patch-level visual representations from the frozen vision encoder:

$$\mathbf{Z}_i = \mathcal{E}_v(\mathbf{x}_i) = [\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^P] \in \mathbb{R}^{P \times d}, \tag{1}$$

where P is the number of image patches and d is the hidden dimension. Here, $\mathcal{E}_v$ denotes the frozen visual-encoder prefix up to the chosen injection layer. EasyBank construction and inference use this same feature interface; the remaining frozen visual layers process the enhanced sequence. The lesion masks are used only for constructing EasyBank and are not required during inference.

We project each lesion mask onto the patch grid and divide patches into lesion-related and normal anatomical sets:

$$r_i^p = \frac{1}{|\Omega_p|} \sum_{\mathbf{u} \in \Omega_p} \mathbf{m}_i(\mathbf{u}), \quad \mathcal{P}_i = \{p \mid r_i^p > \tau\}, \quad \mathcal{N}_i = \{1, \ldots, P\} \setminus \mathcal{P}_i, \tag{2}$$

where Ω_p denotes the image region corresponding to the p-th patch, r_i^p is the lesion occupancy ratio, and τ is the occupancy threshold.

EasyBank contains two complementary reference spaces. The first is a global lesion-related prototype space $\mathcal{C}^L$, which summarizes recurring pathological patterns. Since directly clustering all lesion patches would bias the prototype space toward large lesions, we first aggregate lesion-related patches within each lesion-containing image:

$$\mathbf{h}_i^L = \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \mathbf{z}_i^p, \quad |\mathcal{P}_i| > 0. \tag{3}$$

The normalized image-level lesion representations are then clustered into K_L lesion-related prototypes:

$$\mathcal{C}^L = \mathrm{Cluster}\left(\{\hat{\mathbf{h}}_i^L \mid |\mathcal{P}_i| > 0\}, K_L\right). \tag{4}$$

This design assigns equal weight to each lesion-containing image during prototype construction, preventing large lesions from dominating the prototype space simply because they occupy more patches.

The second reference space consists of location-wise anatomy-aware normal prototypes. Normal anatomical appearances vary substantially across spatial locations; for example, normal lung parenchyma, mediastinum, pleura, and abdominal organs may have very different visual representations. Therefore, using a single global normal prototype space would mix heterogeneous normal structures and provide ambiguous counterfactual references. EasyBank instead constructs a normal prototype subspace for each patch location:

$$\mathcal{C}_p^A = \mathrm{Cluster}\left(\{\hat{\mathbf{z}}_i^p \mid p \in \mathcal{N}_i,\ i = 1, \ldots, N\}, K_A\right), \quad p = 1, \ldots, P. \tag{5}$$

Each $\mathcal{C}_p^A$ summarizes normal anatomical appearances at the same patch location and provides a location-matched reference for later counterfactual comparison.

In addition to prototype centers, EasyBank stores a lesion-support memory $\mathcal{L}_k$ and a morphology prior $\mathbf{M}_k$ for each lesion-related prototype. The support memory provides lesion-related reference embeddings for residual enhancement, while the morphology prior describes the spatial coherence of lesion patterns associated with the prototype. The final EasyBank is summarized as:

$$\mathcal{B} = \left\{ \mathcal{C}^L, \{\mathcal{C}_p^A\}_{p=1}^P, \{\mathcal{L}_k, \mathbf{M}_k\}_{k=1}^{K_L} \right\}. \tag{6}$$

Detailed clustering objectives, support-memory construction, and morphology-prior estimation are provided in Appendix Sec. A.2.

3.3 EasyTag: Counterfactual Lesion-Relevant Patch Selection

Given an inference image, EasyTag selects patches that are likely to contain subtle lesion evidence. For each patch representation $\mathbf{z}^p$, EasyTag compares it with two types of references in EasyBank: the global lesion-related prototypes $\mathcal{C}^L$ and the anatomy-aware normal prototypes $\mathcal{C}_p^A$ at the same patch location. The former measures whether the patch resembles pathological patterns, while the latter evaluates whether the patch can be explained by normal anatomy at the corresponding location.

Specifically, EasyTag retrieves the top-M nearest lesion-related prototypes and top-M nearest anatomy-aware normal prototypes for each patch, and computes the lesion similarity score s_p^L and normal consistency score s_p^A. The counterfactual lesion relevance score is defined as:

$$a_p = \sigma\left(\frac{s_p^L - s_p^A}{\tau_c} \right), \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function and τ_c is a temperature parameter. A high score indicates that the patch is close to lesion-related prototypes but poorly explained by its location-matched normal references. In this sense, EasyTag implements counterfactual reasoning: it asks whether a suspicious patch still appears abnormal after being compared with normal anatomical appearances from the same location.

Patch-wise scores may be noisy when lesion evidence is weak. Moreover, subtle lesions often appear as spatially coherent local structures rather than isolated patches. To include weak but morphology-consistent lesion cues, EasyTag calibrates the initial scores using morphology priors stored in EasyBank. We first select a high-confidence seed set $\mathcal{S}_0$ based on $\{a_p\}_{p=1}^P$. For each seed patch p, we identify its nearest lesion-related prototype:

$$k^*(p) = \arg\max_k \mathrm{sim}(\mathbf{z}^p, \mathbf{c}_k^L). \tag{8}$$

The morphology prior associated with this prototype is then used to propagate confidence from seed patches to spatially coherent neighboring patches:

$$\tilde{a}_q = a_q + \lambda \max_{p \in \mathcal{S}_0} \left[a_p \mathbf{M}_{k^*(p)}(q-p) \right], \quad \mathcal{S}^C = \mathrm{TopK}_q\, \tilde{a}_q, \tag{9}$$

where λ controls the calibration strength, $\mathbf{M}_{k^*(p)}(q-p)$ denotes the morphology-prior value at the relative offset from seed patch p to patch q, and $\mathcal{S}^C$ is the final candidate set. This calibration allows EasyTag to select sparse but spatially coherent lesion evidence while avoiding blind amplification of normal tissues. The detailed retrieval procedure and score computation are provided in Appendix Sec. A.3.

3.4 EasyAmplifier: Morphology-Guided Residual Enhancement

After EasyTag selects lesion-relevant candidate patches, EasyAmplifier strengthens their representations before they are passed to downstream VLM components. The goal is not to replace the original visual embeddings, but to inject lesion-related residual directions into patches that are supported by both counterfactual evidence and morphology priors. This design preserves the anatomical context encoded by the frozen vision encoder while increasing the contribution of subtle lesion cues to the global image embedding.

For each selected candidate patch $p \in \mathcal{S}^C$, EasyAmplifier retrieves a lesion-related reference from the support memory $\mathcal{L}_{k^*(p)}$ associated with its recalled lesion prototype. This reference provides a

Models	ReX			LIDC			Abdomen		
	Stat.	**Sel.**	**Gen.**	**Stat.**	**Sel.**	**Gen.**	**Stat.**	**Sel.**	**Gen.**
LLaVA-Med	0.00	1.11	3.93	0.00	31.58	33.92	1.16	14.42	36.60
RadFM	46.94	7.78	2.96	0.00	21.05	13.75	10.47	11.54	9.60
Lingshu	7.48	25.56	0.33	20.45	33.83	16.20	37.21	53.85	25.04
Hulu-Med	64.63	27.78	0.88	28.79	30.83	23.87	51.16	52.88	39.89
MedGemma	40.82	24.44	4.55	25.00	21.05	35.94	27.91	45.19	26.11
MedGemma1.5	42.86	23.33	4.41	20.45	27.82	41.93	15.12	49.04	38.18
MedKLIP	0.00	17.78	-	0.00	27.07	-	0.00	28.85	-
KAD	0.00	13.33	-	0.00	16.54	-	0.00	5.77	-
EasyLens	**66.67**	**31.11**	**5.15**	**30.30**	**36.09**	**45.86**	**52.33**	**55.77**	**40.67**

Table 1: Comparison on ReX, LIDC, and Abdomen. Stat. denotes regional counting, Sel. denotes region selection, and Gen. denotes the lesion-token probability probe defined in Appendix B.2.

prototype-consistent pathological direction in the frozen embedding space. To extend enhancement beyond isolated high-confidence patches, EasyAmplifier computes, for every patch q, the strongest morphology-consistent support from the selected candidates:

$$p^*(q) = \arg\max_{p \in \mathcal{S}^C} \tilde{a}_p \mathbf{M}_{k^*(p)}(q-p), \quad w_q = \max_{p \in \mathcal{S}^C} \tilde{a}_p \mathbf{M}_{k^*(p)}(q-p). \tag{10}$$

Here, $p^*(q)$ identifies the selected candidate that provides the strongest morphology-supported evidence for patch q, and w_q measures the strength of this support. If w_q is sufficiently large, patch q is considered part of a morphology-consistent lesion region.

