C3-Net: Gaze-Guided Multimodal Learning for Classification and Report Generation

Anonymous ACL submission

Abstract

Radiology report interpretation demands simultaneous visual attention, spatial reasoning, and clinical language understanding — capabilities that existing AI systems address in isolation. We present **C3-Net** (Cognitive Causal Chain Network), a two-stage multimodal framework for chest X-ray classification and report generation following the expert cognitive pathway of Image → Gaze → Text → Diagnosis. In Stage 1, C3-Net trains a multimodal classifier on chest X-ray images (ViT-Base/16), expert radiologist gaze fixations (CNN+BiLSTM), and clinical transcriptions (MedGemma-4B). A dedicated `GazePredictor` reconstructs expert spatial attention from image features alone, enabling gaze-guided inference without eye-tracking at test time. Cross-modal grounding is achieved via a `TextImageAlignment` module employing cross-attention over ViT patch embeddings.

In Stage 2, frozen Stage 1 representations condition a MedGemma decoder for report generation. Evaluated on MIMIC-Eye across four modality configurations, C3-Net achieves an AUC of 0.952, outperforming image-only by +11.2% ($p < 0.001$, DeLong's test). Report generation reaches a CheXBERT similarity score of 0.72, with gaze masking ablations confirming that spatially guided attention is critical for clinically accurate generation.

1 Introduction

Chest X-ray interpretation is among the most common and consequential radiological tasks, yet it remains highly demanding. Experienced radiologists direct visual attention to diagnostically relevant regions, integrate prior clinical context, and produce structured findings in natural language, often under significant time pressure. AI systems capable of replicating this reasoning pipeline could reduce reporting burden and improve diagnostic consistency, particularly in under-resourced settings.

Two complementary research directions have made meaningful progress toward this goal. Gaze-guided learning has shown that radiologist fixation patterns encode diagnostically meaningful spatial priors that improve classification beyond what image supervision alone provides (Ma et al., 2023; Wang et al., 2024). Vision-language models pretrained on large biomedical corpora provide strong text encoders for clinical report understanding and generation (Sellergren et al., 2025). However, these directions have largely been pursued independently. Gaze-guided methods typically require eye-tracking data at inference time, an impractical constraint for clinical deployment, while vision-language approaches do not incorporate explicit expert spatial attention supervision. Furthermore, existing systems typically address classification or generation in isolation, without a unified framework that achieves both.

We present **C3-Net** (Cognitive Causal Chain Network), a two-stage multimodal framework that addresses these limitations directly. In Stage 1, C3-Net trains a multimodal classifier jointly on image, gaze, and text modalities alongside a `GazePredictor` that learns to reconstruct expert attention maps from image features alone, eliminating the inference-time gaze requirement without a separate student network. In Stage 2, frozen Stage 1 representations condition a MedGemma-4B decoder for radiology report generation. MedGemma serves as a unified backbone for both text encoding (Stage 1) and report generation (Stage 2), with its built-in SigLIP vision tower replaced by our gaze-weighted ViT pipeline.

Our main contributions are:

- **GazeGuidedFusion:** a module combining gaze-normalized patch pooling with learnable cross-attention using gaze temporal features as queries, producing spatially grounded image representations aligned with radiologist

fixation patterns.

- **GazePredictor:** a lightweight MLP trained with MSE loss to reconstruct expert gaze heatmaps from mean-pooled ViT patch features, enabling image-only inference through the full gaze-guided pipeline.
- **TextImageAlignment:** a cross-attention module that grounds clinical text tokens to ViT patch regions, learning which diagnostic terms correspond to which image areas.
- **Unified MedGemma backbone:** MedGemma-4B serves as both the text encoder for multimodal classification and the generation backbone for report synthesis, with comprehensive evaluation on MIMIC-Eye across four modality configurations validated via DeLong's test over five random seeds.

2 Related Work

Gaze-Guided Medical Image Analysis: Eye-tracking by radiologists has emerged as a productive supervisory signal throughout the full learning pipeline. Early work demonstrated that fixation patterns could supervise patch selection in Vision Transformers to correct shortcut learning (Ma et al., 2023), while GazeGNN (Wang et al., 2024) constructed unified graph embeddings over gaze and image patches for real-time inference without pre-generated attention maps. Wang et al. (Wang et al., 2025) further pushed gaze supervision upstream, aligning model representations with human visual attention at the self-supervised pre-training stage. Despite these advances, all prior methods require gaze data at inference time or rely on a separate student network for deployment. C3-Net addresses this directly — a dedicated `GazePredictor` reconstructs expert attention from image features alone, enabling full gaze-guided classification at inference without any eye-tracking input.

Multimodal Alignment in Medical Imaging: Contrastive vision-language pre-training (Radford et al., 2021) has been widely adapted to the medical domain via ConVIRT (Zhang et al., 2022), MedCLIP (Wang et al., 2022), BioViL (Boecking et al., 2022), and GLoRIA (Huang et al., 2021). EGMA (Ma et al., 2024) extended this by incorporating gaze heatmaps as an auxiliary alignment signal during contrastive pre-training, achieving state-of-the-art results on CheXpert and RSNA. Our work differs in adopting a two-stage framework where gaze supervision is embedded into the classification backbone rather than the pre-training objective, and cross-modal text-image grounding follows ALBEF (Li et al., 2021), which aligns vision and language representations via momentum distillation before fusion.

Knowledge Distillation for Multimodal AI: Knowledge distillation (Hinton et al., 2015) has been applied in the medical domain to transfer from multimodal teachers to unimodal students (MIND (Guerra-Manzanares and Shamout, 2025)) and across specialist heads for chest X-ray classification (Song et al., 2025). MST-KDNet (Zhu et al., 2025) addressed missing modality scenarios via multi-scale transformer distillation. C3-Net eliminates the need for a separate student network entirely through a learned `GazePredictor`, enabling full three-modality training with image-only inference.

