EXPERTROUTE: DISTILLING RADIOLOGIST VISUAL SEARCH INTO COMPACT EVIDENCE FOR CHEST X-RAY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

For chest X-ray report generation, visual context is useful only if it preserves the evidence needed for faithful diagnostic description. We introduce ExpertRoute, which distills radiologists' visual-search behavior into an image-conditioned expert prior and combines it with independent representation-space semantic coverage under an exact visual-context budget. The expert prior is learned once from eye tracking and requires no gaze at inference. On a frozen LLaVA-Rad backbone, ExpertRoute reduces the visual context entering the LLM from 1,369 to 192 tokens (14.0%). Despite this reduction, it improves RadGraph and GREEN over full-context inference across three evaluation cohorts: MIMIC Validation, a locked MIMIC held-out cohort, and OpenI. On the locked held-out cohort, these RadGraph and GREEN gains are supported by patient-cluster bootstrap. Across the evaluated token budgets, ExpertRoute exhibits a non-monotonic response, with Micro-F1, Macro-F1, and RadGraph all reaching their best observed values at 192 tokens. At this operating point, the compact visual context also reduces per-sample generation latency by 16.2% relative to full-context inference. These results show that report quality depends on how visual evidence is composed, not simply on how much context is retained.

1 INTRODUCTION

Multimodal language models are increasingly capable of generating open-ended radiology reports (Li et al., 2023; Zambrano Chaves et al., 2025), but their computational cost grows with the length of the visual context (Cao et al., 2023; Chen et al., 2024). In the frozen LLaVA-Rad backbone studied here, a 518×518 chest radiograph is represented by a 37×37 grid of 1,369 visual patch tokens before language-model decoding, following the patch-based representation widely used in vision transformers (Dosovitskiy et al., 2021). A natural response is to reduce the visual context. In our implementation, routing occurs after the multimodal projector, so the reduction specifically targets the visual sequence presented to the LLM rather than the upstream vision encoder. In radiology, however, this raises a harder question: under a constrained context, which visual evidence must remain available to support a faithful diagnostic description?

This distinction matters because radiographic evidence is heterogeneous. Some findings are focal and spatially compact, others are distributed, and normality itself may require broad anatomical coverage. A selector optimized only for model-derived saliency or representation similarity can preserve a useful summary of the image without necessarily preserving the same regions that domain experts prioritize during interpretation. Conversely, expert attention alone can be too concentrated to represent the complete image. Under a finite context budget, report quality therefore depends not only on how many visual tokens are retained, but also on how expert-prior and representation-space evidence are balanced.

Visual-token reduction has been studied in vision transformers through adaptive pruning, sampling, and merging (Li et al., 2024; Ryoo et al., 2021; Rao et al., 2021; Liang et al., 2022; Yin et al., 2022; Fayyaz et al., 2022; Bolya et al., 2023), and extended to multimodal models by PuMer, FastV, TRIM, and LLaVA-PruMerge (Cao et al., 2023; Chen et al., 2024; Song et al., 2025; Shang et al., 2025). Existing multimodal methods largely infer importance from the model itself. VisionZip compresses

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

dominant and contextual visual representations (Yang et al., 2025); SparseVLM progressively removes visual tokens inside the language model (Zhang et al., 2025d); SCOPE combines saliency with representation-space coverage (Deng et al., 2025); VisPruner uses vision-encoder attention together with token diversity (Zhang et al., 2025a); and CDPruner maximizes instruction-conditioned diversity (Zhang et al., 2025b). These approaches provide strong model-derived baselines. Radiology additionally provides a source of supervision independent of the model's own attention geometry: radiologists' visual-search behavior.

We use this signal as a form of expert-prior distillation. Eye tracking from the REFLACX chest X-ray dataset (Bigolin Lanfredi et al., 2022) is used to train a lightweight gaze scorer that predicts an image-conditioned visual-search distribution from the image alone. At inference, the scorer identifies a compact set of expert-prior tokens, while an independent semantic selector preserves broad representation-space coverage. ExpertRoute then composes the two ranked sources under an exact token budget. The central design principle is complementary rather than competitive: semantic coverage prevents an expert prior from becoming spatially myopic, while gaze supervision provides an expert-derived prioritization signal beyond the frozen model's own saliency and similarity scores.

The experiments expose a broader property of visual context that motivates the method. Across the evaluated budgets, ExpertRoute reaches its best observed Micro-F1, Macro-F1, and RadGraph at 192 tokens. At the same operating point, it is the strongest of the four budget-swept methods on all three Validation metrics and achieves higher report-level RadGraph and GREEN than the full 1,369-token context on MIMIC Validation, locked MIMIC held-out, and OpenI, while Full retains stronger finding-level CheXbert F1 on the locked cohort. Thus, the observed operating point reflects a quality trade-off rather than universal superiority of compact context, motivating explicit control over how expert-derived priorities and semantic coverage are composed.

Our contributions are threefold. First, we introduce a lightweight hierarchical gaze scorer that distills radiologist visual-search behavior from eye tracking into an image-conditioned expert prior, requiring no gaze at inference. Second, we propose ExpertRoute, an exact-budget router that preserves independently ranked expert and semantic evidence and resolves their interaction only by deterministic index-level de-duplication. Third, we characterize this evidence bottleneck across token budgets, representative context-reduction baselines, a locked held-out cohort, external OpenI evaluation, and systems measurements.

2 Related Work

Chest X-ray report generation. Radiology report generation has progressed from hierarchical encoder–decoder models to memory-driven, cross-modal, factual-consistency, and knowledge-aware systems (Jing et al., 2018; Chen et al., 2020; 2021; Miura et al., 2021; Yang et al., 2022). Large-scale resources such as MIMIC-CXR and OpenI have supported this progression (Johnson et al., 2019; Demner-Fushman et al., 2016; Zambrano Chaves et al., 2025), while medical image–text and VLM systems including ConVIRT, GLoRIA, LLaVA-Med, Med-Flamingo, BiomedCLIP, and LLaVA-Rad broaden multimodal medical generation and reasoning (Zhang et al., 2022; Huang et al., 2021; Li et al., 2023; Moor et al., 2023; Zhang et al., 2025c; Zambrano Chaves et al., 2025). Retrieval-based generation provides a complementary direction (Jeong et al., 2024). Evaluation commonly combines finding-level CheXbert with graph- and factuality-oriented RadGraph and GREEN, alongside lexical overlap (Smit et al., 2020; Jain et al., 2021; Ostmeier et al., 2024). We focus on the visual evidence made available to a frozen report generator.

Visual context selection and token reduction. Visual-token reduction builds on adaptive selection, sparsification, sampling, and merging in vision transformers (Ryoo et al., 2021; Rao et al., 2021; Liang et al., 2022; Yin et al., 2022; Fayyaz et al., 2022; Bolya et al., 2023). In VLMs, PuMer, FastV, TRIM, and LLaVA-PruMerge introduced text-informed pruning, attention-based pruning, CLIP relevance, and pruning-plus-merging (Cao et al., 2023; Chen et al., 2024; Song et al., 2025; Shang et al., 2025). More recent methods include VisionZip, SparseVLM, SCOPE, VisPruner, and CDPruner, which respectively emphasize compact contextual representations, progressive pruning, saliency–coverage trade-offs, visual diversity, and instruction-conditioned diversity (Yang et al., 2025; Zhang et al., 2025d; Deng et al., 2025; Zhang et al., 2025a;b). These methods differ in criterion and insertion point. Our method contributes an orthogonal source of information—a learned expert

visual-search prior—and evaluates whether it remains useful when strong model-derived coverage and diversity baselines are available.