EasyAmplifier then retrieves a lesion-related reference embedding $\mathbf{r}_q$ from the support memory of the strongest recalled prototype and updates the patch representation through a score-weighted residual enhancement:

$$\bar{\mathbf{z}}^q = \begin{cases} \mathbf{z}^q + \alpha w_q(\mathbf{r}_q - \mathbf{z}^q), & w_q > \eta, \\ \mathbf{z}^q, & w_q \le \eta, \end{cases} \tag{11}$$

where α controls the amplification strength and η prevents low-confidence patches from being modified. The residual direction $(\mathbf{r}_q - \mathbf{z}^q)$ moves the patch toward a lesion-related reference while retaining its original embedding as the base representation. Thus, high-confidence and morphology-supported lesion cues are enhanced more strongly, whereas unrelated anatomical regions remain unchanged.

Finally, EasyAmplifier outputs the enhanced visual sequence $\bar{\mathbf{Z}} = [\bar{\mathbf{z}}^1, \ldots, \bar{\mathbf{z}}^P]$, which replaces the original patch sequence before being passed to the subsequent medical VLM components. Since EasyLens relies only on prototype retrieval, counterfactual scoring, morphology-guided weighting, and residual enhancement, it requires no gradient-based optimization, no inference-time lesion annotations, and no modification of the frozen VLM.

4 Experiments

4.1 Datasets and Evaluation Tasks

We evaluate EasyLens on two dataset groups spanning the lesion spectrum described in the introduction. The first group focuses on *subtle lesions*, for which we construct a unified benchmark from ReXGroundingCT, LIDC-IDRI, and AbdomenAtlas 3.0 Mini, denoted as ReX, LIDC, and Abdomen, respectively. This benchmark contains three task types: regional statistics (Stat.), region selection (Sel.), and lesion-aware report generation (Gen.). The second group contains more salient or regular lesion datasets, including MIMIC-CXR, Kvasir-SEG, and BKAI-Polyp, where we evaluate standard medical report generation using BLEU-1, BLEU-4, METEOR, and ROUGE-L. These datasets verify that EasyLens does not over-specialize to subtle abnormalities and preserves general report-generation ability when lesions are visually more apparent. Detailed dataset construction, task definitions, evaluation metrics, and benchmark distributions are provided in Appendix Sec. C.1.

Models	w/ EasyLens	ReX				Kvasir-SEG			
		Stat.	Sel.	Gen.	Inf.(s)	BL-1	MTR	RG-L	Inf.(s)
LLaVA-Med	✗	0.00	1.11	3.93	1.03	11.49	22.69	**8.61**	3.33
	✓	0.00	1.11	**32.37**	1.13	**12.22**	**24.35**	8.56	2.80
RadFM	✗	46.94	7.78	2.96	2.68	1.60	1.31	2.41	0.57
	✓	**48.98**	**10.00**	**4.12**	2.75	**2.28**	**1.44**	**2.61**	0.49
Lingshu	✗	7.48	25.56	0.33	0.46	20.06	10.81	20.38	0.96
	✓	7.48	**27.78**	**8.51**	0.49	**26.69**	**13.89**	**30.00**	0.63
MedGemma	✗	40.82	24.44	4.55	0.78	**13.99**	25.38	15.59	1.43
	✓	**46.26**	**25.56**	**6.73**	1.48	13.81	**26.00**	**16.19**	1.81
MedGemma1.5	✗	42.86	23.33	4.41	1.80	9.56	15.87	15.39	1.87
	✓	**66.67**	**31.11**	**5.15**	2.38	**10.30**	**16.55**	**16.18**	2.01

Table 2: Comparison of different medical VLMs with and without the proposed module on ReX and Kvasir-SEG datasets.

4.2 Implementation and Experimental Setup

Unless otherwise specified, EasyLens denotes MedGemma1.5 equipped with the proposed frozen-backbone inference-time adapter in Table 1, while Table 2 applies the same adapter to each medical VLM backbone. EasyLens keeps the visual encoder, multimodal projector, and language decoder fixed, and only calibrates three interface parameters, i.e., residual strength β, candidate-token budget k, and seed budget s, as detailed in Appendix B.1. We report subtle-lesion results in Table 1, backbone transfer in Table 2, general and salient-lesion report generation in Appendix C.3, and interface ablations in Appendix C.6. For report-generation tasks, we use raw benchmark prompts, a 160-token generation limit, and the evaluation-only lesion-aware probe described in Appendix B.2.

4.3 Quantitative Evaluation

EasyLens yields consistent improvements across all subtle-lesion evaluation settings. Table 1 reports results on ReX, LIDC, and Abdomen, covering regional counting, lesion-aware region selection, and the lesion-token probability probe. MedGemma1.5 with EasyLens achieves the highest scores in all nine dataset–task combinations among the baselines listed in Table 1. Relative to the strongest competing baseline in each setting, EasyLens improves ReX by 2.04, 3.33, and 0.60 points on Stat., Sel., and Gen., respectively. The corresponding gains are 1.51, 2.26, and 3.93 points on LIDC, and 1.17, 1.92, and 0.78 points on Abdomen. These results indicate that EasyLens improves the recognition of small, low-contrast, and spatially sparse lesions that are difficult for existing medical VLMs to capture reliably.

EasyLens transfers effectively across frozen medical VLM backbones. Table 2 further evaluates EasyLens when attached to different frozen backbones, including LLaVA-Med, RadFM, Lingshu, MedGemma, and MedGemma1.5. On ReX, EasyLens improves the lesion-token probe score for all five backbones, with particularly large gains for LLaVA-Med and Lingshu, increasing Gen. from 3.93 to 32.37 and from 0.33 to 8.51, respectively. For the strongest backbone, MedGemma1.5, EasyLens improves Stat./Sel./Gen. from 42.86/23.33/4.41 to 66.67/31.11/5.15. These results show that the proposed module is not tied to a specific backbone architecture.

EasyLens preserves general report-generation capability beyond subtle-lesion tasks. As shown in Appendix C.3, EasyLens remains comparable to MedGemma1.5 on MIMIC-CXR and improves performance on Kvasir-SEG and BKAI-Polyp across all non-zero lexical metrics. Together with the backbone-transfer results in Table 2, these findings suggest that amplifying lesion-relevant visual evidence does not substantially disrupt the original reporting behavior of the frozen VLM, while providing consistent benefits for subtle-lesion perception.

4.4 Ablation Study

EasyTag provides training-free selection competitive with supervised alternatives. As shown in Table 3, EasyTag improves MedGemma1.5 on ReX, increasing Stat. from 42.86 to 66.67, Sel.

EasyTag Selector Ablation				EasyAmplifier Amplifier Ablation			
Models	**ReX**			**Models**	**ReX**		
	Stat.	**Sel.**	**Gen.**		**Stat.**	**Sel.**	**Gen.**
MedGemma1.5	42.86	23.33	4.41	MedGemma1.5	42.86	23.33	4.41
w/ MLP Selector	69.39	27.78	4.13	w/o Residual	64.63	**33.33**	4.89
w/ GT Masks	**71.43**	**32.22**	**5.36**	w/o Morphology	57.14	28.89	3.17
w/ EasyTag	66.67	31.11	5.15	w/ EasyAmplifier	**66.67**	31.11	**5.15**

Table 3: Ablation studies of the EasyTag and EasyAmplifier on the ReX dataset.