Radiology Report Generation: Report generation from chest X-rays has seen rapid progress, from memory-driven transformers (Chen et al., 2020) to grounded generation linking sentences to anatomical regions (MAIRA-2 (Bannur et al., 2024)) and inter-report consistency improvements (ICON (Hou et al., 2024)). A persistent challenge is that BLEU-4 and ROUGE-L fail to reflect clinical accuracy. We therefore adopt CheXBERT (Smit et al., 2020) and BERTScore (Zhang et al., 2019) as primary generation metrics. We use MedGemma-4B (Sellergren et al., 2025) as both text encoder and generation backbone, discarding its built-in vision tower in favour of our gaze-weighted ViT.

3 Dataset

3.1 MIMIC-Eye

We use MIMIC-Eye (Hsieh et al., 2023), a multimodal dataset that integrates MIMIC-CXR (Johnson et al., 2024) chest X-rays with expert radiologist gaze recordings and transcriptions from two sources, REFLACX (Bigolin Lanfredi et al., 2022) and EyeGaze (Karargyris et al., 2020). The raw dataset contains 3,192 patient folders, 6,292 chest X-ray images, and 3,052 REFLACX reading sessions.

3.2 Preprocessing and Filtering

Each sample requires all three modalities, image, gaze, and text, to be present. Samples missing

Source	Patients	With Fixations	With Transcription
REFLACX only	2,154	3,032	3,032
EyeGaze only	993	1,038	1,080
Both	45	—	—

Table 1: Dataset composition across REFLACX and EyeGaze sources.

fixations (20) or transcriptions (5) are discarded, yielding 4,064 usable samples.

Images: Images are resized to 224×224 and normalized using ImageNet statistics.

Gaze: Gaze fixations are converted into two representations:

- A patch-aligned spatial heatmap, constructed by placing a Gaussian at each fixation point weighted by duration, smoothed, normalized to $[0, 1]$, and downsampled to the ViT patch grid via 16×16 block averaging.
- A temporal fixation sequence padded to 50 fixations, where each fixation is encoded as $(x, y, \tilde{d})$, with $x, y \in [0, 1]$ denoting normalized spatial coordinates and $\tilde{d}$ the log-normalized fixation duration.

$$\tilde{d} = \frac{\log(1 + d)}{\log(1 + 5000)} \tag{1}$$

Text: Text transcriptions consist of radiologist dictations in clinical shorthand (e.g., *"there is mild bibasilar subsegmental atelectasis. no effusions or pneumothorax. no acute osseous abnormality."*) and are tokenized using the MedGemma tokenizer to a maximum length of 256 tokens.

Labels: Labels are binary (Normal=0, Abnormal=1) and derived from REFLACX pathology annotations and EyeGaze metadata.

3.3 Data Splits

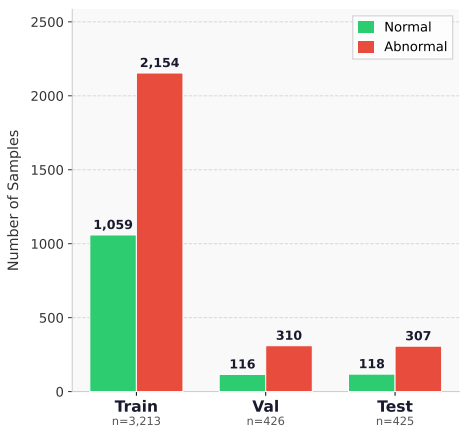

Figure 1: Class distribution across dataset splits.

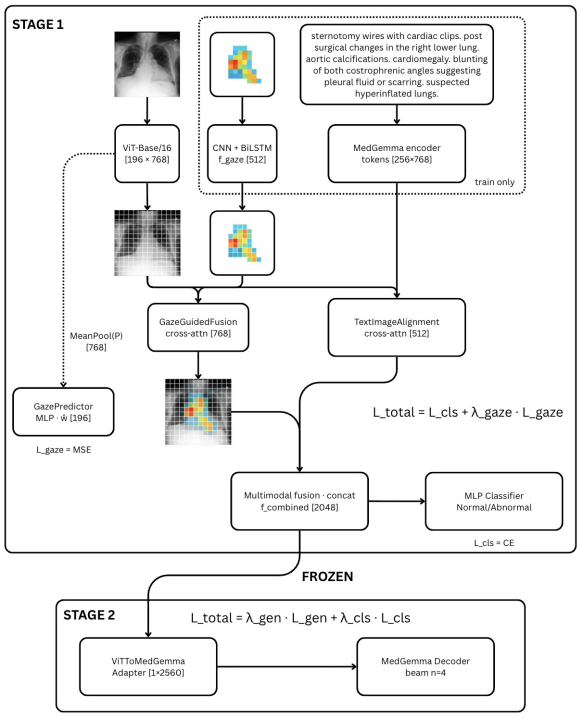

Figure 2: C3-Net architecture.

The dataset is imbalanced and it is addressed during training using inverse-frequency weighted cross-entropy loss (He and Garcia, 2009), where each class weight is computed as:

$$w_c = \frac{1}{N_c} \tag{2}$$

with N_c denoting the number of training samples in class c. AUC is used as the primary evaluation metric due to this imbalance.

4 Methodology

We present **C3-Net** (Cognitive Causal Chain Network), a multimodal framework for chest X-ray classification and radiology report generation. The system follows the expert cognitive pathway of *Image* $\rightarrow$ *Gaze* $\rightarrow$ *Text* $\rightarrow$ *Diagnosis*, and operates in two stages. Stage 1 trains a multimodal classifier with joint gaze prediction. Stage 2 freezes Stage 1 and fine-tunes a MedGemma decoder for report generation. A deployment constraint is that only images are available at inference time. Gaze and

text are used exclusively during training. Radiologist transcriptions serve dual roles: as text encoder input during Stage 1 multimodal classification, and as generation targets during Stage 2 decoder training.