Radiologist gaze and expert supervision. Radiologist gaze has been used as weak spatial supervision for localization and diagnostic attention. EGD-CXR and REFLACX provide chest X-ray eye-tracking resources with aligned gaze and reporting information (Karargyris et al., 2020; Bigolin Lanfredi et al., 2022). Recent CoGaze incorporates radiologist gaze as a probabilistic prior during chest X-ray vision–language pretraining, without requiring gaze for downstream inference (Liu et al., 2026). Our objective is different: instead of shaping the backbone representation, gaze supervises a compact image-conditioned scorer that is frozen after training and used explicitly to route visual evidence under a fixed LLM context budget.

3 METHOD

3.1 OVERVIEW: CONSTRUCTING A COMPACT EVIDENCE SET

Let the frozen vision encoder map an input chest radiograph I to final-layer patch representations $H^{(L)} \in \mathbb{R}^{N \times d_v}$, where L denotes the final vision-transformer layer, $N = 37 \times 37 = 1{,}369$ is the number of patch tokens, and d_v is the vision-feature dimension. We denote the full patch-index set by $V = \{1, \ldots, N\}$. The multimodal projector maps all patch embeddings into the language-model embedding space. Our router selects an exact subset

$$S \subseteq V, \qquad |S| = B, \tag{1}$$

where S is the set of retained patch indices and B is the visual-context budget. The projected tokens indexed by S are restored to raster order and passed to the frozen LLM. The principal operating point is $B = 192$. The prompt is fixed to *"Describe the chest X-ray findings in detail."* The vision encoder, projector, and report generator are frozen throughout; only the gaze scorer is learned. Figure 1 summarizes the overall ExpertRoute pipeline from offline expert-prior distillation to exact-budget visual evidence routing and report generation.

3.2 DISTILLING RADIOLOGIST VISUAL SEARCH

Gaze targets. REFLACX fixations supervise an image-conditioned visual-search prior. A fixation contributes an image-pixel-space Gaussian weighted by its duration, clipped to $[0.05, 2.0]$ seconds. The pixel heatmap is area-resized to the 37×37 patch grid. For a nonzero map, normalization produces $\mathbf{q} = (q_1, \ldots, q_N)$, where q_i is the target mass at patch i and $\sum_i q_i = 1$. These densities provide weak spatial supervision instead of lesion masks. Further construction details are given in Appendix A.

Hierarchical spatial scorer. The scorer combines four frozen vision hidden states with the frozen projector output. Let $\mathcal{K} = \{3, 6, 9, 12, \mathrm{proj}\}$ index these five sources, $X_i^{(k)}$ denote the feature of patch i from source k, and ϕ_k be a source-specific LayerNorm followed by a linear projection to 256 dimensions. Vision features have dimension 768 and projected features have dimension 4,096. Source fusion is

$$\tilde{h}_i = \sum_{k \in \mathcal{K}} w_k \phi_k(X_i^{(k)}), \qquad w_k = \frac{\exp(a_k)}{\sum_{\ell \in \mathcal{K}} \exp(a_\ell)}, \tag{2}$$

where a_k is a learned scalar and w_k its normalized source weight. A learned linear map $\psi : \mathbb{R}^2 \to \mathbb{R}^{256}$ embeds normalized patch coordinates $c_i = (x_i, y_i) \in [0, 1]^2$. The descriptors $\tilde{h}_i + \psi(c_i)$ are reshaped into $U^{(0)} \in \mathbb{R}^{37 \times 37 \times 256}$. Two spatial residual blocks refine this grid:

$$U^{(t)} = \mathcal{R}_t\Big(U^{(t-1)}\Big) = U^{(t-1)} + F_t\Big(U^{(t-1)}\Big), \qquad t \in \{1, 2\}, \tag{3}$$

where F_t denotes the residual transformation comprising a depthwise 3×3 convolution, a pointwise 1×1 convolution, GELU, and spatial dropout, while $\mathcal{R}_t$ denotes the complete residual block including the skip connection. A 1×1 prediction layer then produces gaze logits g_i from $U^{(2)}$. Figure 4 illustrates this construction. The scorer has 1,988,358 trainable parameters (1.99M); the backbone and projector remain frozen.

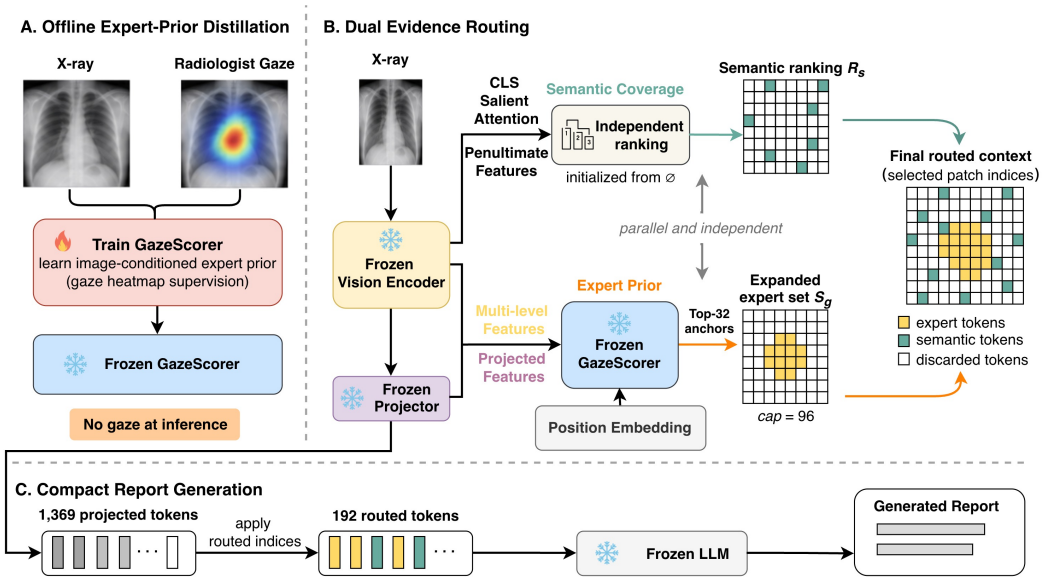

Figure 1: **Overview of ExpertRoute.** Radiologist visual-search behavior is first distilled offline into a frozen image-conditioned expert prior. At inference, ExpertRoute constructs two independent evidence sources: an expert-prior route from the frozen GazeScorer and a semantic-coverage route initialized from the empty set. The expert-selected indices are preserved, while non-overlapping patches are appended from the semantic ranking until the exact visual-context budget is reached. All 1,369 patches are encoded and projected, but only the final 192 routed visual tokens are passed to the frozen LLM for report generation. No gaze annotation is required at inference.

The predicted gaze distribution is $p_i = \exp(g_i)/\sum_{j\in V}\exp(g_j)$. Training minimizes

$$\mathcal{L}_{\text{gaze}} = D_{\text{KL}}(\mathbf{q}\,\|\,\mathbf{p}) + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}, \qquad \lambda_{\text{rank}} = 0.1, \tag{4}$$

where D_{KL} is target-to-prediction KL divergence.

The auxiliary loss applies a margin of 0.1 to min–max-normalized gaze logits for high- versus low-gaze patches. KL uses probability-normalized targets, while ranking uses min–max-normalized targets; the exact sampled-pair objective is given in Eq. 9 of Appendix A. The gaze-scorer checkpoint used in all downstream experiments is selected solely by validation shuffled-AUC (sAUC) and then frozen for report generation.

At the principal $B = 192$ setting, the 32 highest-scoring patches define gaze anchors. Anchors are processed in descending score order, with patch index breaking ties. Each anchor expands to its radius-1 grid neighborhood, including diagonal neighbors and clipped at image boundaries. Unique indices are collected in anchor-priority and within-neighborhood raster order, with $B_g^{\max} = \lfloor B/2 \rfloor = 96$ at the principal budget. Because neighborhoods overlap and are clipped at image boundaries, $|S_g|$ is image-dependent and can remain below the cap; on MIMIC Validation it averages 78.24 tokens at $B = 192$. The resulting index set S_g seeds the final composition, and the semantic branch fills the remaining budget.