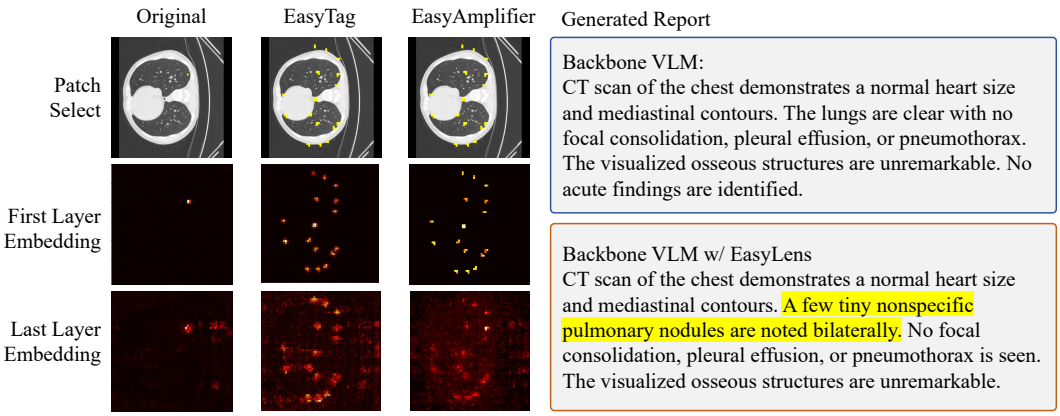


Figure 3: Case study of EasyLens on subtle-lesion perception.

from 23.33 to 31.11, and Gen. from 4.41 to 5.15. The MLP Selector is a trainable patch-scoring module supervised to identify lesion-relevant tokens. Although it obtains a slightly higher Stat. score of 69.39, its Sel. and Gen. scores decrease to 27.78 and 4.13. This comparison shows that EasyTag achieves competitive performance without training, while providing stronger lesion-aware evidence for selection and generation. The GT-mask setting selects lesion-overlapping tokens using ground-truth masks and serves only as a mask-guided upper-reference. Its results, 71.43 on Stat., 32.22 on Sel., and 5.36 on Gen., are close to EasyTag, further indicating that EasyTag recovers most useful lesion evidence without inference-time masks.

EasyAmplifier benefits from residual preservation and morphology-aware enhancement. Table 3 also evaluates the amplification design. Removing the residual formulation reduces Stat. from 66.67 to 64.63 and Gen. from 5.15 to 4.89, while Sel. increases from 31.11 to 33.33. Removing morphology modeling causes a broader drop, reducing Stat. to 57.14, Sel. to 28.89, and Gen. to 3.17. These results suggest that morphology-aware enhancement is the main source of subtle-lesion amplification, while the residual path helps preserve the original visual semantics. Additional interface-level ablations are provided in Appendix C.6.

4.5 Case Study

Qualitative evidence for lesion-relevant token selection and amplification. Figure 3 provides a qualitative example of a CT case containing a subcentimeter pulmonary nodule. The selected tokens produced by EasyTag overlap with the suspected nodule region, supporting the claim that the training-free selector can identify micro-lesion-related visual evidence from frozen representations. After EasyAmplifier, the response associated with the selected lesion patches becomes stronger and more spatially concentrated on the nodule region, indicating that morphology-guided residual enhancement increases the lesion-semantic contribution of the relevant patches. This visual change is consistent with the generated report, where the EasyLens-enhanced model mentions the tiny pulmonary nodule, while the baseline model omits this finding.

5 Conclusion

We presented **EasyLens**, a training-free plug-and-play amplifier that improves subtle-lesion perception in frozen medical VLMs. EasyLens builds a pathology-anatomy prototype space, selects lesion-

relevant tokens through counterfactual prototype reasoning, and strengthens them with morphology-guided residual enhancement. Experiments across subtle-lesion benchmarks and multiple VLM backbones show consistent gains in regional counting, region selection, and lesion-aware report generation without updating model parameters. These results show that **EasyLens** can expose and amplify weak lesion evidence already encoded in frozen visual tokens, enabling medical VLMs to better recognize and report clinically important micro-lesions without model-specific retraining.

Acknowledgments

The authors gratefully acknowledge the financial support through the National Key Research and Development Program of China (2023YFC2411300), the National Natural Science Foundation of China (82595964), the Fundamental Research Funds for the Central Universities (YG2025ZD26, JC202622).

References

[1] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

[2] Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9494–9509, 2024.

[3] Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xueping Wang, Meikang Qiu, and Hang Li. Mimo: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24732–24741, 2025.

[4] 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*, 2025.

[5] Zijie Cheng, Ariel Yuhan Ong, Siegfried K Wagner, David A Merle, Lie Ju, Hanyuan Zhang, Ruinian Chen, Linze Pang, Boxuan Li, Tiantian He, et al. Understanding the robustness of vision-language models to medical image artefacts. *NPJ Digital Medicine*, 8(1):727, 2025.

[6] Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. MedRAX: Medical reasoning agent for chest x-ray. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 15661–15676. PMLR, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/fallahpour25a.html`.

[7] Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. *arXiv preprint arXiv:2504.20930*, 2025.

[8] Chengyu Fang, Heng Guo, Zheng Jiang, Chunming He, Xiu Li, and Minfeng Xu. Photon: Speedup volume understanding with efficient multimodal large language models. *arXiv preprint arXiv:2603.25155*, 2026.

[9] Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better tokens for better 3d: Advancing vision-language modeling in 3d medical imaging. *arXiv preprint arXiv:2510.20639*, 2025.

[10] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984, 2024.

[11] Kinhei Lee, Peiyuan Jing, Zhenxuan Zhang, Yue Yang, Tao Wang, Dominic C Marshall, Yingying Fang, and Guang Yang. Seeing through experts eyes a foundational vision language model trained on radiologists gaze and reasoning. *arXiv preprint arXiv:2604.14316*, 2026.

[12] Qingqiu Li, Xiaohan Yan, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, and Shujun Wang. Anatomical structure-guided medical vision-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 80–90. Springer, 2024.

[13] Qingqiu Li, Zihang Cui, Seongsu Bae, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, Junjun He, et al. Aor: Anatomical ontology-guided reasoning for medical large multimodal model in chest x-ray interpretation. *arXiv preprint arXiv:2505.02830*, 2025.

[14] Che Liu, Zhongwei Wan, Yuqi Wang, Hui Shen, Haozhe Wang, Kangyu Zheng, Mi Zhang, and Rossella Arcucci. Argus: benchmarking and enhancing vision-language models for 3d radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16448–16460, 2025.

[15] Jiarun Liu, Hong-Yu Zhou, Cheng Li, Weijian Huang, Hao Yang, Yong Liang, Guangming Shi, Hairong Zheng, and Shanshan Wang. Mlip: medical language-image pre-training with masked local representation learning. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[16] Kang Liu, Zhuoqi Ma, Siyu Liang, Yunan Li, Xiyue Gao, Chao Liang, Kun Xie, and Qiguang Miao. Seeing like radiologists: Context-and gaze-guided vision-language pretraining for chest x-rays. *arXiv preprint arXiv:2603.26049*, 2026.

[17] Lingxiao Luo, Bingda Tang, Xuanzhong Chen, Rong Han, and Ting Chen. Vividmed: Vision language model with versatile visual grounding for medicine. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1800–1821, 2025.

[18] Yifan Mao, Fei-Fei Xue, Ruixuan Wang, Jianguo Zhang, Wei-Shi Zheng, and Hongmei Liu. Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–538. Springer, 2020.

[19] Jonggwon Park, Soobum Kim, Byungmu Yoon, and Kyoyun Choi. Radzero: Similarity-based cross-attention for explainable vision-language alignment in radiology with zero-shot multi-task capability. *arXiv e-prints*, pages arXiv–2504, 2025.

[20] Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. *arXiv preprint arXiv:2402.09262*, 2024.

[21] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[22] Matthew Tivnan, Tzu-Cheng Lee, Ruoqiao Zhang, Kirsten Boedeker, Liang Cai, Jeremias Sulam, and J Webster Stayman. Task-driven ct image quality optimization for low-contrast lesion detectability with tunable neural networks. In *Medical Imaging 2023: Physics of Medical Imaging*, volume 12463, pages 338–343. SPIE, 2023.

[23] Jun Wang, Lixing Zhu, Xiaohan Yu, Abhir Bhalerao, and Yulan He. Improving medical visual representation learning with pathological-level cross-modal alignment and correlation exploration. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[24] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022.

[25] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21372–21383, 2023.

[26] Kohei Yamamoto and Tomohiro Kikuchi. Totalfm: An organ-separated framework for 3d-ct vision foundation models. *arXiv preprint arXiv:2601.00260*, 2026.

[27] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023.

[28] Yabin Zhang, Chong Wang, Yunhe Gao, Jiaming Liu, Maya Varma, Justin Xu, Sophie Ostmeier, Jin Long, Sergios Gatidis, Seena Dehkharghani, et al. A reasoning-enabled vision-language foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2604.00493*, 2026.

[29] Yi Zhang, Yiji Mao, Xuanyu Lu, Xingyu Zou, Hao Huang, Xinyang Li, Jiayue Li, and Haixian Zhang. From single to universal: tiny lesion detection in medical imaging. *Artificial Intelligence Review*, 57(8):192, 2024.

[30] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR, 2022.

[31] Yucheng Zhou, Lingran Song, and Jianbing Shen. Improving medical large vision-language models with abnormal-aware feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12994–13011, 2025.

[32] Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. Mmedpo: Aligning medical vision-language models with clinical-aware multimodal preference optimization. *arXiv preprint arXiv:2412.06141*, 2024.