4.1 Image Encoder

We use ViT-Base/16 (Dosovitskiy et al., 2020) with ImageNet pretrained weights, splitting each image into $N = 196$ non-overlapping 16×16 patches. The encoder produces two outputs:

$$\text{ViT}(x) = [f_{\text{cls}}, P] \tag{3}$$

where $f_{\text{cls}} \in \mathbb{R}^{768}$ is the CLS token (global image summary) and $P \in \mathbb{R}^{196\times768}$ are the patch token features, used downstream for gaze-guided attention and text-image alignment, respectively. Full parameter counts and compute details are provided in Appendix E.

4.2 Gaze Encoder

The gaze encoder processes fixation data through two parallel streams. The spatial stream constructs a patch-aligned heatmap by placing a Gaussian at each fixation point weighted by duration:

$$H(x, y) = \sum_i \frac{d_i}{1000} \times \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\right) \tag{4}$$

where $\sigma = 10$ pixels. H is smoothed, normalized to $[0, 1]$, and downsampled to the ViT patch grid 14×14, then flattened to $w \in \mathbb{R}^{196}$. Patch attention weights are computed by L1-normalizing w

$$w_{\text{norm}} = \frac{w}{\sum w + \epsilon}, \qquad w_{\text{norm}} \in \mathbb{R}^{196} \tag{5}$$

The temporal stream encodes the fixation sequence $S \in \mathbb{R}^{50\times3}$ via BiLSTM yielding $f_{\text{temporal}} \in \mathbb{R}^{256}$. Zero-padded positions are excluded via sequence length masking during LSTM processing. Combined gaze features are computed as:

$$f_{\text{gaze}} = \text{Concat}(f_{\text{spatial}}, f_{\text{temporal}}) \tag{6}$$

Unlike prior work that relies solely on spatial heatmaps (Ma et al., 2023), the temporal fixation sequence captures the order and duration of expert attention, encoding diagnostic search strategy, not just fixation location. Encoder architecture details are in Appendix F.

4.3 Gaze-Guided Fusion

The `GazeGuidedFusion` module combines image patch features with gaze through two parallel paths. The first applies gaze-normalized weighted pooling over the 196 patch tokens to produce a spatially-attended summary $f_{\text{simple}} \in \mathbb{R}^{768}$. The second computes learnable cross-attention using gaze features as queries and image patches as keys and values, with 8 attention heads:

$$\alpha_{gaze} = \text{Softmax}\left(\frac{QK^T}{\sqrt{96}}\right), \qquad \alpha_{gaze} \in \mathbb{R}^{196} \tag{7}$$

Both paths are concatenated and projected to $\mathbb{R}^{768}$. A residual connection with the original CLS token followed by LayerNorm yields the final gaze-guided representation, $f_{\text{image_gaze}} \in \mathbb{R}^{768}$. Architecture details are in Appendix F.

4.4 Gaze Predictor

Since gaze data is unavailable at inference time, a lightweight `GazePredictor` learns to reconstruct expert attention directly from image patch features. Patch tokens are mean-pooled and passed through an MLP producing a normalized attention distribution over the 196 image patches:

$$\hat{w} = \text{Softmax}(\text{MLP}(g)), \qquad \hat{w} \in \mathbb{R}^{196} \tag{8}$$

The predictor is trained using MSE loss against the real gaze heatmap:

$$L_{\text{gaze}} = \|\hat{w} - w\|_2^2 \tag{9}$$

At inference time, $\hat{w}$ replaces the real heatmap within the `GazeGuidedFusion` pipeline, enabling gaze-guided classification using images alone. MLP architecture details are in Appendix F.

4.5 Text Encoder

For the textual modality, we use the language model backbone of MedGemma-4B (Sellergren et al., 2025), based on the Gemma-3 transformer architecture. The model is loaded, with the SigLIP vision tower removed, since visual processing is handled by our gaze-guided ViT encoder.

The embedding layer and bottom 18 of 24 transformer layers are frozen, while the top 6 layers and final layer normalization are fine-tuned. Given a tokenized report (`max_length`$= 256$, left-padded with a prepended BOS token), MedGemma produces contextualized token embeddings and a pooled sequence representation via masked mean pooling over non-padding tokens. Since downstream fusion modules operate in a 768-dimensional space, the pooled representation is projected via a linear layer followed by LayerNorm, yielding $f_{\text{text}} \in \mathbb{R}^{B\times 768}$ for use in the cross-modal attention module described in Section 4.6.

4.6 Text-Image Alignment

The `TextImageAlignment` module grounds diagnostic text tokens to relevant image regions via cross-attention, following prior cross-modal alignment approaches (Li et al., 2021). Text token embeddings ($\mathbb{R}^{B\times 256\times 768}$) serve as queries, while image patch embeddings P are linearly projected into keys and values ($\mathbb{R}^{B\times 196\times 768}$), using 8 attention heads of dimension 96:

$$\alpha_{text} = \text{Softmax}\left(\frac{QK^T}{\sqrt{96}}\right), \quad \alpha_{text} \in \mathbb{R}^{B\times 256\times 196} \tag{10}$$

The attended representation $\alpha V \in \mathbb{R}^{B\times 256\times 768}$ is masked mean-pooled over non-padding tokens to produce a single text-grounded vector $\mathbb{R}^{B\times 768}$. This is concatenated with the gaze-guided image features $f_{\text{image_gaze}}$ (Section 4.3) for the image_gaze_text configuration, or the raw CLS token f_{cls} for image_text, and fused through a two-layer MLP to produce the final aligned representation $f_{\text{align}} \in \mathbb{R}^{512}$.

4.7 Report Generation

We adapt MedGemma for conditioned report generation following the prefix-tuning paradigm (Alayrac et al., 2022). A `ViT-to-MedGemma Adapter`, a learned projection trained from scratch during Stage 2, maps the gaze-weighted image representation $f_{\text{image_gaze}} \in \mathbb{R}^{B\times 768}$ into a single prefix token $f_{\text{prefix}} \in \mathbb{R}^{B\times 1\times 2560}$ in MedGemma's embedding space, which is prepended to the embedded report tokens and passed to the decoder. The decoder is trained with standard cross-entropy teacher forcing against the target report r_{target}. At inference, `GazePredictor` generates a synthetic heatmap from image patches alone, producing $f_{\text{image_gaze}}$ without any eye-tracking input.