3.3 Independent Semantic Coverage

The second branch follows the SCOPE selection principle in the frozen vision space (Deng et al., 2025). For each patch index $i \in V$, let A_i denote its head-summed CLS-to-patch saliency from the penultimate vision layer and let z_i denote its frozen penultimate-layer patch feature, normalized to unit length for cosine computation. Let $C \subseteq V$ denote the semantic selector's current working set. For nonempty C, the independent branch uses zero-baseline representation-space coverage

$$f(C) = \sum_{u\in V} \max\left\{0, \max_{s\in C}\cos(z_u, z_s)\right\}, \tag{5}$$

with $f(\varnothing) = 0$. Here u indexes every patch whose representation should be covered, s indexes a patch already in C, and $\cos(\cdot, \cdot)$ denotes cosine similarity. For a candidate patch $v \in V \setminus C$, its marginal coverage gain is $\Delta(v; C) = f(C \cup \{v\}) - f(C)$. Following SCOPE with the fixed saliency exponent $\beta = 1$, candidates are greedily scored by

$$s_{\text{sem}}(v; C) = \Delta(v; C) A_v, \tag{6}$$

where A_v is the candidate's saliency. Starting from $C = \varnothing$, the implementation materializes the first B elements of the independent semantic ordering,

$$R_s = \text{GreedyPrefix}_B(V, \varnothing) = (r_1, r_2, \ldots, r_B), \tag{7}$$

where $r_t \in V$ is the patch index chosen at greedy step t. The semantic ordering is independent of S_g: gaze-derived indices neither initialize nor reweight this ranking. The implementation precomputes the full $N \times N$ cosine-similarity matrix and materializes only a budget-limited prefix. With T greedy steps, the selector requires $O(N^2 d_v + TN^2)$ time and $O(N^2)$ similarity storage; ExpertRoute and SCOPE use $T = B$, while the gaze-conditioned comparator uses only its remaining semantic quota. Tie-breaking and exact implementation details are given in Appendix B.

3.4 Expert–semantic composition

Starting from the expert-prior set S_g, we traverse the ordered semantic ranking R_s and append the highest-ranked indices that are not already in S_g until the exact budget is reached. Formally,

$$S_{\text{sem}} = \text{Set}\big[\text{Prefix}_{B-|S_g|}(\text{Filter}(R_s, V \setminus S_g))\big], \qquad S = S_g \cup S_{\text{sem}}, \tag{8}$$

where $\text{Filter}(R_s, V \setminus S_g)$ removes from the ordered list R_s all indices already contained in S_g, $\text{Prefix}_m(\cdot)$ returns the first m remaining indices, and Set converts the prefix to the semantic-addition index set S_{sem}. Thus $|S| = B$ exactly, all indices are unique and valid, and each branch retains its own ranking semantics. The final indices in S are sorted back to raster order before LLM prefill. This design intentionally avoids treating expert-selected evidence as already semantically covered: expert priority and representation-space redundancy remain distinct signals, and their interaction occurs only through exact index overlap.

For controlled comparison, GazeRoute uses the same learned gaze scorer but fills the remaining budget with deterministic 4×4 stratified spatial completion rather than semantic coverage. At $B = 192$, it expands the top-32 gaze anchors with radius-1 neighborhoods and does not use the 96-token gaze cap of the expert–semantic routes.

GazeSemanticRoute instead initializes semantic coverage from S_g: $m_u = \max_{s \in S_g} \cos(z_u, z_s)$, scores candidates by $A_v \sum_u \max(0, \cos(z_u, z_v) - m_u)$, and updates m_u after each selection. ExpertRoute differs by preserving an independent empty-set semantic ranking and resolving overlap only at final composition. Full allocation and tie-breaking rules are given in Appendix B.

3.5 Routing placement and compute

ExpertRoute, GazeRoute, and GazeSemanticRoute route after the multimodal projector, so all 1,369 patches are encoded and projected while only 192 visual tokens enter the LLM. SCOPE and VisionZip intervene earlier, whereas SparseVLM+ progressively prunes inside the LLM. The evaluated implementations use these different insertion points, and therefore report realized wall-clock latency and GPU memory in addition to nominal token counts.

4 Experiments

4.1 Setup

Backbone and cohorts. All methods use the same frozen LLaVA-Rad backbone and report-generation task protocol (Zambrano Chaves et al., 2025) with the BiomedCLIP-CXR-518 vision encoder family (Zhang et al., 2025c). Method development and budget analysis use a fixed MIMIC Validation cohort of 2,706 images from 477 patients. Confirmatory evaluation uses a locked MIMIC held-out cohort of 580 frontal case records from 197 patients; external evaluation uses 3,285 OpenI

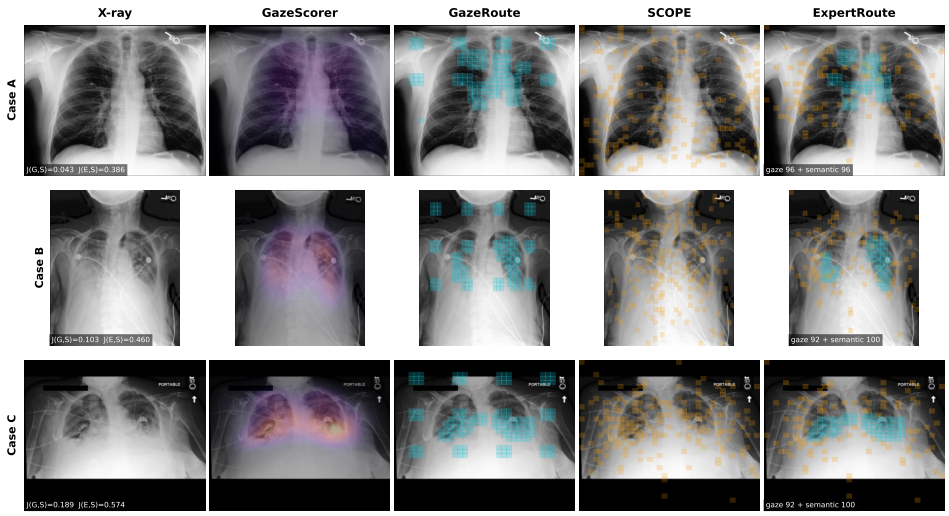

Figure 2: **Qualitative comparison of frozen routing outputs.** Across representative MIMIC Validation cases, GazeRoute emphasizes regions prioritized by the distilled expert prior, while SCOPE provides broader representation-space coverage. ExpertRoute combines these complementary behaviors by preserving expert-derived tokens and adding non-overlapping tokens from an independently constructed semantic ranking.

image records. Gaze-prior learning uses REFLACX only, with 2,083/368/581 train/validation/test records after patient-level repartitioning; scorer training/validation patients are disjoint from both downstream MIMIC cohorts. Because eye-tracking supervision is frontal, lateral Validation images are analyzed separately in Appendix C.3. The prompt, report generator, projector, and decoding configuration are fixed within matched comparisons; split construction and hardware details are provided in Appendix A.

Baselines. We compare full-context inference (Full), the two internal variants GazeRoute and GazeSemanticRoute, and recent external context-reduction baselines: SCOPE (Deng et al., 2025), VisionZip (Yang et al., 2025), SparseVLM+ (Zhang et al., 2025d), CDPruner (Zhang et al., 2025b), and VisPruner (Zhang et al., 2025a). CDPruner and VisPruner are integrated from their released algorithms at a fixed 192-token budget without dataset-specific tuning; VisPruner uses $r = 0.5$ with equal importance/diversity allocation. SparseVLM+ is evaluated with its progressive 1,369→205→136→80 schedule and is reported separately from fixed input-budget methods. Further LLaVA-Rad adaptation details are given in Appendix A.