A Additional Method Details

A.1 Preliminary Analysis of Representation Dilution

Appendix Fig. 4 provides the empirical motivation for EasyLens. The analysis shows that subtle-lesion cues are not completely absent in frozen medical VLM representations. Instead, their patch-level evidence can be weakly encoded but becomes less separable from normal anatomical regions during visual-token aggregation. This observation motivates a training-free representation enhancement strategy that operates directly on patch-level visual embeddings.

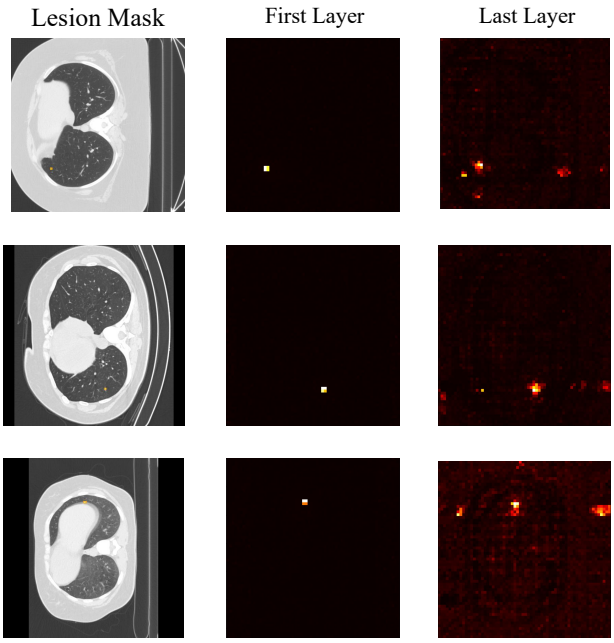


Figure 4: Subtle-lesion representation dilution in medical VLMs.

A.2 Detailed Construction of EasyBank

EasyBank is a non-parametric pathology-anatomy prototype space constructed from patch-level representations extracted by a frozen medical VLM. Given a prototype construction set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{m}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ is a medical image and $\mathbf{m}_i \in \{0,1\}^{H\times W}$ is its lesion mask, we extract visual patch representations:

$$\mathbf{Z}_i = \mathcal{E}_v(\mathbf{x}_i) = [\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^P] \in \mathbb{R}^{P\times d}. \tag{12}$$

Patch-level lesion assignment. Let Ω_p denote the image region covered by the p-th visual patch. We project the lesion mask onto the patch grid by computing the lesion occupancy ratio:

$$r_i^p = \frac{1}{|\Omega_p|} \sum_{\mathbf{u}\in\Omega_p} \mathbf{m}_i(\mathbf{u}). \tag{13}$$

Patches are divided into lesion-related and normal anatomical regions:

$$\mathcal{P}_i = \{p \mid r_i^p > \tau\}, \quad \mathcal{N}_i = \{1, \ldots, P\} \setminus \mathcal{P}_i, \tag{14}$$

where τ is the lesion occupancy threshold.

Lesion-related prototype construction. Directly clustering all lesion patches would bias prototype construction toward salient lesions, since they contribute more patch embeddings. To reduce this

imbalance, we represent each lesion-containing image using the mean embedding of its lesion-related patches:

$$\mathbf{h}_i^L = \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \mathbf{z}_i^p, \quad |\mathcal{P}_i| > 0. \tag{15}$$

The normalized lesion representations are collected into:

$$\mathcal{B}^L = \{\hat{\mathbf{h}}_i^L \mid |\mathcal{P}_i| > 0,\ i = 1, \ldots, N\}. \tag{16}$$

We cluster this buffer into K_L lesion-related prototypes:

$$\mathcal{C}^L = \{\mathbf{c}_1^L, \mathbf{c}_2^L, \ldots, \mathbf{c}_{K_L}^L\}. \tag{17}$$

The clustering objective is:

$$\min_{\{\mathbf{c}_k^L\}_{k=1}^{K_L}} \sum_{\mathbf{h} \in \mathcal{B}^L} \min_k \left\|\mathbf{h} - \hat{\mathbf{c}}_k^L\right\|_2^2. \tag{18}$$

Each lesion-related prototype summarizes a recurring pathological representation pattern discovered from frozen VLM embeddings.

Lesion-support memory. For each lesion-related prototype, we retain a support memory containing patch-level lesion embeddings from construction samples assigned to that prototype. Let

$$\pi(i) = \arg\min_k \left\|\hat{\mathbf{h}}_i^L - \hat{\mathbf{c}}_k^L\right\|_2^2 \tag{19}$$

denote the prototype assignment of image i. The support memory for prototype k is:

$$\mathcal{L}_k = \{\hat{\mathbf{z}}_i^p \mid \pi(i) = k,\ p \in \mathcal{P}_i\}. \tag{20}$$

This memory provides lesion-related reference embeddings for EasyAmplifier without adding trainable parameters.

Morphology prior estimation. Subtle lesions are often spatially coherent rather than isolated at a single patch. For each lesion-related prototype, we estimate a morphology prior from the patch-level lesion supports of its assigned construction samples. For a relative patch offset δ, the prior is computed as:

$$\mathbf{M}_k(\delta) = \frac{1}{Z_k} \sum_{\pi(i)=k} \sum_{p \in \mathcal{P}_i} \mathbf{1}[p + \delta \in \mathcal{P}_i], \tag{21}$$

where $\mathbf{1}[\cdot]$ is the indicator function and Z_k is a normalization factor. We normalize $\mathbf{M}_k$ to $[0, 1]$. This prior estimates how likely neighboring patches are to belong to the same lesion pattern when a patch recalls prototype k.

Anatomy-aware normal prototype construction. Normal anatomical appearances vary substantially across spatial locations. Therefore, EasyBank constructs location-wise normal prototype subspaces rather than using a single global normal prototype set. For each patch location p, we collect normal embeddings from the same position:

$$\mathcal{B}_p^A = \{\hat{\mathbf{z}}_i^p \mid p \in \mathcal{N}_i,\ i = 1, \ldots, N\}. \tag{22}$$

Each location-specific buffer is clustered into K_A anatomy-aware normal prototypes:

$$\mathcal{C}_p^A = \{\mathbf{c}_{p,1}^A, \mathbf{c}_{p,2}^A, \ldots, \mathbf{c}_{p,K_A}^A\}. \tag{23}$$

The clustering objective is:

$$\min_{\{\mathbf{c}_{p,k}^A\}_{k=1}^{K_A}} \sum_{\mathbf{z} \in \mathcal{B}_p^A} \min_k \left\|\mathbf{z} - \hat{\mathbf{c}}_{p,k}^A\right\|_2^2. \tag{24}$$

The final EasyBank is:

$$\mathcal{B} = \left\{\mathcal{C}^L, \{\mathcal{C}_p^A\}_{p=1}^P, \{\mathcal{L}_k, \mathbf{M}_k\}_{k=1}^{K_L}\right\}. \tag{25}$$

It provides lesion-related prototypes, anatomy-aware normal references, lesion-support memories, and morphology priors for EasyTag and EasyAmplifier.

A.3 Detailed Counterfactual Prototype Reasoning in EasyTag

Given an inference image, the frozen vision encoder produces patch-level representations:

$$\mathbf{Z} = [\mathbf{z}^1, \mathbf{z}^2, \ldots, \mathbf{z}^P]. \tag{26}$$

EasyTag evaluates each patch from two complementary perspectives. A lesion-relevant patch should be close to lesion-related prototypes in $\mathcal{C}^L$, while being poorly explained by anatomy-aware normal prototypes $\mathcal{C}_p^A$ from the same spatial location.

For patch $\mathbf{z}^p$, EasyTag retrieves its top-M nearest lesion-related prototypes and top-M nearest anatomy-aware normal prototypes:

$$\mathcal{R}_p^L = \mathrm{TopM}(\mathcal{C}^L, \mathbf{z}^p), \quad \mathcal{R}_p^A = \mathrm{TopM}(\mathcal{C}_p^A, \mathbf{z}^p). \tag{27}$$

The lesion similarity score and normal consistency score are:

$$s_p^L = \frac{1}{M} \sum_{\mathbf{c} \in \mathcal{R}_p^L} \mathrm{sim}(\mathbf{z}^p, \mathbf{c}), \quad s_p^A = \frac{1}{M} \sum_{\mathbf{c} \in \mathcal{R}_p^A} \mathrm{sim}(\mathbf{z}^p, \mathbf{c}). \tag{28}$$

The initial counterfactual lesion relevance score is:

$$a_p = \sigma \left(\frac{s_p^L - s_p^A}{\tau_c} \right). \tag{29}$$

A normal patch is expected to have high normal consistency with $\mathcal{C}_p^A$, whereas a lesion-relevant patch should exhibit higher pathological similarity and lower consistency with its normal anatomical reference.