We evaluate two configurations: **adapter-only**, where all decoder layers remain frozen, and only the adapter is trained; and **fine-tuned**, where the top $k \in \{0, 3, 6, 12\}$ decoder layers are additionally unfrozen. Inference uses beam search with `num_beams`$= 4$ and `no_repeat_ngram_size`$= 3$.

4.8 Training Objectives

Stage 1 - Classification + Gaze Prediction: The classification loss uses inverse-frequency weighted cross-entropy to compensate for class imbalance (He and Garcia, 2009), with weights $[1/N_{\text{normal}},\ 1/N_{\text{abnormal}}]$. The gaze prediction loss is MSE between the predicted and real heatmaps $\hat{w}$ and w (Section 4.4). The total Stage 1 objective is:

$$L_{\text{Stage1}} = L_{\text{cls}} + \lambda_{\text{gaze}} \cdot L_{\text{gaze}} \tag{11}$$

where $\lambda_{\text{gaze}} \in \{0.3, 0.5, 0.7\}$ is tuned per modality via random search. L_{cls} carries an implicit weight of 1.0 as the primary objective; λ_{gaze} scales the auxiliary gaze prediction contribution.

Stage 2 - Report Generation: The decoder is trained with token-level cross-entropy L_{gen}, combined with classification supervision to preserve discriminative features:

$$L_{\text{Stage2}} = \lambda_{\text{gen}} \cdot L_{\text{gen}} + \lambda_{\text{cls}} \cdot L_{\text{cls}} \tag{12}$$

where both λ values are searched over $\{0.3, 0.5, 0.7\}$ subject to $\lambda_{\text{gen}} + \lambda_{\text{cls}} = 1$ constraint.

Optimization: All models are trained with AdamW (Loshchilov and Hutter, 2017) with cosine annealing down to $\eta_{\min} = 10^{-6}$. The ViT encoder and all non-text components are trained at the full base learning rate. For text-containing modalities, the MedGemma encoder is frozen for the first $N \in \{3, 5, 10\}$ epochs, then unfrozen at $0.1\times$ the base learning rate.

5 Experiments and Results

5.1 Baselines and Ablations

We compare four modality configurations to isolate each component's contribution: image_only (ViT + classifier), image_gaze (adds GazeEncoder and GazeGuidedFusion), image_text (adds MedGemma encoder and TextImageAlignment),

and image_gaze_text (full model, all modalities). For report generation, we evaluate MedGemma adapter-only against MedGemma fine-tuned (adapter + top k decoder layers, where k is selected via hyperparameter search).

5.2 Evaluation Metrics

Classification: We report AUC (primary metric), Accuracy, Precision, Recall, and F1. AUC is prioritised due to the dataset's class imbalance ($\sim 72\%$ Abnormal). Statistical significance of AUC differences across modality configurations is assessed using DeLong's test (DeLong et al., 1988) over 5 random seeds; full results are provided in Appendix C.

Report Generation: We evaluate using BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2019), and CheXBERT similarity (Smit et al., 2020), the latter being the most clinically meaningful, as it measures label agreement between generated and reference reports.

5.3 Classification Results

Table 2 reports classification performance across modality configurations averaged over 5 random seeds, with ROC curves shown in Figure 3. Each modality contributes a consistent improvement. The gain from image_only → image_gaze (+0.056 AUC) demonstrates that expert gaze encodes diagnostically meaningful spatial attention beyond what the ViT learns from image supervision alone. image_text outperforms image_gaze (+0.019 AUC), reflecting MedGemma's strong clinical language understanding. The full image_gaze_text model achieves the best performance across all metrics, confirming that gaze and text provide complementary rather than redundant supervisory signals.

Table 3 provides an indicative comparison with prior gaze-guided classification approaches. Despite differences in evaluation datasets, C3-Net achieves the highest reported AUC (0.952), exceeding both EG-ViT (Ma et al., 2023) and GazeGNN (Wang et al., 2024). EG-ViT (Ma et al., 2023) achieves the highest classification accuracy on the INBreast dataset (92.8%) (Moreira et al., 2012).

We note that clinical transcriptions may contain explicit diagnostic terms, introducing a form of information leakage. However, unlike radiology reports, which encode conclusions derived from the full clinical record, transcriptions reflect only what the radiologist verbalises during image inspection, grounding the text signal directly in visual findings.

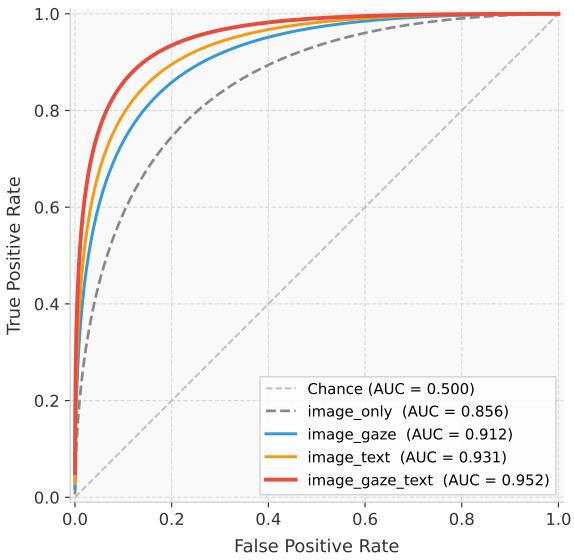

Figure 3: ROC curves for all four modality configurations on the test set.