Metrics and inference. Finding-level agreement is measured with CheXbert Micro- and Macro-F1 over 14 findings (Smit et al., 2020). Report-level structure and factuality are measured with RadGraph F1 (Jain et al., 2021) and GREEN (Ostmeier et al., 2024); BLEU and ROUGE-L provide lexical overlap. Micro-F1, Macro-F1, and RadGraph F1 are the three primary clinical metrics used in the multi-budget analysis. Statistical inference for the reported locked held-out comparisons uses 5,000 paired patient-cluster bootstrap resamples; implementation details appear in Appendix A.

Figure 2 qualitatively illustrates the complementary routing behavior of GazeRoute, SCOPE, and ExpertRoute on representative Validation cases.

4.2 Main MIMIC evaluation

Table 1 places the full-context reference, model-derived compact-context baselines, gaze-supervised routing variants, and ExpertRoute in one matched MIMIC comparison. On Validation, ExpertRoute attains the highest Micro-F1 and Macro-F1 among compact-context methods and the second-highest RadGraph, while VisionZip has the highest RadGraph and GREEN point estimates. On the locked held-out cohort, VisPruner is strongest on finding-level Micro/Macro F1, whereas ExpertRoute

Table 1: Main MIMIC comparison. LLaVA-Rad (Full) is the 1,369-token reference. Model-derived baselines are listed first, followed by gaze-supervised routing variants and ExpertRoute. Publication venues are shown below published baseline names. Bold and underline denote the best and second-best compact-context point estimates within each cohort.

		MIMIC Validation				Locked held-out			
Method	Tokens	Micro↑	Macro↑	RG↑	GREEN↑	Micro↑	Macro↑	RG↑	GREEN↑
LLaVA-Rad (Full) Nat. Commun.'25	1,369	0.5642	0.3778	0.2552	0.4066	0.5639	0.3985	0.2245	0.2899
Model-derived compact-context baselines									
SCOPE NeurIPS'25	192	0.5412	0.3562	0.2823	0.4273	0.5386	0.3543	0.2505	**0.3123**
VisionZip CVPR'25	192	<u>0.5438</u>	0.3603	**0.2943**	**0.4319**	<u>0.5426</u>	0.3603	0.2473	0.2967
SparseVLM+ ICML'25	prog.	0.4603	0.2990	0.2576	0.3880	0.4308	0.3049	0.2188	0.2535
CDPruner NeurIPS'25	192	0.5210	0.3339	0.2751	0.4198	0.5060	0.3304	0.2416	0.2901
VisPruner ICCV'25	192	0.5356	0.3490	0.2839	<u>0.4315</u>	**0.5502**	**0.3876**	0.2513	0.3027
Gaze-supervised routing variants									
GazeRoute	192	0.5305	<u>0.3614</u>	0.2807	0.4130	0.5251	0.3650	0.2454	0.2791
GazeSemanticRoute	192	0.5354	0.3529	0.2871	0.4274	0.5341	0.3766	<u>0.2543</u>	0.3077
ExpertRoute	192	**0.5464**	**0.3620**	<u>0.2875</u>	0.4298	0.5370	<u>0.3778</u>	**0.2557**	<u>0.3120</u>

achieves the highest RadGraph F1, the second-highest Macro-F1, and a GREEN point estimate within 0.0003 of SCOPE.

The comparison with Full exposes the quality trade-off directly: Full retains higher CheXbert finding-level F1, while the 192-token ExpertRoute route improves report-level RadGraph and GREEN. On held-out data, the ExpertRoute–Full RadGraph gain is $+0.0311$ (95% CI $[0.0183, 0.0428]$) and the GREEN gain is $+0.0221$ ($[0.0063, 0.0374]$), with both intervals excluding zero. Together, the two MIMIC cohorts show that compact routing changes the quality profile of the generated reports rather than merely approximating Full with fewer tokens.

4.3 External OpenI generalization

ExpertRoute transfers its compact evidence representation to OpenI with only 14.0% of the full LLM visual context (Table 2). Relative to Full, RadGraph increases from 0.2691 to 0.3063 and GREEN from 0.4973 to 0.5638—absolute gains of 0.0372 and 0.0665. The same route also exceeds GazeRoute and SCOPE on both report-level metrics, supporting the value of composing expert-prior and representation-space evidence beyond the development cohort.

Against CDPruner and VisPruner, ExpertRoute is higher on all six reported OpenI metrics. Across the full compact-context set, SCOPE is strongest on finding-level F1 and VisionZip on RadGraph/GREEN, whereas ExpertRoute is second on Macro-F1, RadGraph, and GREEN and remains close to the best GREEN point estimate (0.5638 versus 0.5665). This supports a balanced external profile rather than a single-metric optimum.

4.4 Budget-dependent evidence composition

We evaluate GazeRoute, SCOPE, GazeSemanticRoute, and ExpertRoute at $B \in \{32, 64, 128, 192, 256\}$, where B is the number of visual tokens passed to the LLM. GazeRoute uses the same fixed-anchor gaze policy with spatial completion, the two expert–semantic routes use gaze cap $\lfloor B/2 \rfloor$, and SCOPE uses $K = B$; complete allocation rules are given in Appendix B.

Figure 3 reports all three primary clinical metrics. At $B = 32$, SCOPE leads Micro-F1, Macro-F1, and RadGraph, indicating the value of broad semantic coverage under extreme compression. At $B = 64$ and 128, the leading method becomes metric-dependent. At $B = 192$, ExpertRoute is the

Table 2: OpenI external evaluation ($N = 3285$). LLaVA-Rad (Full) is the reference. Model-derived baselines are listed first, followed by gaze-supervised variants and ExpertRoute. Publication venues are shown below published baseline names. B-4 and R-L denote BLEU-4 and ROUGE-L. Bold and underline denote the best and second-best compact-context point estimates.

Method	Micro↑	Macro↑	RG↑	GREEN↑	B-4↑	R-L↑
LLaVA-Rad (Full) Nat. Commun.'25	0.5203	0.3361	0.2691	0.4973	0.0420	0.2236
Model-derived compact-context baselines						
SCOPE NeurIPS'25	**0.5292**	**0.3338**	0.2988	0.5601	0.0438	0.2283
VisionZip CVPR'25	0.4974	0.2985	**0.3133**	**0.5665**	**0.0473**	<u>0.2325</u>
SparseVLM+ ICML'25	0.4803	0.2999	0.2911	0.5438	0.0458	0.2275
CDPruner NeurIPS'25	0.4992	0.3099	0.2909	0.5460	0.0439	0.2262
VisPruner ICCV'25	0.4931	0.2863	0.2956	0.5626	0.0453	0.2254
Gaze-supervised routing variants						
GazeRoute	0.4986	0.3076	0.3000	0.5535	0.0447	**0.2368**
GazeSemanticRoute	<u>0.5146</u>	0.3209	0.3054	0.5629	<u>0.0473</u>	0.2318
ExpertRoute	0.5102	<u>0.3265</u>	<u>0.3063</u>	<u>0.5638</u>	0.0466	0.2320

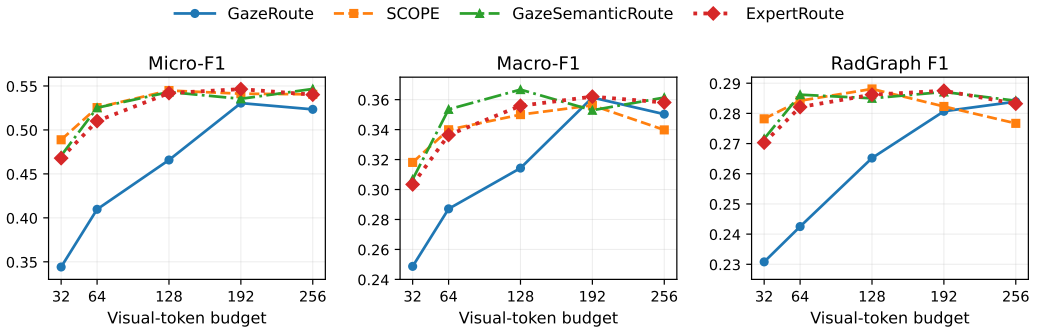

Figure 3: Primary clinical metrics across visual-token budgets on MIMIC Validation. At 32 tokens, SCOPE leads Micro-F1, Macro-F1, and RadGraph; at 192 tokens, ExpertRoute leads all three simultaneously. ExpertRoute itself peaks at 192 on each primary metric and declines at 256, while the leading method at intermediate and larger budgets depends on the evidence-composition rule.

only route that leads all three metrics simultaneously, whereas at $B = 256$ GazeSemanticRoute leads all three. For ExpertRoute itself, Micro-F1, Macro-F1, and RadGraph all reach their best observed values at 192 and decline at 256. Thus 192 is the best observed operating point for the evaluated ExpertRoute policy, while the preferred evidence-composition rule varies with context budget.