Morphology-guided score calibration. Patch-wise counterfactual scores may be noisy when lesion evidence is weak. To include spatially coherent but low-confidence lesion evidence, EasyTag calibrates scores using morphology priors. We first select high-confidence seed patches:

$$\mathcal{S}_0 = \mathrm{TopK}_0\{a_p\}_{p=1}^P. \tag{30}$$

For each seed patch p, its nearest lesion-related prototype is:

$$k^*(p) = \arg\max_k \mathrm{sim}(\mathbf{z}^p, \mathbf{c}_k^L). \tag{31}$$

The calibrated score for patch q is:

$$\tilde{a}_q = a_q + \lambda \max_{p \in \mathcal{S}_0} \left[a_p \mathbf{M}_{k^*(p)}(q - p) \right], \tag{32}$$

where λ controls the strength of morphology-guided calibration. The final candidate set is:

$$\mathcal{S}^C = \mathrm{TopK}_{q \in \{1,\ldots,P\}} \tilde{a}_q. \tag{33}$$

This procedure encourages EasyTag to select spatially coherent lesion regions rather than isolated high-scoring patches.

A.4 Detailed Morphology-Guided Residual Enhancement in EasyAmplifier

EasyAmplifier strengthens lesion-relevant representations selected by EasyTag. Although the selected patches may already contain lesion evidence, their original embeddings can still be weak due to representation dilution during visual-token aggregation. EasyAmplifier therefore enhances them through residual updates in the visual embedding space, rather than replacing the original features.

Candidate-level residual direction. For each selected candidate patch $p \in \mathcal{S}^C$, EasyTag provides its calibrated score $\tilde{a}_p$ and recalled lesion prototype $k^*(p)$. EasyAmplifier retrieves the closest lesion-related reference from the corresponding support memory:

$$\mathbf{r}_p = \arg\max_{\mathbf{u} \in \mathcal{L}_{k^*(p)}} \mathrm{sim}(\mathbf{z}^p, \mathbf{u}). \tag{34}$$

The residual direction is:

$$\Delta_p = \mathbf{r}_p - \mathbf{z}^p. \tag{35}$$

The candidate patch can be enhanced as:

$$\mathbf{z}^{p,+} = \mathbf{z}^p + \alpha \tilde{a}_p \Delta_p, \quad p \in \mathcal{S}^C, \tag{36}$$

where α controls the residual amplification strength. This update moves the candidate patch toward a lesion-related reference while preserving its original anatomical context.

Morphology-guided propagation. Enhancing only selected candidates may miss nearby patches that contain weaker but spatially consistent lesion evidence. EasyAmplifier therefore propagates residual enhancement according to the morphology prior recalled by each selected candidate. For every patch q, we compute:

$$p^*(q) = \arg\max_{p \in \mathcal{S}^C} \tilde{a}_p \mathbf{M}_{k^*(p)}(q-p), \quad w_q = \max_{p \in \mathcal{S}^C} \tilde{a}_p \mathbf{M}_{k^*(p)}(q-p). \tag{37}$$

When $w_q > \eta$, patch q is treated as morphologically supported by the selected lesion evidence. Its lesion-related reference is retrieved from the support memory of the strongest recalled prototype:

$$\mathbf{r}_q = \arg\max_{\mathbf{u} \in \mathcal{L}_{k^*(p^*(q))}} \mathrm{sim}(\mathbf{z}^q, \mathbf{u}). \tag{38}$$

The enhanced representation is:

$$\bar{\mathbf{z}}^q = \mathbf{z}^q + \alpha w_q(\mathbf{r}_q - \mathbf{z}^q), \quad w_q > \eta. \tag{39}$$

For patches with insufficient morphology support, EasyAmplifier keeps the original representation unchanged:

$$\bar{\mathbf{z}}^q = \mathbf{z}^q, \quad w_q \le \eta. \tag{40}$$

Enhanced visual sequence. The final enhanced patch sequence is:

$$\bar{\mathbf{Z}} = [\bar{\mathbf{z}}^1, \bar{\mathbf{z}}^2, \ldots, \bar{\mathbf{z}}^P]. \tag{41}$$

This sequence replaces the original visual embedding sequence before being passed to the subsequent medical VLM components. Since the enhancement uses only prototype retrieval, morphology-guided propagation, and residual injection, it requires no gradient-based optimization, no inference-time lesion annotations, and no modification of the pretrained model.

B Additional Implementation Details

B.1 Hyperparameter and Interface Calibration

EasyLens is designed as a plug-and-play adapter for frozen medical VLMs. Its portability comes from separating the shared lesion-enhancement mechanism from lightweight interface calibration. The shared mechanism is unchanged across datasets and backbones: EasyTag selects lesion-relevant visual tokens, and EasyAmplifier applies residual lesion-aware enhancement to morphology-supported tokens. The calibrated parameters only determine the operating point of this fixed mechanism.

Interface parameters. The interface parameters are β, k, and s. The residual strength β controls the magnitude of the EasyAmplifier update. The candidate-token budget k corresponds to K in the method, and the seed budget s corresponds to K_0. Seeds are the globally highest-scoring patches in Eq. (30). Support retrieval uses a single nearest neighbor within the recalled prototype memory, as defined in Eqs. (34) and (38); it is not an additional tunable budget.

Let $\mathbf{z}_i$ denote the original visual embedding of token i and w_i its morphology-supported weight. The candidate set supplies support to neighboring patches, so the final update applies to every sufficiently supported patch, as in Eqs. (39)–(40):

$$\bar{\mathbf{z}}_i = \begin{cases} \mathbf{z}_i + \beta w_i(\mathbf{r}_i - \mathbf{z}_i), & w_i > \eta, \\ \mathbf{z}_i, & w_i \le \eta, \end{cases}$$

where $\mathbf{r}_i$ is the nearest support reference from the prototype recalled by the strongest supporting candidate, and w_i is the morphology-supported weight defined in Appendix A.4. This residual formulation preserves the original visual representation while amplifying lesion-relevant evidence. In the method section, the residual coefficient is denoted by α; in experiments, we use β to denote its implementation value.

Operating regimes. We use three task-conditioned operating regimes. Counting tasks use a high-recall setting because missing a subtle lesion directly changes the answer. Selection tasks use a balanced precision–recall setting because excessive token amplification may introduce distractor regions. Report-generation tasks use a conservative setting because the decoder must preserve global context and language fluency.

For regular lesion report generation, the operating point is further calibrated according to lesion scale and morphology. Datasets with broader or more heterogeneous lesion regions require a larger token budget. For backbone adaptation, the same discrete grid is reused across VLMs, but the final operating point is calibrated to the visual-token geometry of each backbone. This calibration does not update any backbone parameters and does not change the EasyLens architecture.

Common settings. Unless otherwise specified, we use the common settings in Table 4. All selected configurations are fixed before evaluation and shared by all test samples in the corresponding dataset–task or backbone–evaluation cell. No per-instance hyperparameter selection is used.

Table 4: Common EasyLens settings used in the main experiments.

Setting	Value
Vision injection layer	1 (first vision block)
Candidate selection	global top-k by $\tilde{a}_q$
Seed selection	global top-s by a_p
Support reference	one nearest neighbor in recalled memory
Morphology calibration weight λ	1.0
Amplification gate	$w_q > \eta$
Max new tokens for count/select	96
Max new tokens for report generation	160
Report prompt mode	`benchmark_raw`
Report probe window	first 3 decoding steps

B.2 Lesion-aware Report-generation Probe

For subtle-lesion report generation, lexical-overlap metrics alone may not fully capture whether the model recognizes the target lesion. A generated report can be fluent and globally plausible while still omitting the subtle abnormality. Therefore, in addition to generating free-form descriptions with the raw benchmark prompt, we use a fixed lesion-aware decoding probe to measure whether lesion-relevant tokens receive probability mass at the beginning of generation.

The probe is evaluation-only. It does not modify the prompt, decoder, visual encoder, EasyLens module, or generation procedure. For all report-generation experiments, we use the raw benchmark prompt, set the maximum generation length to 160 tokens, and compute lesion-token statistics over the first $T = 3$ decoding steps.

Let x_i be the input image and q_i be the report-generation prompt for sample i. Given the enhanced visual representation $\bar{Z}_i$, the frozen VLM defines a next-token distribution:

$$p_{i,t}(v) = p_\theta(v \mid q_i, \bar{Z}_i, y_{i,<t}),$$

where v is a vocabulary token and t is the decoding step. For each sample, we define a strict lesion-token set $\mathcal{L}_i$ from the benchmark target description. This set contains tokens corresponding to the lesion category, morphology, or lesion-relevant clinical description.