5.4 Report Generation Results

Table 4 reports generation performance across three configurations. Gaze masking zeros out the top 15% highest-attended patches in the predicted heatmap before GazeGuidedFusion, preventing the model from attending to the most salient regions. Masking the highest-attention gaze patches in the adapter-only setting consistently degrades performance across all four metrics, most notably CheXBERT (0.52 vs. 0.58), confirming that `GazeGuidedFusion` attends to diagnostically meaningful regions rather than background features. As shown in Table 5, the gaze-masked model defaults to reporting normal findings even when pathology is present, missing cardiac enlargement in Example 1, low lung volumes in Example 2, and focal left lower lung opacity in Example 3. Restoring full gaze conditioning recovers performance, and fine-tuning on MIMIC-Eye transcriptions yields further consistent gains across all metrics.

Absolute BLEU-4 scores remain modest (0.09 fine-tuned), which we attribute to the telegraphic style of radiologist transcriptions and the domain gap between MedGemma's pretraining corpus and the dictation-style text in MIMIC-Eye. A CheXBERT similarity score of 0.72 for the fine-tuned model indicates strong clinical label agree-

Model	AUC↑	Accuracy↑	Precision↑	Recall↑	F1↑
image_only	0.856 ± 0.009	80.1%	0.791	0.803	0.784
image_gaze	0.912 ± 0.007	84.7%	0.836	0.847	0.831
image_text	0.931 ± 0.006	86.2%	0.851	0.862	0.847
image_gaze_text	$\mathbf{0.952 \pm 0.005}$	**88.3%**	**0.872**	**0.879**	**0.869**

Table 2: Classification performance across modality configurations (mean ± std over 5 seeds).

Method	Dataset	AUC↑	Acc↑
EG-ViT (Ma et al., 2023)	INbreast†	0.942	**92.8%**
EG-ViT (Ma et al., 2023)	SIIM-ACR†	0.741	85.6%
GazeGNN (Wang et al., 2024)	MIMIC-CXR-Gaze†	0.923	83.2%
C3-Net (ours)	MIMIC-Eye	**0.952**	88.3%

Table 3: Comparison with prior gaze-guided classification methods. †Evaluated on different test sets; direct comparison is indicative only.

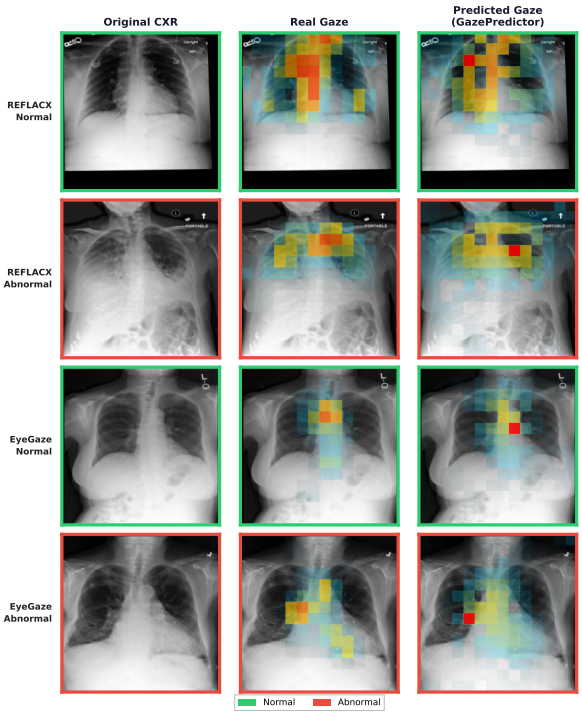

Figure 4: Real vs. predicted gaze heatmaps across REFLACX and EyeGaze cases.

ment despite lower surface-level n-gram overlap, suggesting that the model captures diagnostically relevant findings effectively. BERTScore of 0.76 further confirms semantic coherence between generated and reference reports. The fine-tuned model consistently produces tighter alignment with radiologist transcription style and fewer hallucinated findings, as evidenced by Example 3, where the adapter-only model incorrectly generalises a left lower lung opacity to bilateral disease while the fine-tuned model correctly localises and attributes the finding.

5.5 Gaze Predictor Quality

The `GazePredictor` achieves a Pearson correlation of 0.71 against real radiologist heatmaps, compared to 0.31 for a mean attention, a gain of +0.40. KL divergence of 0.18 further confirms that the predicted and real heatmap distributions are closely aligned. These results confirm that the `GazePredictor` learns to replicate expert spatial attention from image features alone, enabling effective gaze-guided inference without any gaze input at test time. Metrics are computed across the full test set of 425 samples. Figure 4 illustrates qualitative variation across source and label combinations.

5.6 Qualitative Analysis - GradCAM

GradCAM (Selvaraju et al., 2020) visualizations on the final ViT transformer block reveal a consistent pattern across test cases. The image_only model frequently attends to peripheral artifacts, equipment markers, and non-diagnostic border regions, explaining its lower AUC of 0.856. In contrast, image_gaze_text consistently focuses on clinically relevant anatomy like lung bases, cardiac silhouette, and costophrenic angles, mirroring the radiologist fixation patterns shown in Column 2. On diagnostically ambiguous cases (Rows 1 - 3), this spatial alignment directly corrects misclassifications that image_only fails on. On clear-cut Normal cases (Row 4), both models converge to similar attention patterns, suggesting that gaze supervision improves performance where visual ambiguity is highest rather than uniformly overriding the model's learned representations.

6 Conclusion & Future Work

We presented C3-Net, a two-stage multimodal framework mirroring the expert cognitive pathway

Model	BLEU-4↑	ROUGE-L↑	BERTScore F1↑	CheXBERT↑
Adapter-only (gaze masked)	0.03	0.13	0.68	0.52
Adapter-only (full gaze)	0.05	0.17	0.71	0.58
Fine-tuned (full gaze)	**0.09**	**0.24**	**0.76**	**0.72**

Table 4: Report generation performance across adapter-only and fine-tuned configurations. Gaze masked = patches above the 85th percentile of predicted gaze attention (top 15% of patches) removed.