4.5 Complementary Expert–Semantic Composition

The frozen 2,706-case Validation routing summary directly characterizes the 192-token composition. The expert branch contributes 78.24 tokens on average and the independent semantic branch adds 113.76; composition examines 125.83 semantic candidates and skips 12.07 overlaps per case (9.57%). Thus the expert set is preserved while semantic ranking supplies genuinely new indices until the exact budget is filled.

The resulting selection remains distinct from either component route: mean Jaccard overlap is 0.521 with SCOPE, 0.304 with GazeRoute, and 0.857 with GazeSemanticRoute, localizing the remaining difference to independent versus gaze-conditioned semantic ranking. At the same 192-token budget, ExpertRoute improves Validation Micro-F1, Macro-F1, and RadGraph over both GazeRoute and

SCOPE. Additional intersection statistics and frontal-view diagnostics are reported in Appendix B and Appendix C.3.

4.6 Quality–efficiency trade-off

ExpertRoute reduces the LLM visual sequence by 86.0%, from 1,369 to 192 tokens. On the same 200-case Validation manifest, three matched repetitions yield mean generation latencies of 3,150.9 ms for Full and 2,640.7 ms for ExpertRoute, a 16.2% reduction (1.19× speedup); the reduction is observed in all three runs. The synchronized timing includes vision encoding, projection, online routing, and LLM generation. complete systems measurements appear in Appendix D.1.

5 Discussion

The experiments reveal a budget-dependent quality response rather than a universal ordering among routing rules. The reversal across context sizes is consistent with a change in the role of redundancy control: independent ranking can preserve a second evidence source at moderate budgets, whereas gaze-conditioned coverage may become more useful as the expert allocation occupies more of the context. We therefore interpret 192 as the best observed operating point for the evaluated ExpertRoute policy, not as a universal optimum.

Compact routing also changes the quality profile rather than uniformly improving every metric. At the principal setting, report-level RadGraph and GREEN improve over Full across all three cohorts, while finding-level CheXbert F1 remains lower on the locked cohort. Per-finding changes are heterogeneous rather than systematically concentrated in focal abnormalities: pneumothorax and consolidation improve, pleural effusion is essentially unchanged, and the largest reductions occur for atelectasis, fracture, and lung opacity (Appendix C.1). These automated-label results do not establish clinical safety, but they argue against a simple interpretation of uniform localized-evidence loss.

Finally, the gaze scorer is an offline expert-prior distillation module rather than a persistent annotation requirement. It is trained once on REFLACX and frozen thereafter, so inference uses only the image. The lateral-view analysis shows that the same frozen route retains the overall report-level trade-off outside the frontal gaze-supervision regime (Appendix C.3); this is a robustness observation rather than causal evidence of gaze transfer across views.

6 Limitations

The gaze scorer is learned from REFLACX and is most directly aligned with frontal chest radiographs, broader view-position and backbone transfer remain important extensions. Because ExpertRoute routes after the multimodal projector, compute savings are concentrated in the LLM visual context. The semantic selector also precomputes the full pairwise similarity matrix; incremental or lazy-greedy variants could reduce routing overhead. Gaze-target/encoder alignment is approximate but quantified in Appendix A.4, and automatic metrics do not replace prospective radiologist evaluation.

7 Conclusion

ExpertRoute distills radiologists' visual-search behavior into a reusable expert prior and composes it with independent semantic coverage to construct a compact visual evidence set for chest X-ray report generation. Its strongest observed multi-metric operating point uses 192 of 1,369 visual tokens. At this setting, ExpertRoute achieves higher RadGraph and GREEN than full-context inference on MIMIC Validation, locked MIMIC held-out, and OpenI, while reducing per-sample generation latency by 16.2%; the held-out report-level gains over Full are supported by patient-cluster bootstrap. Together with the budget sweep and controlled routing variants, these results show that the composition of visual evidence matters as much as the amount of context retained.

REPRODUCIBILITY PROTOCOL

The principal evaluated configuration uses a 192-token visual-context budget with both the GazeScorer and report generator frozen. All held-out comparisons use aligned case IDs and reference reports, and all method-specific implementations are evaluated with the same report-generation prompt. At the principal operating point, ExpertRoute produces exactly 192 unique valid visual-token indices for every evaluated sample. The budget analyses use the corresponding frozen evaluation results without post-hoc retuning or score substitution. Fixed-budget methods are compared with the same 192-token visual context entering the LLM, while progressive SparseVLM+ is reported separately because its token count changes across layers.

AI USE STATEMENT

Generative AI tools were used as research-assistance tools for literature discovery and organization, manuscript drafting and language editing. The tools were not treated as an autonomous source of empirical evidence: all reported quantitative results and statistical analyses were obtained from the authors' experimental pipeline and checked against the underlying outputs. All AI-assisted suggestions were reviewed and revised by the authors, who made the final decisions regarding the method, experiments, claims, and presentation and take full responsibility for the content of this work.

REFERENCES

Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F. Auffermann, Jessica Chan, Phuong-Anh T. Duong, Vivek Srikumar, Trafton Drew, Joyce D. Schroeder, and Tolga Tasdizen. REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific Data*, 9(1):350, 2022. doi: 10.1038/s41597-022-01441-z.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Qingqing Cao, Bhargavi Paranjape, and Hannaneh Hajishirzi. PuMer: Pruning and merging tokens for efficient vision language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12890–12903. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.721.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision – ECCV 2024*, pp. 19–35, 2024. doi: 10.1007/978-3-031-73004-7_2.

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

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5904–5914. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.459.

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

Jinhong Deng, Wen Li, Joey Tianyi Zhou, and Yang He. SCOPE: Saliency-coverage oriented token pruning for efficient multimodel LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5394.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *Computer Vision – ECCV 2022*, pp. 396–414, 2022. doi: 10.1007/978-3-031-20083-0_24.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3922–3931. IEEE, 2021.

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

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Subathra Adithan, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pp. 978–990. PMLR, 2024.

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

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. doi: 10.1038/s41597-019-0322-0.

Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy Wu, Matthew Tong, Arjun Sharma, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth Krupinski, and Mehdi Moradi. Eye gaze data for chest x-rays. *PhysioNet*, 2020. doi: 10.13026/qfdz-zr67. Version 1.0.0.

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

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

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.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5288–5304. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.416.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 2023.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.21.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13937–13949, 2021.

Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12786–12797, 2021.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025. doi: 10.1109/ICCV51701.2025.02122.

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

Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X. Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multi-modal LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7614–7623. Association for Computational Linguistics, 2025.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025.

Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80:102510, 2022. doi: 10.1016/j.media.2022.102510.