At each decoding step, we compute the lesion-token probability mass:

$$\ell_{i,t} = \sum_{v \in \mathcal{L}_i} p_{i,t}(v).$$

To avoid rewarding diffuse probability mass over a large vocabulary, we also use a strict top-K version:

$$\ell_{i,t}^{\text{strict}} = \sum_{v \in \mathcal{L}_i \cap \text{TopK}(p_{i,t})} p_{i,t}(v),$$

where $\mathrm{TopK}(p_{i,t})$ denotes the set of highest-probability vocabulary tokens at step t.

The final lesion-aware report-generation probe score is computed over the first three decoding steps:

$$\mathrm{Probe}(i) = \frac{100}{T} \sum_{t=1}^{T} \ell_{i,t}^{\mathrm{strict}}, \quad T = 3.$$

The dataset-level score is the average over all report-generation samples:

$$\mathrm{Probe} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{Probe}(i).$$

We focus on the first three decoding steps because early tokens usually determine the main clinical content of a generated report. If the model fails to assign probability mass to lesion-relevant tokens at this stage, later fluent continuation often cannot recover the missed subtle abnormality. This probe therefore measures lesion-awareness at the point where the model commits to the report content, while remaining independent of any training or inference-time modification.

C Additional Experimental Details

C.1 Dataset Details and Benchmark Construction

We organize the evaluation datasets into two groups. The first group targets *subtle lesions* and is built from ReXGroundingCT, LIDC-IDRI, and AbdomenAtlas 3.0 Mini, denoted as ReX, LIDC, and Abdomen, respectively. Following the terminology in the introduction, these cases focus on abnormalities whose visual evidence is spatially sparse, low-contrast, or weakly distinguishable from surrounding anatomical structures. The second group contains more salient or regular lesion datasets, including MIMIC-CXR, Kvasir-SEG, and BKAI-Polyp. These datasets are used to evaluate whether EasyLens preserves standard medical report-generation ability when pathological findings are visually more apparent.

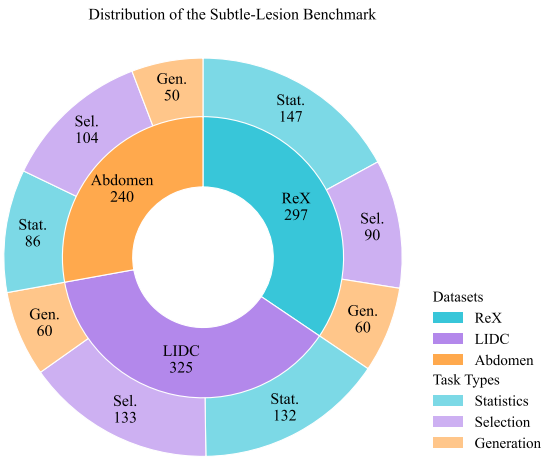


Figure 5: Dataset and task distribution of the subtle-lesion benchmark.

For the subtle-lesion group, we construct a unified benchmark with three task types: regional statistics (Stat.), region selection (Sel.), and lesion-aware report generation (Gen.). Fig. 5 summarizes the distribution of the subtle-lesion benchmark across datasets and task types.

For Stat. and Sel., each image is divided into a 3×3 grid. Grid regions containing pathological patches are treated as positive regions. Stat. evaluates whether the model can correctly recover the number of positive regions, while Sel. evaluates whether it can identify which regions contain pathological evidence. These two tasks test complementary aspects of subtle-lesion perception: Stat. emphasizes regional evidence aggregation, whereas Sel. emphasizes spatial discrimination between lesion cues and visually similar anatomical background.

For subtle-lesion Gen., we evaluate whether subtle pathological evidence becomes more accessible during report generation. Instead of relying only on the final generated report, we measure the probability assigned to lesion-related tokens during decoding. This design is important because the enhanced visual embeddings may still be ignored by the decoder if the corresponding subtle-lesion tokens do not enter the final top-ranked generation path. Therefore, lesion-token probability provides a more direct signal of whether encoder-side enhancement actually increases the decoder's awareness of subtle pathological evidence.

Models	ReX Acc.	LIDC Acc.	Abdomen Acc.
MedGemma1.5	6.67	13.33	18.00
w/ EasyLens	**16.67**	**26.67**	**32.00**
Hulu-Med	5.00	8.33	14.00
w/ EasyLens	**18.33**	**20.00**	**32.00**

Table 5: Diagnosis accuracy (%) in subtle-lesion report generation.

For the salient or regular lesion group, we evaluate standard report generation on MIMIC-CXR, Kvasir-SEG, and BKAI-Polyp. Unlike the subtle-lesion Gen. task, this evaluation uses the final generated reports rather than lesion-token probabilities, because the goal is to test general report quality when the target findings are more visually apparent. We report BLEU-1, BLEU-4, METEOR, and ROUGE-L, which evaluate lexical overlap and sentence-level similarity between generated reports and reference reports. These results complement the subtle-lesion benchmark by testing whether EasyLens preserves general medical image interpretation and report-generation ability beyond subtle-lesion scenarios.

C.2 Diagnosis Accuracy in Medical Report Generation

We further evaluate whether the enhanced subtle-lesion evidence is reflected in the final generated reports. The lesion-token probability probe described in Sec. B.2 provides an auxiliary measure of lesion awareness during decoding. However, an increase in lesion-token probability does not necessarily imply that the corresponding abnormality is included in the final report. We therefore complement this probe with an output-level evaluation of diagnosis accuracy on ReX, LIDC, and Abdomen, using MedGemma1.5 and Hulu-Med as the backbones.

For each sample, a fixed GPT-based evaluator determines whether the generated report correctly includes the corresponding ground-truth subtle-lesion description. Diagnosis accuracy is defined as the percentage of generated reports satisfying this criterion, with the same evaluator used across all compared methods. Unlike the decoding probe, this metric evaluates the actual generated text and directly measures whether the target pathological finding is conveyed in the report. It focuses specifically on target-lesion reporting rather than the overall correctness or completeness of the entire report.

As shown in Table 5, EasyLens consistently improves diagnosis accuracy across both backbones and all three datasets. For MedGemma1.5, accuracy increases from 6.67%, 13.33%, and 18.00% to 16.67%, 26.67%, and 32.00% on ReX, LIDC, and Abdomen, respectively. For Hulu-Med, the corresponding results improve from 5.00%, 8.33%, and 14.00% to 18.33%, 20.00%, and 32.00%. These results provide complementary output-level evidence that EasyLens helps the frozen medical VLMs include subtle pathological findings in their generated reports, rather than yielding improvements only in the auxiliary lesion-token probability probe.

C.3 Additional Evaluation on General and Salient-Lesion Report Generation

This section complements the main quantitative results by evaluating the benignness of EasyLens beyond subtle-lesion benchmarks. The main paper focuses on subtle-lesion perception, where EasyLens brings consistent gains across ReX, LIDC, and Abdomen. Here, we further evaluate whether the same enhancement mechanism affects standard report generation on general or more salient lesion cases. Table 6 reports results on MIMIC-CXR, Kvasir-SEG, and BKAI-Polyp.

EasyLens preserves general chest X-ray reporting ability. On MIMIC-CXR, EasyLens remains highly comparable to its MedGemma1.5 backbone. It slightly improves BL-1 from 21.24 to 21.38, while BL-4, MTR, and RG-L remain nearly unchanged. This shows that enhancing subtle visual evidence does not disturb the global thoracic context or the original language-generation behavior of the frozen backbone.

EasyLens also remains effective on salient lesion report generation. On Kvasir-SEG, EasyLens improves MedGemma1.5 from 9.56/15.87/15.39 to 10.30/16.55/16.18 on BL-1/MTR/RG-L. On BKAI-Polyp, it further improves the same backbone from 8.12/17.05/13.56 to 8.31/17.54/13.62. Since BL-4 is zero for all methods on the two polyp datasets, the comparison mainly relies on BL-1,

Dataset	Metric	LLaVA-Med	RadFM	Hulu-Med	MedGemma	MedGemma1.5	EasyLens
MIMIC-CXR	BL-1	0.74	0.11	6.97	14.39	21.24	**21.38**
	BL-4	0.04	0.01	1.11	1.73	**3.64**	3.63
	MTR	4.68	2.24	11.63	13.67	**17.72**	17.63
	RG-L	8.37	4.37	14.45	16.89	**18.98**	18.91
Kvasir-SEG	BL-1	11.49	1.60	**16.14**	13.99	9.56	10.30
	BL-4	0.00	0.00	0.00	0.00	0.00	0.00
	MTR	22.69	1.31	12.49	**25.38**	15.87	16.55
	RG-L	8.61	2.41	**21.40**	15.59	15.39	16.18
BKAI-Polyp	BL-1	11.89	1.86	**15.01**	14.25	8.12	8.31
	BL-4	0.00	0.00	0.00	0.00	0.00	0.00
	MTR	24.21	1.43	14.20	**26.04**	17.05	17.54
	RG-L	8.30	2.03	**21.26**	16.32	13.56	13.62

Table 6: Comparison of report generation performance across different medical VLMs on MIMIC-CXR, Kvasir-SEG, and BKAI-Polyp datasets.