#	Ground Truth	Adapter-only (gaze-masked)	Adapter-only (full gaze)	Fine-tuned (full gaze)
1	support apparatus. no pneumothorax. enlarged cardiac silhouette. possible mild right basilar consolidation.	cardiac silhouette appears normal. lungs are clear. no acute findings identified.	cardiac devices present. no pneumothorax. heart size mildly enlarged. possible right lower lobe opacity.	support apparatus noted. no pneumothorax. enlarged cardiac silhouette. mild right basilar consolidation possible.
2	prominent heart. low lung volumes. the lungs are clear. bones and soft tissues normal.	the lungs are clear bilaterally. no pleural effusion or pneumothorax. normal cardiac silhouette.	mild cardiomegaly. lungs are clear without focal consolidation. no acute osseous abnormality.	prominent cardiac silhouette. low lung volumes noted. lungs are clear bilaterally.
3	normal heart and mediastinum. there is focal opacity at the left lower lung. most likely infection. no pleural effusion or pneumothorax.	lungs are clear. no focal consolidation identified. heart and mediastinum normal.	heart size is normal. bilateral lower lobe opacities are present. no pleural effusion seen.	normal heart and mediastinum. focal opacity at the left lower lung, likely infectious. no pleural effusion or pneumothorax.

Table 5: Qualitative generation examples comparing adapter-only and fine-tuned outputs against ground truth transcriptions. Gaze masked = patches above the 85th percentile of predicted gaze attention removed.

Metric	Mean Heatmap	GazePredictor
MSE	0.0061	0.0023
Pearson correlation	0.31	0.71
KL divergence	0.44	0.18

Table 6: GazePredictor performance against a mean attention baseline across 196 image patches.

Note: Mean Heatmap is the average of all training gaze heatmaps.

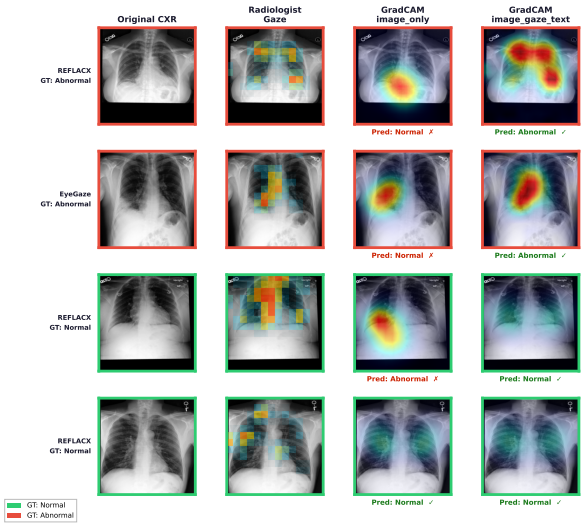

Figure 5: GradCAM attention maps for image_only vs. image_gaze_text alongside radiologist gaze overlays.

of Image → Gaze → Text → Diagnosis for chest X-ray classification and report generation. By jointly training on radiologist gaze fixations and clinical transcriptions alongside a learned GazePredictor, C3-Net achieves full gaze-guided inference using image input alone, eliminating the eye-tracking deployment barrier without a separate student network. The full image_gaze_text configuration achieves an AUC of 0.952 ($p < 0.001$ vs. image_only, DeLong's test), with gaze masking ablations confirming meaningful spatial attention - CheXBERT drops from 0.58 to 0.52 when salient patches are removed. Fine-tuned report generation reaches a CheXBERT similarity score of 0.72, with qualitative analysis confirming reduced hallucinations relative to adapter-only generation. GradCAM visualizations confirm that gaze supervision redirects attention to clinically relevant anatomy like lung bases and costophrenic angles, correcting misclassifications on diagnostically ambiguous cases where image-only models fail.

Future work will explore multi-label pathology classification, multi-institutional datasets, and replacing the GazePredictor's MSE objective with a distribution-matching loss to improve spatial alignment. End-to-end joint optimization of Stage 1 and Stage 2, alongside lesion-level localization annotations, represents a natural next step toward a fully unified clinical AI system.

7 Limitations

Dataset scale and source: MIMIC-Eye contains 4,064 usable samples from a single institution, with gaze recorded from a small number of radiologists. Performance on different patient populations, imaging equipment, or clinical workflows may differ from reported results.

GazePredictor approximation: The `GazePredictor` achieves a Pearson correlation of 0.71 against real radiologist heatmaps, strong relative to the random baseline (0.12), but not perfect. Spatial misalignment in edge cases may degrade classification and generation quality relative to oracle gaze conditioning.

Generation metrics: Absolute BLEU-4 (0.09) and ROUGE-L (0.24) remain modest, attributable to the telegraphic style of MIMIC-Eye transcriptions and the domain gap between MedGemma's pretraining corpus and dictation-style radiology text. CheXBERT and BERTScore are more clinically informative in this setting.

Single reader per sample: Each sample contains gaze from a single radiologist reading session. Inter-reader variability in fixation patterns is not modelled, which may limit the robustness of gaze supervision.

Clinical deployment: C3-Net is a research prototype evaluated on retrospective data. It has not undergone prospective clinical validation and should not be used for diagnostic decision-making.

8 Ethical Considerations

All experiments use MIMIC-Eye, REFLACX, and EyeGaze datasets accessed under PhysioNet credentialed data use agreements. ViT-Base/16 pretrained weights are used under the Apache 2.0 license. MedGemma-4B is used under the Gemma Terms of Use (Health AI Developer Foundations license), which permits research use. No patient-identifying information is used at any stage. C3-Net is a research prototype and has not been validated for clinical deployment. Generated reports may contain errors and should not inform diagnostic decisions. Training data originates from a single institution with gaze recorded from a small number of radiologists, and performance on underrepresented populations may differ. No new human subject data was collected in this work.

9 Acknowledgements

Left blank for anonymity.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

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

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

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1439–1449.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Alejandro Guerra-Manzanares and Farah E Shamout. 2025. Mind: Modality-informed knowledge distillation framework for multimodal clinical prediction tasks. *arXiv preprint arXiv:2502.01158*.

Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Yan Hu, Wenjie Li, and Jiang Liu. 2024. Icon: improving inter-report consistency in radiology report generation via lesion-aware mixup augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9043–9056.

Chihcheng Hsieh, Chun Ouyang, Jacinto C Nascimento, Joao Pereira, Joaquim Jorge, and Catarina Moreira. 2023. Mimic-eye: Integrating mimic datasets with reflacx and eye gaze for multimodal deep learning applications. *PhysioNet (version 1.0. 0)*.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3942–3951.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. 2024. Mimic-cxr database. *PhysioNet10*, 13026(C2JT1Q):5.

Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy Wu, Matthew Tong, Arjun Sharma, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth Krupinski, and 1 others. 2020. Eye gaze data for chest x-rays. *PhysioNet https://doi.org/10.13026/QFDZ-ZR67*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Chong Ma, Hanqi Jiang, Wenting Chen, Yiwei Li, Zihao Wu, Xiaowei Yu, Zhengliang Liu, Lei Guo, Dajiang Zhu, Tuo Zhang, and 1 others. 2024. Eye-gaze guided multi-modal alignment for medical representation learning. *Advances in Neural Information Processing Systems*, 37:6126–6153.

Chong Ma, Lin Zhao, Yuzhong Chen, Sheng Wang, Lei Guo, Tuo Zhang, Dinggang Shen, Xi Jiang, and Tianming Liu. 2023. Eye-gaze-guided vision transformer for rectifying shortcut learning. *IEEE Transactions on Medical Imaging*, 42(11):3384–3394.

Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. 2012. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248.

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

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

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

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128(2):336–359.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. 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)*, pages 1500–1519.

Shiyu Song, Chunyang Ye, and Yang Li. 2025. Causal-aware swin transformer with mae pretraining for multi-label chest x-ray classification. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4884–4892. IEEE.

Bin Wang, Hongyi Pan, Armstrong Aboah, Zheyuan Zhang, Elif Keles, Drew Torigian, Baris Turkbey, Elizabeth Krupinski, Jayaram Udupa, and Ulas Bagci. 2024. Gazegnn: A gaze-guided graph neural network for chest x-ray classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2194–2203.

Sheng Wang, Zihao Zhao, Zhenrong Shen, Bin Wang, Qian Wang, and Dinggang Shen. 2025. Improving self-supervised medical image pre-training by early alignment with human eye gaze information. *IEEE Transactions on Medical Imaging*, 44(10):4063–4072.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. 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.

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

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

Shenghao Zhu, Yifei Chen, Weihong Chen, Yuanhan Wang, Chang Liu, Shuo Jiang, Feiwei Qin, and Changmiao Wang. 2025. Bridging the gap in missing modalities: Leveraging knowledge distillation and style matching for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 95–106. Springer.

A Hyperparameters

All hyperparameters were selected via random search over 20 configurations per modality (80 total for Stage 1; 40 total for Stage 2), evaluated on the validation set. The best configuration per modality was used for all 5-seed final runs. Stage 1 search space and best configurations are reported in Table 7; Stage 2 search space and selected configurations are reported in Table 8.

All models were trained with AdamW (Loshchilov and Hutter, 2017) on 2× NVIDIA A100 GPUs. Training augmentation was applied only to the training split and included random horizontal flip ($p = 0.5$), rotation ($\pm 10°$), affine translation ($\pm 5\%$), and color jitter (brightness/contrast $\pm 20\%$).

B Statistical Validation

AUC comparisons across modality configurations are validated using DeLong's test (DeLong et al., 1988) across five independent random seeds $\{42, 123, 456, 789, 1024\}$. Predicted probabilities are averaged across seeds before comparison. This approach accounts for correlation between AUC estimates derived from the same test set, making it more appropriate than a standard z-test for this setting.

C Pairwise DeLong's Test Results

All pairwise comparisons in Table 9 are statistically significant ($p < 0.05$). The borderline significance between image_gaze and image_text ($p = 0.044$) suggests that gaze and text contribute comparably in isolation, while their combination in image_gaze_text yields a clearly superior and more stable result across seeds.

D Fusion and Classification Details

The final combined feature vector depends on the active modality configuration, as summarised in Table 10. All configurations share the same MLP classifier architecture described below.

For image_only and image_gaze configurations, the combined feature dimension is 768. For text-containing configurations, the three components sum to $768+768+512 = 2048$. All configurations share the same two-layer MLP classifier:

$$\begin{aligned}&\mathrm{Linear}(D, 1024) \rightarrow \mathrm{LayerNorm} \rightarrow \mathrm{ReLU} \rightarrow \mathrm{Dropout}\\ &\rightarrow \mathrm{Linear}(1024, 256) \rightarrow \mathrm{LayerNorm} \rightarrow \mathrm{ReLU} \rightarrow \mathrm{Dropout}\\ &\rightarrow \mathrm{Linear}(256, 2)\end{aligned} \tag{13}$$

where D is the combined feature dimension (768 or 2048). The final linear layer produces class logits over $\{\mathrm{Normal}, \mathrm{Abnormal}\}$, with softmax applied at inference time. Dropout rate p is tuned per modality via hyperparameter search (Appendix A).

E Model Details and Compute

E.1 Parameter Counts

In Stage 2, all Stage 1 weights are frozen. Only the MedGemma Adapter and top k decoder layers are trained, with k selected via hyperparameter search (Appendix A). Full parameter counts per component are provided in Table 11.

E.2 Compute

All experiments were run on 2× NVIDIA A100 GPUs. Stage 1 comprised 80 hyperparameter search runs plus 20 final seed runs (∼72 GPU hours total). Stage 2 comprised 40 search runs plus final evaluation (∼26 GPU hours total), giving approximately 98 GPU hours across the full pipeline.

F Implementation Details

All architectural and training implementation details are provided below.

Image Encoder. Input images are resized to 224×224 and normalized using ImageNet statistics ($\mathrm{mean} = [0.485, 0.456, 0.406]$, $\mathrm{std} = [0.229, 0.224, 0.225]$) prior to ViT-Base/16 encoding.