Hongxu Yin, Arash Vahdat, Jose M. Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10809–10818, 2022. doi: 10.1109/CVPR52688.2022.01054.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025. doi: 10.1038/s41467-025-58344-x.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu-Hsin Wei, Tristan Naumann, Muhao Chen, Matthew Lungren, Akshay Chaudhari, Serena Yeung, Curtis Langlotz, Sheng Wang, and Hoifung Poon. LLaVA-Rad MIMIC-CXR Annotations. *PhysioNet*, January 2025. doi: 10.13026/4ma4-k740. URL `https://doi.org/10.13026/4ma4-k740`. Version 1.0.0.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20857–20867, 2025a. doi: 10.1109/ICCV51701.2025.01939.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-0855.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2025c. doi: 10.1056/AIoa2400640.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74840–74857. PMLR, 2025d.

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 *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 2–25. PMLR, 2022.

A Implementation and evaluation protocol

This section collects the implementation details of the expert-prior scorer together with the evaluation and statistical conventions used throughout the paper. The scorer is trained only on REFLACX and is frozen before all downstream routing and report-generation experiments.

A.1 GazeScorer architecture

Figure 4 illustrates the five-source gaze scorer described in Section 3.2.

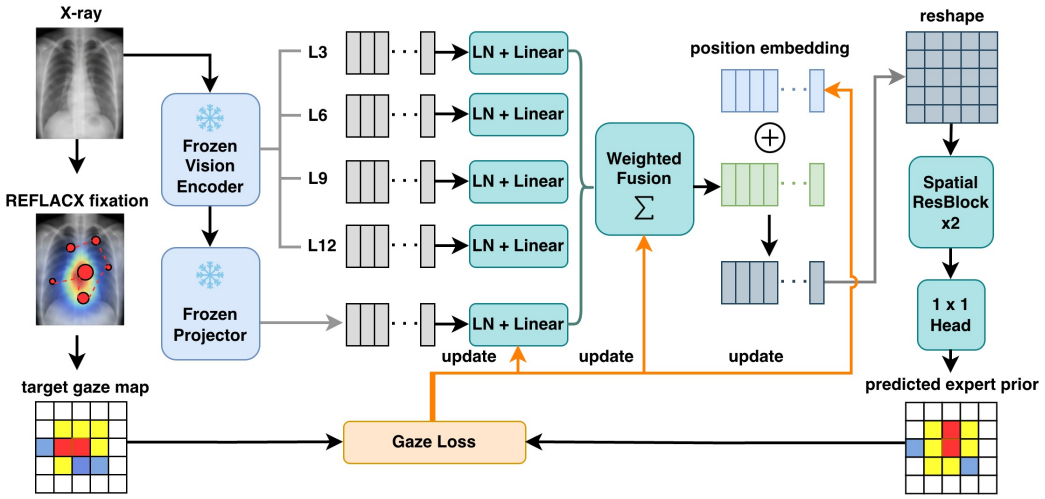

Figure 4: **Learning an image-conditioned expert prior.** Four vision hidden states and projected features are adapted to a shared 256-dimensional space, fused with learned softmax weights, and augmented with position. Two internally residual spatial blocks and a prediction head produce gaze logits. The scorer is trained offline and frozen for report generation.

A.2 Gaze supervision and training objective

Gaze target construction. The pixel heatmap uses an axis-aligned Gaussian for each valid fixation, with standard deviations $\sigma_x = 1.5\eta_x$ and $\sigma_y = 1.5\eta_y$ in pixels. Here η_x, η_y are the pixels-per-degree values used by the target builder; missing values use its fallback of 50. Contributions are evaluated within three standard deviations and weighted by clipped fixation duration. Invalid or out-of-image fixations are skipped. The pixel-to-patch reduction is implemented with OpenCV area interpolation (`INTER_AREA`) to produce the 37×37 supervision grid, followed by normalization when nonzero; a record with no valid fixation produces a zero map.

Target processing and ranking supervision. Let u_i be the nonnegative heatmap value at patch i after the `INTER_AREA` reduction to the 37×37 training grid. A later generic training utility uses bilinear interpolation with `align_corners=False` only when a target-shape adjustment is required; in the frozen A5 configuration the target and scorer output are already 37×37, so this utility does not define an additional pixel-to-patch resampling step. The processor constructs the KL target $q_i = u_i/(\sum_j u_j + \epsilon)$ and the ranking target $t_i = (u_i - \min_j u_j)/(\max_j u_j - \min_j u_j + \epsilon)$, where $\epsilon > 0$ is a numerical stabilizer. Invalid targets are excluded by the training validity mask. KL is computed from log-softmax logits and $\mathbf{q}$, summing across patches and averaging across the valid batch.

For image b, define $P_b = \{i : t_i \geq 0.2 \max_j t_j\}$ and $N_b = \{i : t_i \leq 0.05 \max_j t_j\}$. For nondegenerate targets with both sets nonempty, positive and negative indices are sampled independently with replacement using the device RNG, forming a pair list $\mathcal{P}_b$ of length $\min(5000, |P_b||N_b|)$. The

ranking scores are

$$\tilde{g}_i = \frac{g_i - \min_j g_j}{\max_j g_j - \min_j g_j + \epsilon}, \qquad \mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{B}_+|} \sum_{b \in \mathcal{B}_+} \frac{1}{|\mathcal{P}_b|} \sum_{(i,j) \in \mathcal{P}_b} \max(0, 0.1 - \tilde{g}_i + \tilde{g}_j), \quad (9)$$

where $\mathcal{B}_+$ contains batch elements with valid pairs. Empty-pair samples are skipped, and the ranking term is zero when $\mathcal{B}_+$ is empty. The frozen configuration uses $\lambda_{\text{rank}} = 0.1$.

A.3 Optimization and training configuration

The final A5 GazeScorer is optimized with AdamW using an initial learning rate of 1×10^{-4}, weight decay 1×10^{-5}, and batch size 8 for at most 50 epochs. A ReduceLROnPlateau scheduler monitors validation loss with reduction factor 0.5 and patience 5. Gradients are clipped to a maximum norm of 1.0, and the spatial residual blocks use dropout probability $p = 0.2$. Early stopping uses the validation sAUC criterion with patience 10. Training uses random seed 42 and no data augmentation. The LLaVA-Rad backbone, including the vision encoder and multimodal projector, remains frozen throughout; only the 1.99M-parameter GazeScorer is optimized.

A.4 Checkpoint selection and target alignment

Gaze-domain evaluation and checkpoint selection. For a nondegenerate evaluator prediction a and min–max-normalized target t, define $s_i = (a_i - \min_j a_j)/\sum_k (a_k - \min_j a_j)$ and $q_i^{\text{eval}} = t_i / \sum_j t_j$. Similarity is $\text{SIM} = \sum_i \min(s_i, q_i^{\text{eval}})$. The Top-192 mass is the target probability mass contained in the 192 highest-scoring predicted patches. Metrics are computed per image. The checkpoint used for all downstream experiments is epoch 10, which attains the highest validation sAUC (0.6215); the preserved REFLACX test split is not used for model selection.

Spatial alignment audit. Gaze heatmaps are reduced directly from image coordinates to the patch grid, whereas encoder inputs additionally undergo center cropping, yielding an approximate alignment. A retrospective audit over 307,900 valid fixations found a median displacement of 0.709 patch (P90: 2.083 patches). Because the gaze supervision is spatially smoothed, the original and geometry-corrected density maps remain highly concordant (median Pearson $r = 0.956$; cosine similarity $= 0.969$), and the median displacement is 0.232 times the effective Gaussian width. All reported experiments use the frozen original targets.

A.5 GazeScorer architecture ablation

The controlled REFLACX Validation ablation in Table 3 isolates the contributions of the GazeScorer components. Architecture variants are compared using KL divergence (KL), correlation coefficient (CC), similarity (SIM), and Top-192 target mass, which characterize distributional agreement and coverage of the highest-ranked patches. Multi-layer fusion improves KL and similarity over the token-MLP baseline, and spatial refinement provides a further gain. Explicit positional encoding is most useful when combined with hierarchical and spatial features, yielding the final A5 architecture used in all downstream routing experiments.