MTR, and RG-L. These results support the conclusion in Sec. 4.3: EasyLens improves subtle-lesion perception without sacrificing broader medical report-generation ability.

C.4 Plug-and-Play Interface Calibration

Plug-and-play deployment allows lightweight interface calibration without modifying the host VLM. EasyLens is attached to frozen medical VLMs as an inference-time adapter. Across all backbones, the VLM parameters are frozen, the decoding procedure is unchanged, and the EasyTag–EasyAmplifier architecture remains identical. The only model-dependent step is interface calibration, where a small number of scalar operating parameters are selected from a shared discrete grid.

Interface calibration is necessary because different VLMs have different visual token geometries. Medical VLMs vary in embedding scale, lesion-background separability, and the spatial distribution of lesion evidence. A residual strength that is suitable for one encoder may be too weak for another encoder with lower lesion contrast, or too strong for an encoder whose lesion tokens are already sharply localized. Similarly, some backbones concentrate lesion evidence in a few patches, while others distribute it across more visual tokens. Therefore, these calibrated parameters should be interpreted as adapter-interface parameters, not as model-specific learned components.

The calibrated parameters control only the strength and coverage of the same shared mechanism. The interface parameters are β, k, and s. The residual strength β controls the update magnitude. The candidate budget $k = K$ controls how many patches are retained after morphology calibration, and the seed budget $s = K_0$ controls how many globally highest-scoring seeds initialize calibration. The support reference is the single nearest neighbor specified in Appendix A.4. These parameters are low-dimensional, interpretable, and selected from a small discrete grid.

EasyLens does not tune parameters per test example. For each dataset–task or backbone–task setting, the operating point is fixed before evaluation and then applied to all test samples. Thus, the calibration is an interface-level deployment choice rather than instance-level optimization. No backbone parameters are updated, no model-specific EasyLens architecture is introduced, and no task-specific instruction tuning is performed.

Task-conditioned regimes reflect different precision–recall requirements. For status or counting-oriented tasks, EasyLens uses a high-recall regime because missing a subtle lesion may directly change the answer. For lesion selection, EasyLens uses a balanced regime that preserves enough candidate regions while avoiding excessive distractor amplification. For report generation, EasyLens uses a conservative-generation regime with stable token coverage and moderate residual strength, so that lesion awareness is improved without destabilizing global context or language fluency.

Regular lesion report generation uses the same calibration principle. MIMIC-CXR adopts a conservative setting because reports require broad thoracic context. Kvasir-SEG uses a moderate setting because the lesion is localized and visually salient. BKAI-Polyp uses a larger coverage setting because polyp appearance and boundary morphology are more heterogeneous. These choices follow the same interface-calibration principle used for subtle-lesion tasks.

Layer	ReX			LIDC			Abdomen			Macro
	Stat.	Sel.	Gen.	Stat.	Sel.	Gen.	Stat.	Sel.	Gen.	
1	**66.67**	**31.11**	**5.15**	**30.30**	**36.09**	**45.86**	**52.33**	**55.77**	**40.67**	**40.44**
6	48.98	24.44	4.74	24.24	30.08	44.28	45.35	50.00	38.77	34.54
11	46.94	25.56	5.04	22.73	30.08	41.36	45.35	49.04	37.98	33.79
16	44.22	24.44	4.51	18.18	30.83	41.51	45.35	51.92	37.34	33.14
21	46.26	23.33	4.41	17.42	31.58	43.09	48.84	50.96	37.89	33.75
26	57.14	25.56	5.02	21.21	30.83	40.75	46.51	47.12	38.56	34.74

Table 7: Ablation of the vision layer used by EasyLens on ReX, LIDC, and Abdomen. Early injection achieves the best overall performance.

Setting	ReX			LIDC			Abdomen		
	Stat.	Sel.	Gen.	Stat.	Sel.	Gen.	Stat.	Sel.	Gen.
topk_32_seed4	46.26	26.67	**5.17**	**26.52**	30.08	42.40	47.67	49.04	**39.79**
topk_64_seed8	45.58	26.67	3.93	21.21	30.08	42.52	**52.33**	48.08	37.96
topk_128_seed16	54.42	31.11	5.15	**26.52**	29.32	43.16	48.84	**52.88**	39.76
topk_256_seed32	66.67	28.89	4.04	25.76	**31.58**	**45.46**	**52.33**	**52.88**	38.18
topk_512_seed48	**71.43**	**32.22**	3.86	25.00	29.32	44.76	48.84	51.92	38.87

Table 8: Ablation of the EasyTag candidate token budget k and seed budget s on ReX, LIDC, and Abdomen. Different tasks prefer different operating points due to distinct precision–recall requirements.

C.5 Common Implementation Settings

Default implementation settings follow the shared method definition. In the default MedGemma1.5 setting, EasyLens is applied after the first vision block, with layer numbers starting at 1. The bank and inference features are taken at the same interface. Seed selection and candidate selection use the global rankings in Appendix A.3, with budgets $s = K_0$ and $k = K$. Morphology calibration uses $\lambda = 1.0$, and support retrieval uses the prototype-conditioned nearest-neighbor rule in Appendix A.4. The final residual update is gated by $w_q > \eta$. For VQA-style tasks, the maximum number of generated tokens is 96. For report-generation tasks, it is 160, and the lesion-token scoring window uses three decoding steps.

C.6 Additional Interface Ablations

These ablations complement the main component ablation in Sec. 4.4. The main paper verifies the effectiveness of EasyTag and EasyAmplifier. Here, we further analyze how the same EasyTag–EasyAmplifier mechanism should be interfaced with frozen VLMs, including the visual injection position and the candidate token budget. These experiments do not introduce additional task-specific modules; they only study the operating interface of the fixed EasyLens design.

C.6.1 Visual Injection Position

Early visual injection is the most effective for preserving subtle lesion evidence. Table 7 compares injecting EasyLens into different vision layers. Injecting at Layer 1 achieves the best results across all nine subtle-lesion metrics and the highest macro score of 40.44. Moving the injection to deeper layers consistently reduces the macro score, e.g., 34.54 at Layer 6, 33.79 at Layer 11, and 33.14 at Layer 16. Although Layer 26 partially recovers the ReX Stat. score, it remains clearly weaker than Layer 1 on the overall macro score and on LIDC/Abdomen.

This supports using EasyLens as an early visual-interface adapter. Subtle lesions are low-contrast and spatially localized. If enhancement is applied too late, patch-level lesion evidence may already be compressed by the vision encoder and multimodal alignment layers. Early residual injection amplifies lesion-relevant visual tokens before this compression, while still allowing the frozen VLM to perform its original downstream reasoning.



Figure 6: Qualitative comparison of selected patches under different candidate and seed budgets.

Joint candidate and seed budgets reflect a precision–recall trade-off. Table 8 varies the EasyTag candidate budget k together with the seed budget s. No single budget dominates all datasets and metrics. On ReX, a larger budget improves Stat. and Sel., while Gen. is better with a compact budget. On LIDC, a larger budget benefits selection and generation, with the joint setting $(k, s) = (256, 32)$ achieving the best Sel. and Gen. On Abdomen, selection prefers a moderate-to-large budget, while generation again benefits from a smaller candidate set.

This behavior is consistent with task-conditioned calibration. Regional counting and selection tasks often require high recall over candidate lesion regions, especially when multiple subtle abnormalities may be present. Report generation, however, must preserve lesion evidence without introducing excessive background distractors. When k is too small, the selector may miss weak lesion evidence; when k is too large, additional candidates may dilute the lesion signal. The joint candidate/seed-budget ablation therefore supports the use of a fixed EasyLens mechanism with lightweight interface calibration rather than a one-size-fits-all operating point.

C.6.2 Qualitative Analysis of Candidate Token Budget

The topk case study provides a visual explanation for the budget ablation. Figure 6 visualizes how different candidate budgets affect lesion evidence selection. A small budget tends to focus on the most confident local responses, which can reduce distractors but may miss weak or spatially scattered lesions. A moderate budget preserves more lesion-related alternatives and is therefore more suitable for selection-oriented tasks. An excessively large budget may introduce background structures with similar local appearance, diluting the lesion signal during generation.

