# Less Is More? Selective Visual Attention to High-Importance Regions for Multimodal Radiology Summarization

Anonymous ACL submission


## Abstract

Automated radiology report summarization aims to distill verbose findings into concise clinical impressions, but existing multimodal models often struggle with visual noise and fail to meaningfully improve over strong text-only baselines in the FINDINGS → IMPRESSION transformation. We challenge two prevailing assumptions: (1) that more visual input is always better, and (2) that multimodal models add limited value when findings already contain rich image-derived detail. Through controlled ablations on MIMIC-CXR benchmark, we show that selectively focusing on pathology-relevant visual patches rather than full images yields substantially better performance. We introduce **ViTAS**, **V**isual-**T**ext **A**ttention **S**ummarizer, a multi-stage pipeline that combines ensemble-guided MedSAM2 lung segmentation, bidirectional cross-attention for multi-view fusion, Shapley-guided adaptive patch clustering, and hierarchical visual tokenization feeding a ViT. ViTAS achieves SOTA results with 29.25% BLEU-4 and 69.83% ROUGE-L, improved factual alignment in qualitative analysis, and the highest expert-rated human evaluation scores. Our findings demonstrate that *less but more relevant* visual input is not only sufficient but superior for multimodal radiology summarization. Code is publicly available at TBA.


## 1 Introduction

Radiology report summarization is essential for distilling verbose diagnostic findings into concise, actionable impressions, thereby supporting clinical decision-making and reducing radiologist workload (Miura et al., 2021). With the rapid growth of medical imaging volumes, automated multimodal summarization-integrating both textual findings and radiological images-has emerged as a promising direction to produce more clinically faithful and contextually grounded impressions (Kim et al., 2023a; Wang et al., 2023a; Nicolson et al., 2023a; Delbrouck et al., 2021; Zhang et al., 2024).

Two influential assumptions have shaped recent research in this area. First, many works implicitly follow a "the more, the better" heuristic: the more visual input provided to the model (full images, multiple views, higher resolution), the better the summarization performance (Wang et al., 2023a). This assumption has rarely been systematically challenged in the radiology domain, even though irrelevant regions frequently introduce visual noise that can dilute cross-modal alignment and degrade factual accuracy (DeGrave et al., 2021; Rajaraman et al., 2024; Niehoff et al., 2023). Second, several studies argue that multimodal models add limited value when transforming richly detailed FINDINGS sections (already written by radiologists with comprehensive image interpretation) into concise IMPRESSION sections. According to this view, if the FINDINGS section contains sufficient detail, it should already encode all key radiographic information, and the primary limitation of multimodal models lies in their failure to learn meaningful cross-modal interactions between text and visual representations (Sim et al., 2025).

We present the first study on MIMIC-CXR that directly tests *both* assumptions in the same controlled setting. Our results show that: contrary to the first assumption, feeding full uncurated images (even multi-view) is *not* optimal-selective incorporation of only the most pathology-relevant visual patches yields substantially larger gains than simply increasing the amount of visual input; contrary to the second assumption, properly integrated multi-view visual context *does* meaningfully improve FINDINGS → IMPRESSION summarization quality across all major automatic and clinical metrics. To achieve this, we propose ViTAS, a multi-stage framework that systematically removes noise and focuses the model on clinically salient information. The pipeline begins with robust ensemble-guided lung segmentation using MedSAM2 (Ma et al., 2025), followed by dual-backbone Swin

Transformer V2 processing with bidirectional spatial cross-attention for multi-view fusion. Post-hoc gradient-weighted attention maps combined with Shapley value-based view contribution quantification enable dynamic, pathology-adaptive selection of high-importance patch clusters. These selected clusters are then hierarchically tokenized (global + cluster + dynamic extra patches) and fused with textual findings in a ViT + T5 multimodal transformer to generate structured impressions. Our main contributions are:

- We introduce an end-to-end pipeline combining ensemble MedSAM2 lung segmentation, bidirectional mid-level cross-attention, view-adaptive patch clustering, and hierarchical visual tokenization. The approach achieves SOTA automatic metrics, substantially improves factual alignment and reduces hallucination risk by qualitative analysis, and obtains the highest expert-rated scores in human evaluation across readability, factual correctness, informativeness, and completeness.

- We show that selective, attention- and Shapley-guided patch extraction outperforms both full-image and naive ROI-based multimodal baselines, providing the first clear evidence in radiology summarization that *less (but more relevant) visual input is better*.

- We demonstrate, through controlled ablations on MIMIC-CXR, that multimodal models *can* significantly outperform strong text-only baselines in the FINDINGS $\rightarrow$ IMPRESSION task when visual noise is effectively removed-challenging the view that multimodal gains are inherently limited in previous setting.