Table 3: GazeScorer architecture ablation on REFLACX Validation. Architecture variants are compared using distributional agreement and top-ranked target-coverage metrics. **Bold** and underline denote best and second-best values.

ID	Architecture	Params	KL↓	CC↑	SIM↑	Top-192↑
A0	token MLP	2.10M	0.339	0.721	0.696	0.444
A1	A0 + position	2.10M	0.354	0.708	0.689	0.436
A2	multi-layer fusion	1.98M	0.315	0.739	0.707	0.452
A3	2-D spatial residual	1.19M	0.323	0.733	0.704	0.450
A4	multi-layer + spatial	1.99M	<u>0.311</u>	**0.746**	<u>0.708</u>	<u>0.456</u>
A5	+ position (full)	1.99M	**0.310**	<u>0.745</u>	**0.713**	**0.457**

A.6 Downstream evaluation setup and baseline integration

REFLACX train and validation records are pooled and repartitioned by patient with seed 42 and target training fraction 0.85; the official test partition is preserved. Records flagged as discarded eye tracking are excluded. All experiments are conducted on NVIDIA GeForce RTX 3090 GPU.

For the external baselines, CDPruner and VisPruner follow their released algorithms at a fixed 192-token budget without dataset-specific tuning. VisPruner uses $r = 0.5$ with equal importance/diversity allocation. CDPruner uses projected visual features and a mean instruction-token embedding for conditional relevance followed by greedy MAP-DPP selection; VisPruner uses reconstructed vision-layer CLS-to-patch attention. SparseVLM+ follows the recorded progressive 1,369→205→136→80 schedule.

A.7 Evaluation metrics and statistical inference

Finding-level metrics. Main-table CheXbert scores use the 14-label mapping in which positive is 1 and negative, uncertain, and blank are 0. The supplementary "+" variants additionally map uncertain to 1. The five-label subset comprises cardiomegaly, edema, consolidation, atelectasis, and pleural effusion. Undefined precision/recall contributions use the evaluator's zero-division value of zero.

Report metrics and aggregation. RadGraph is evaluated using F1-RadGraph with partial reward. Per-record precision, recall, and F1 are computed first and then averaged separately; the reported mean F1 need not equal the harmonic mean of mean precision and mean recall. The held-out analysis averages over case records, with the same records and references used for all methods. Patient-cluster bootstrap resamples patients with replacement and concatenates all their records, including repetitions when a patient is drawn multiple times. CheXbert F1 is recomputed from the resampled labels, whereas RadGraph and GREEN use the resampled frozen per-record scores. The 5,000 replicates use seed 20260826 and percentile 2.5/97.5 limits.

B Routing configurations and budget analysis

This section specifies the exact allocation rules used by the four budget-swept routes and then reports diagnostics that characterize how their evidence sets differ across budgets.

B.1 Allocation rules

The evaluated budgets are $B \in \{32, 64, 128, 192, 256\}$. GazeRoute uses a Top-32 gaze-anchor candidate ranking, radius 1, no separate gaze cap, and 4×4 stratified spatial completion. SCOPE selects B patches from empty-set semantic coverage. Both GazeSemanticRoute and ExpertRoute likewise start from Top-32 gaze-anchor candidates and radius 1, with respective gaze caps 16, 32, 64, 96, and 128. Anchor expansion is truncated at the cap before completion, so Top-32 candidates do not imply 32 retained anchors. GazeSemanticRoute initializes coverage with the gaze set; ExpertRoute traverses an independently initialized ranking and skips exact-index overlaps. All routes restore final raster order. The budget sweep evaluates these allocation policies; cross-budget token sets are not assumed to be nested.

For GazeRoute at $B = 192$, each of the 16 spatial strata supplies center-out candidates in round-robin order. Each candidate expands to a radius-1 neighborhood and already selected indices are skipped. If a neighborhood exceeds the remaining budget, its new indices are sorted by descending gaze score, with patch index breaking ties, and truncated. A global descending-gaze fallback fills any remaining capacity. Final indices are restored to raster order. This spatial completion does not use semantic features, SCOPE saliency, or cosine coverage.

For semantic selection, greedy score ties are broken by the smallest patch index. The implementation precomputes the full $N \times N$ cosine matrix and materializes a budget-limited prefix. In GazeSemanticRoute, coverage is initialized from S_g via $m_u = \max_{s \in S_g} \cos(z_u, z_s)$, candidates are scored by $A_v \sum_u \max(0, \cos(z_u, z_v) - m_u)$, and m_u is updated after each choice; the initial maximum is not clamped to zero.

B.2 Routing composition diagnostics

At $B = 192$, the expert branch contributes 78.24 tokens on average and the semantic branch adds 113.76. Composition examines 125.83 semantic candidates and skips 12.07 overlapping indices per case (mean skip ratio 9.57%). The final set has mean intersection/Jaccard of 131.15/0.521 with SCOPE, 89.30/0.304 with GazeRoute, and 177.11/0.857 with GazeSemanticRoute, supporting partial rather than complete overlap between the evidence sources.

On the 1,835-image frontal Validation subset, ExpertRoute attains Micro-F1 0.5691, RadGraph 0.2747, and GREEN 0.4024.

B.3 Complete budget sweep

Table 4 reports the complete primary clinical metrics used for the budget analysis. Each row reports the frozen result for its corresponding method and budget.

Table 4: MIMIC Validation quality across visual-token budgets. Bold and underline denote best and second-best point estimates within each budget.

Metric	B	SCOPE	GazeRoute	GazeSemanticRoute	ExpertRoute
Micro-F1↑	32	**0.4888**	0.3441	<u>0.4707</u>	0.4679
	64	**0.5254**	0.4097	<u>0.5252</u>	0.5101
	128	**0.5448**	0.4658	<u>0.5431</u>	0.5419
	192	<u>0.5412</u>	0.5305	0.5354	**0.5464**
	256	<u>0.5405</u>	0.5234	**0.5466**	0.5401
Macro-F1↑	32	**0.3181**	0.2487	<u>0.3068</u>	0.3034
	64	<u>0.3399</u>	0.2871	**0.3535**	0.3364
	128	0.3500	0.3143	**0.3667**	<u>0.3559</u>
	192	0.3562	<u>0.3614</u>	0.3529	**0.3620**
	256	0.3398	0.3503	**0.3614**	<u>0.3581</u>
RadGraph F1↑	32	**0.2782**	0.2308	<u>0.2716</u>	0.2703
	64	<u>0.2842</u>	0.2425	**0.2862**	0.2821
	128	**0.2881**	0.2652	0.2850	<u>0.2862</u>
	192	0.2823	0.2807	<u>0.2871</u>	**0.2875**
	256	0.2767	<u>0.2839</u>	**0.2842**	0.2832

C Diagnostic analyses

C.1 Locked held-out diagnostics

The following analyses complement the primary held-out results with alternative finding-level metrics (Table 5), per-finding behavior, and RadGraph precision–recall decomposition (Table 7).

Per-finding CheXbert analysis. To examine whether the aggregate finding-level drop reflects a systematic loss of focal abnormalities, Table 6 compares Full and ExpertRoute by label. The changes are heterogeneous: pneumothorax and consolidation improve, pleural effusion is essentially unchanged, while the largest reductions occur for atelectasis, fracture, and lung opacity. Because several labels have limited support, these values are descriptive and are not used for significance claims.

Table 5: Supplementary CheXbert metrics on the locked held-out cohort ($N = 580$). Bold and underline rank the listed compact-context methods only; Full is a reference.