This qualitative behavior matches the quantitative trend in Table 8. The visualization shows that EasyLens does not rely on arbitrary token amplification. Instead, the candidate budget controls the amount of visual evidence exposed to the frozen VLM. Proper calibration balances lesion recall and background suppression, which is crucial for subtle-lesion grounding and faithful report generation.

D Limitations

EasyLens has several limitations. First, although it is training-free at inference time, it relies on an offline prototype bank constructed from reference images and lesion masks. Its effectiveness may

therefore depend on the coverage of the prototype bank, especially when target lesions differ in anatomy, scale, or imaging appearance from the reference set. Second, the current implementation uses fixed interface parameters for each dataset–task setting, including the token budget and residual strength. This simple calibration strategy may not be optimal for all cases, particularly when lesion size or visual contrast varies substantially. Third, our experiments are conducted on a limited set of medical VLM backbones and lesion-oriented benchmarks. Broader validation across additional imaging modalities, institutions, and clinical tasks is still needed to assess the generality of the proposed approach.

E Licenses and Terms for Existing Assets

This work uses existing public or credentialed medical datasets and publicly released model checkpoints only for research evaluation. We do not claim ownership of any third-party assets. We cite the original creators of the datasets, models, and codebases used in our experiments, and we use each asset according to its corresponding license, data-use agreement, or access policy. We do not redistribute third-party medical images, reports, masks, or model weights. Users who reproduce our experiments should obtain the original assets from their official sources and comply with the corresponding licenses and access requirements.

Any supplementary material released with this paper contains only our implementation, configuration files, and evaluation scripts. It does not include protected health information, third-party medical images, clinical reports, segmentation masks, or third-party model weights. Reproduction of the reported experiments requires users to independently obtain the relevant datasets and model checkpoints from their original providers and to comply with the corresponding licenses, data-use agreements, and citation requirements.

Table 9: Existing datasets and model assets used in this work. For assets with restricted or gated access, users must obtain access from the original provider.

Asset	Role in this work	License or access terms
ReXGroundingCT	Subtle-lesion CT evaluation and prototype construction	Released under CC BY-NC-SA 4.0 with gated access requiring acceptance of dataset conditions. We use it only for non-commercial research evaluation and do not redistribute the data.
LIDC-IDRI	Lung CT subtle-lesion evaluation and prototype construction	Distributed through TCIA. We follow the TCIA Data Usage Policy and the Creative Commons Attribution 3.0 Unported license terms associated with the collection. We cite the original dataset and do not redistribute the data.
AbdomenAtlas 3.0 Mini	Abdominal CT subtle-lesion evaluation and prototype construction	Released under CC BY-NC-SA 4.0. We use it for non-commercial research evaluation and do not redistribute the data.
MIMIC-CXR	Chest X-ray report-generation evaluation	Accessed through PhysioNet under the PhysioNet Credentialed Health Data License 1.5.0 and the corresponding Data Use Agreement. Access requires credentialed-user approval and required training. We do not share the data, attempt re-identification, or redistribute any images or reports.
Kvasir-SEG	Polyp report-generation evaluation	The official terms restrict use to research and educational purposes, and commercial use requires prior written permission. We cite the dataset paper and do not redistribute the dataset.
BKAI-Polyp / BKAI-IGH NeoPolyp	Polyp report-generation evaluation	Accessed as Kaggle competition data subject to the original competition rules. We use it only for research evaluation and do not redistribute the dataset.
LLaVA-Med	Frozen medical VLM backbone and baseline	The official release states that the data, code, and model checkpoints are intended for research use only, with non-commercial restrictions and additional terms inherited from the underlying LLaMA, Vicuna, and GPT-4 resources. It is not intended for clinical care or clinical decision making.
RadFM	Frozen medical VLM backbone and baseline	The official code repository is released under the MIT License. The public checkpoint page used in our experiments does not specify a separate model-card license; therefore, we use the checkpoint only for research evaluation and do not redistribute the weights.
Lingshu	Frozen medical VLM backbone and baseline	The official Hugging Face model card lists the MIT License. We use the released checkpoint only for research evaluation and not for clinical deployment.
Hulu-Med	Frozen medical VLM backbone and baseline	The official repository is released under the Apache License 2.0. We use the released checkpoint and code only for research evaluation and not for clinical deployment.
MedGemma and MedGemma1.5	Frozen medical VLM backbone and baseline	The model weights are governed by the Health AI Developer Foundations terms of use. The accompanying repository/tutorial code is released under Apache License 2.0. We use MedGemma models only as frozen research backbones and do not treat them as clinical-grade diagnostic systems.
MedKLIP	Encoder-enhancement baseline	We cite the official paper and repository. We did not identify a separate explicit license for the released repository or checkpoint at the time of writing; therefore, we use it only for research comparison and do not redistribute its code or weights.
KAD	Encoder-enhancement baseline	The official code repository is released under the MIT License. Any associated model or data resources are used according to their original access terms, and we do not redistribute third-party assets.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and Introduction clearly state that EasyLens is a training-free plug-and-play subtle-lesion representation amplifier for frozen medical VLMs. The claimed contributions are limited to EasyBank, EasyTag, and EasyAmplifier, and the experimental claims are supported on ReXGroundingCT, LIDC-IDRI, AbdomenAtlas 3.0 Mini, MIMIC-CXR, Kvasir-SEG, and BKAI-Polyp with multiple frozen medical VLM backbones.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Appendix D discusses dependence on the offline mask-derived bank, limited bank and dataset coverage, fixed interface calibration, and the need for broader validation across modalities, institutions, and clinical tasks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Justification: The paper does not present theorem-level theoretical results or formal proofs. The mathematical formulations in Section 3 and Appendix A define the EasyBank, EasyTag, and EasyAmplifier procedures rather than proving theoretical guarantees.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: Sections 3 and 4 describe the proposed algorithm, frozen-backbone setting, evaluation tasks, metrics, and compared backbones. Appendix A provides detailed method definitions, Appendix B specifies interface calibration and common implementation settings, and Appendix C describes dataset/task construction and additional evaluations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The submission includes anonymized code and reproduction instructions in the supplementary material. The experiments use existing public or credentialed medical datasets, and the paper describes the preprocessing, task construction, evaluation metrics, and EasyLens implementation settings needed to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Sections 4.1 and 4.2 describe the datasets, task types, metrics, frozen VLM backbones, raw benchmark prompts, and generation length. Appendix B.1 and Table 4 specify the EasyLens interface parameters and common settings, including injection layer, support retrieval mode, seed initialization, score threshold, retrieval temperature, and maximum generated tokens. Since EasyLens is training-free, optimizer and training schedule details are not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No].

Justification: We report point estimates without error bars, confidence intervals, or statistical significance tests. Deterministic decoding does not quantify uncertainty due to finite evaluation samples or bank construction.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: All experiments all running on A6000 with 48GB memory,and table 2 reports the average inference time.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The research uses existing medical image datasets and evaluates a frozen-model inference-time adapter. It does not collect new patient data or conduct new human-subject experiments. The method is intended for research on medical VLM sensitivity and should not be used as an autonomous clinical diagnostic system without appropriate validation and oversight.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: We discuss the potential benefit of improving sensitivity to clinically important subtle lesions, which may support more reliable medical image interpretation. We also note that EasyLens is not a standalone diagnostic system and should be used with clinician oversight, with attention to validation across populations, institutions, and privacy-sensitive medical data handling.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release a new high-risk foundation model, scraped dataset, image generator, or autonomous clinical system. EasyLens is an inference-time adapter evaluated on existing medical VLMs and datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: The paper cites the original sources of the datasets and model backbones used in the experiments, and the supplementary material lists the corresponding versions, licenses, and access terms in Appendix E. All existing assets are used according to their stated research-use or credentialed-access policies.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The paper introduces a unified subtle-lesion benchmark derived from existing datasets and documents the task construction, dataset/task distributions, evaluation metrics, and usage protocol in Appendix C. The released asset is anonymized and accompanied by documentation for reproduction.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The work does not involve crowdsourcing, user studies, or new human-subject data collection.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The study does not collect new human-subject data and uses existing medical imaging datasets under their corresponding access and usage policies. No new intervention, recruitment, or participant interaction is involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: In addition to writing, editing, and formatting assistance, a GPT-based evaluator is used for the output-level target-lesion reporting evaluation described in Appendix C.2. The frozen VLMs used by EasyLens are described in the main paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.