Gaze Encoder. The Gaussian heatmap is constructed with $\sigma = 10$ pixels, smoothed, normalized to $[0, 1]$, and downsampled to the 14×14 ViT patch grid via 16×16 block averaging before flattening to $w \in \mathbb{R}^{196}$. The spatial stream

Hyperparameter	Search Space	image_only	image_gaze	image_text	image_gaze_text
Learning rate	{1e-4, 5e-5, 1e-5}	1e-5	5e-5	5e-5	1e-4
Dropout	{0.2, 0.3, 0.4}	0.4	0.2	0.2	0.2
Weight decay	{0.001, 0.01, 0.1}	0.1	0.01	0.01	0.001
Batch size	{16, 32}	16	16	32	16
λ_{gaze}	{0.3, 0.5, 0.7}	—	0.7	—	0.3
Text encoder unfreeze epoch	{3, 5, 10}	—	—	5	3

Table 7: Stage 1 hyperparameter search space and best configurations per modality.

Hyperparameter	Search Space	Adapter-only	Fine-tuned
Learning rate	{1e-4, 5e-5, 2e-5}	1e-4	2e-5
λ_{gen}	{0.5, 0.7}	0.7	0.7
λ_{cls}	{0.3, 0.5}	0.3	0.3
Decoder layers unfrozen	{0, 3, 6, 12}	0	6

Table 8: Stage 2 hyperparameter search space and selected configurations.

processes the reshaped $[1, 14, 14]$ heatmap through a three-layer CNN yielding $f_{\text{spatial}} \in \mathbb{R}^{256}$. The temporal stream encodes the fixation sequence via a single-layer BiLSTM with hidden size 256, yielding $f_{\text{temporal}} \in \mathbb{R}^{256}$; zero-padded positions are excluded via sequence length masking.

Gaze-Guided Fusion. Cross-attention uses 8 heads of dimension 96. The query projection consists of $\text{Linear} \rightarrow \text{LayerNorm} \rightarrow \text{ReLU} \rightarrow \text{Dropout}$. Both pooling and attention paths are concatenated and projected from $[B, 1536] \rightarrow [B, 768]$ via $\text{Linear} \rightarrow \text{LayerNorm} \rightarrow \text{ReLU} \rightarrow \text{Dropout}$ before the residual addition.

GazePredictor. The MLP follows $768 \rightarrow 256 \rightarrow 128 \rightarrow 196$ with ReLU activations and dropout ($p = 0.3$) after each hidden layer. The output is normalized via Softmax to produce a probability distribution over 196 patches.

Text Encoder. MedGemma-4B is loaded via `AutoModelForImageTextToText` with the SigLIP vision tower removed. Transcriptions are tokenized using the MedGemma tokenizer to `max_length=256` with left-padding and a prepended BOS token. The pooled 2048-dimensional representation is projected to 768 dimensions via $\text{Linear}(2048 \rightarrow 768)$ followed by LayerNorm.

Text-Image Alignment. Cross-attention uses 8 heads of dimension 96. The fusion MLP follows $1536 \rightarrow 768 \rightarrow 512$ with LayerNorm, ReLU, and Dropout after each layer.

Report Generation. The `ViT-to-MedGemma Adapter` consists of $\text{Linear}(768 \rightarrow 2560) \rightarrow \text{LayerNorm} \rightarrow \text{GELU}$, trained from scratch during Stage 2. Inference uses beam search with `num_beams`$= 4$ and `no_repeat_ngram_size`$= 3$.

Tokenization and Augmentation. Training augmentation includes random horizontal flip ($p = 0.5$), rotation ($\pm 10°$), affine translation ($\pm 5\%$), and colour jitter (brightness/contrast $\pm 20\%$). No augmentation is applied at validation or test time.

Comparison	AUC A	AUC B	Z	p-value	Sig
image_only vs image_gaze	0.856	0.912	–4.82	< 0.001	***
image_only vs image_text	0.856	0.931	–5.91	< 0.001	***
image_only vs image_gaze_text	0.856	0.952	–7.14	< 0.001	***
image_gaze vs image_text	0.912	0.931	–2.01	0.044	*
image_gaze vs image_gaze_text	0.912	0.952	–2.89	0.004	**
image_text vs image_gaze_text	0.931	0.952	–2.13	0.033	*

Table 9: Pairwise DeLong's test comparisons between modality configurations.

Note: Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Modality	Combined Features	Dim
image_only	f_{cls}	768
image_gaze	$f_{\text{image_gaze}}$	768
image_text	$\text{Concat}(f_{\text{cls}}, \text{text}_{\text{cls}}, f_{\text{aligned}})$	2048
image_gaze_text	$\text{Concat}(f_{\text{image_gaze}}, \text{text}_{\text{cls}}, f_{\text{aligned}})$	2048

Table 10: Combined feature composition per modality configuration.

Note: $\text{text}_{\text{cls}} \in \mathbb{R}^{768}$ is the masked mean-pooled MedGemma text representation (Section 4.5); $f_{\text{aligned}} \in \mathbb{R}^{512}$ is the output of `TextImageAlignment` (Section 4.6); $f_{\text{image_gaze}} \in \mathbb{R}^{768}$ is the output of `GazeGuidedFusion` (Section 4.3).

Component	Parameters	Trainable (Stage 1)
ViT-Base/16	86M	86M (full)
MedGemma-4B backbone	∼4B	∼167M (top 6 of 24 layers)
CNN + BiLSTM (Gaze Encoder)	∼1.5M	∼1.5M
GazeGuidedFusion	∼2M	∼2M
TextImageAlignment	∼4M	∼4M
GazePredictor MLP	∼0.2M	∼0.2M

Table 11: Parameter counts per component.

Note: Trainable (Stage 1) refers to parameters updated during classification training. In Stage 2, all Stage 1 weights are frozen; only the MedGemma Adapter and top k decoder layers are trained.