Method	Micro-14+↑	Macro-14+↑	Micro-5↑	Macro-5↑	Micro-5+↑	Macro-5+↑
LLaVA-Rad (Full)	0.5587	0.4324	0.5683	0.4770	0.5939	0.5208
Model-derived compact-context baselines						
SCOPE	**0.5450**	0.4024	0.5511	0.4534	**0.5972**	**0.5124**
SparseVLM+	0.4190	0.3310	0.3210	0.2811	0.3331	0.2920
CDPruner	0.5148	0.3749	0.5180	0.4286	0.5588	0.4998
VisPruner	<u>0.5306</u>	**0.4173**	**0.5644**	<u>0.4695</u>	0.5741	0.5025
Gaze-supervised routing variants						
GazeRoute	0.5190	0.4026	0.5429	0.4627	0.5703	0.5015
GazeSemanticRoute	0.5272	0.3952	<u>0.5571</u>	**0.4857**	<u>0.5787</u>	<u>0.5097</u>
ExpertRoute	0.5267	<u>0.4030</u>	0.5421	0.4667	0.5650	0.5015

Table 6: Held-out CheXbert performance by finding for Full and ExpertRoute. Recall is included to make sensitivity changes visible without overloading the table. ΔF1 is ExpertRoute minus Full.

Finding	Support	Full R	Full F1	ExpertRoute R	ExpertRoute F1	ΔF1
Enlarged Cardiomediastinum	46	0.0652	0.0800	0.0652	0.0896	+0.0096
Cardiomegaly	193	0.6839	0.6316	0.6684	0.6277	-0.0038
Lung Opacity	241	0.4979	0.5195	0.3485	0.4190	-0.1005
Lung Lesion	26	0.1923	0.2632	0.2308	0.2553	-0.0078
Edema	122	0.3852	0.4700	0.4180	0.4397	-0.0303
Consolidation	32	0.0938	0.1200	0.2188	0.2258	+0.1058
Pneumonia	28	0.2500	0.2642	0.1786	0.2128	-0.0514
Atelectasis	150	0.5933	0.4986	0.3400	0.3750	-0.1236
Pneumothorax	18	0.6111	0.5238	0.6111	0.5500	+0.0262
Pleural Effusion	210	0.6619	0.6651	0.6905	0.6651	+0.0001
Pleural Other	22	0.0455	0.0625	0.0909	0.1143	+0.0518
Fracture	31	0.2258	0.2917	0.1290	0.1905	-0.1012
Support Devices	241	0.8299	0.8180	0.8548	0.8306	+0.0126
No Finding	26	0.5769	0.3704	0.4231	0.2933	-0.0770

Table 7: RadGraph precision and recall on the locked held-out cohort. Bold and underline rank the listed compact-context methods only; Full is a reference.

Method	Precision↑	Recall↑	F1↑
LLaVA-Rad (Full)	0.2512	0.2193	0.2245
Model-derived compact-context baselines			
SCOPE	0.2744	0.2480	0.2505
SparseVLM+	0.2557	0.2076	0.2188
CDPruner	0.2657	0.2378	0.2416
VisPruner	**0.2820**	0.2447	0.2513
Gaze-supervised routing variants			
GazeRoute	0.2713	0.2407	0.2454
GazeSemanticRoute	<u>0.2801</u>	<u>0.2514</u>	<u>0.2543</u>
ExpertRoute	0.2794	**0.2550**	**0.2557**

C.2 OpenI finding-level metrics

Alternative CheXbert mappings provide a consistency check on the main external finding-level evaluation (Table 8).

Table 8: Supplementary CheXbert metrics on OpenI ($N = 3285$). Bold and underline rank the listed compact-context methods only; Full is a reference.

Method	Micro-14+↑	Macro-14+↑	Micro-5↑	Macro-5↑	Micro-5+↑	Macro-5+↑
LLaVA-Rad (Full)	0.4865	0.3113	0.4294	0.3659	0.4164	0.3580
Model-derived compact-context baselines						
SCOPE	**0.4889**	**0.3058**	**0.4403**	**0.3807**	**0.4237**	**0.3712**
VisionZip	0.4533	0.2721	<u>0.4385</u>	0.3381	0.4103	0.3254
SparseVLM+	0.4373	0.2645	0.3061	0.2534	0.2888	0.2276
CDPruner	0.4625	0.2967	0.3523	0.2901	0.3523	0.3106
VisPruner	0.4537	0.2670	0.3635	0.2836	0.3538	0.2958
Gaze-supervised routing variants						
GazeRoute	0.4608	0.2898	0.4018	0.3181	0.3865	0.3239
GazeSemanticRoute	0.4713	<u>0.3031</u>	0.4139	0.3413	0.4022	0.3480
ExpertRoute	<u>0.4724</u>	0.3025	0.4133	<u>0.3489</u>	<u>0.4144</u>	<u>0.3576</u>

C.3 View-position robustness

MIMIC Validation contains 1,835 frontal and 871 lateral images. Because gaze supervision is frontal, we report the lateral subset separately rather than treating it as direct validation of gaze prediction. Table 9 shows that ExpertRoute retains the same qualitative trade-off observed in the main evaluation: finding-level F1 remains below Full, while RadGraph and GREEN are higher. The locked held-out cohort is entirely frontal, so no held-out lateral analysis is available.

Table 9: Lateral-only MIMIC Validation results ($N = 871$), computed by subsetting frozen per-sample outputs; no model inference was rerun.

Method	Micro-F1	Macro-F1	RadGraph F1	GREEN
LLaVA-Rad (Full)	0.4937	0.3018	0.2833	0.4721
SCOPE	**0.4781**	0.2765	**0.3151**	**0.4916**
VisionZip	0.3908	0.2070	0.2124	0.2857
SparseVLM+	0.4186	0.2705	0.2961	0.4625
GazeRoute	0.4680	0.2878	0.3048	0.4799
GazeSemanticRoute	0.4654	<u>0.2904</u>	<u>0.3147</u>	0.4873
ExpertRoute	<u>0.4759</u>	**0.2921**	<u>0.3147</u>	<u>0.4875</u>

D Systems evaluation

The following measurements document the realized systems behavior of the evaluated implementations and the protocol used to keep the principal comparisons aligned.

D.1 Efficiency measurement details

All methods use the same 200-case Validation manifest and matched generation protocol on an NVIDIA GeForce RTX 3090. Each method/repeat runs in an isolated subprocess with a 10-case warmup. CUDA synchronization immediately precedes and follows `model.generate()`; model loading and outer file I/O are excluded. All three repetitions show lower generation latency for ExpertRoute than Full. For VisionZip, routing logs confirm exact 1,369-to-192-token compression for every evaluated sample. Its end-to-end latency nevertheless exceeds Full in our LLaVA-Rad adaptation. Since the benchmark measures the complete generation path and cross-method stage

costs are not isolated, we treat this as a realized property of the evaluated integration rather than an intrinsic efficiency claim about the original VisionZip implementation.

Peak allocated memory is reset per sample and measured over the profiling runner's allocation window, which includes model residency and any enabled replay allocations. Table 10 therefore reports recorded system-level aggregates.

Table 10: Recorded systems aggregates on 200 Validation cases with three repetitions. Peak MiB includes resident allocations and any enabled replay profiling; it is not generation-only memory. Cross-method stage costs are not isolated.

Method	LLM visual tokens	Latency (ms)↓	Speedup↑	Peak MiB↓
LLaVA-Rad (Full)	1,369	3150.9	1.00×	23687.6
Model-derived compact-context baselines				
SCOPE	192	2441.5	1.29×	14025.7
VisionZip	192	5538.6	0.57×	13615.5
SparseVLM+	1,369→205→136→80	2988.6	1.05×	14672.3
Gaze-supervised routing variants				
GazeRoute	192	2326.5	1.35×	14081.4
GazeSemanticRoute	192	2504.4	1.26×	14087.1
ExpertRoute	192	2640.7	1.19×	14090.5