## 2 Related Work

With the rapid advancement of multimedia content, multimodal summarization has gained increasing attention to generate concise and informative summaries (Li et al., 2018; Jangra et al., 2021; Overbay et al., 2023; Liang et al., 2023c; Sanabria et al., 2018; Mahasseni et al., 2017; Liu et al., 2024; Mukherjee et al., 2022; Yu et al., 2021). Recent studies have begun to explore incorporating radiology images as inputs to multimodal summarization models, under the assumption that their rich visual information can enhance the quality of generated summaries (Ghosh et al., 2024; Thawakar et al., 2024; Hu et al., 2023; Van Veen et al., 2023; Chen et al., 2023; Delbrouck et al., 2023). Some works focus on pretraining vision-language models such as CLIP (Radford et al., 2021) and ViT (Dosovitskiy et al., 2020), while others address radiology report summarization from findings sections (Zhang et al., 2018) or explore multimodal summarization techniques including guided attention across modalities (Atri et al., 2021), self-supervised opinion summarization with images (Im et al., 2021), and multimodal approaches specifically for radiology reports (Delbrouck et al., 2021).

Several studies have explored modality contributions in multimodal models (Goyal et al., 2017). Parcalabescu and Frank(2023) employ Shapley values by randomly masking image patches and text tokens to quantify changes in output. Liang et al.(2023a) introduced an information-theoretic framework to measure multimodal interactions in datasets and performed human annotation studies on general-domain data (Liang et al., 2023b).

CSTRL (Naznin et al., 2025) shows that learning from salient textual statements improves summarization, highlighting that not all contextual information is equally informative. While textual saliency has been explored, selective visual importance in multimodal radiology summarization remains under-investigated. Although NEURAL (Joshi and Rekik, 2025) and MSCL (Zhao et al., 2023) demonstrate the effectiveness of salient visual regions for report generation, most multimodal summarization approaches still assume a "the more, the better" paradigm or suggest that textual information alone can often be sufficient.

## 3 Method

We propose a multi-stage framework for radiology report summarization that prioritizes clinically relevant information while discarding irrelevant regions. First, lung regions are accurately segmented using an ensemble-guided MedSAM2 approach to focus the model on pulmonary anatomy. Frontal and lateral views are then processed through a dual-backbone Swin Transformer V2 architecture with bidirectional spatial cross-attention, enabling complementary information exchange while preserving view-specific features. Attention-driven post-hoc analysis identifies high-importance patches and quantifies each view's contribution using Shapley values, allowing dynamic selection of the most informative regions. Finally, the selected patch clusters are fused with textual prompts in a multi-

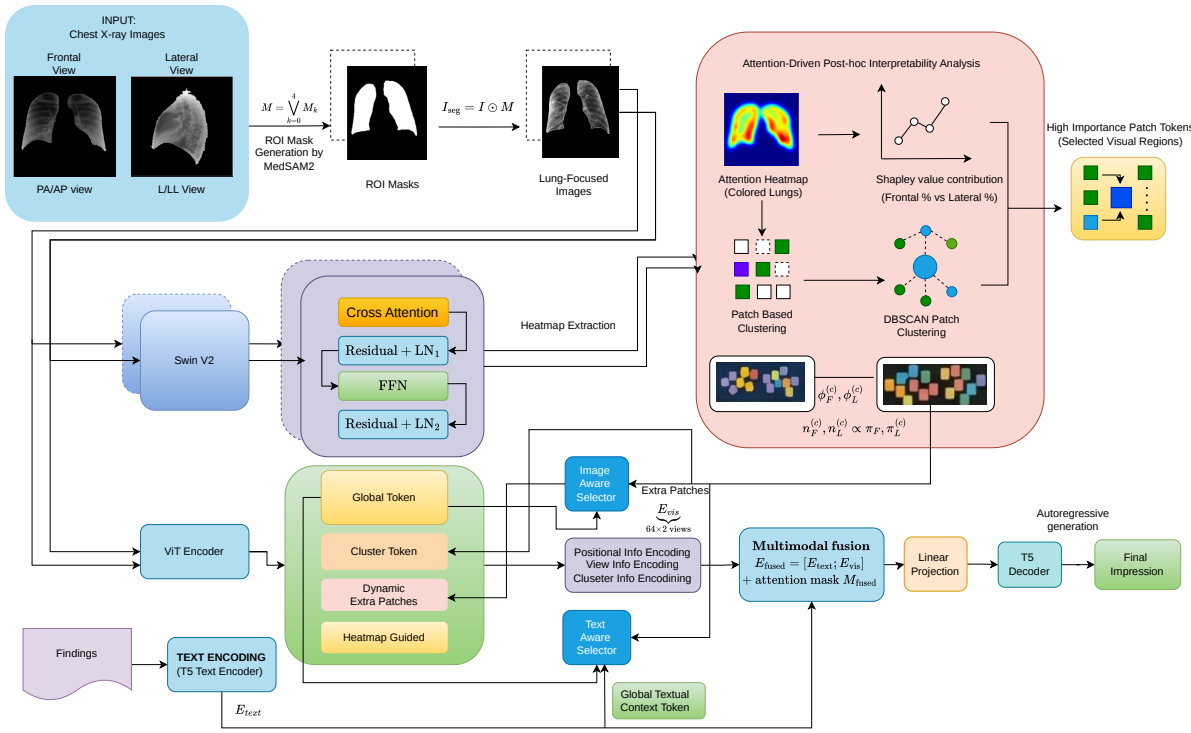


Figure 1: Overview of the **ViTAS** pipeline: Chest X-rays (frontal and lateral) are lung-segmented using MedSAM2, then fused via a dual SwinV2 with cross-attention. Interpretability modules identify important regions, which are selectively tokenized and combined with text embeddings in a T5 decoder to generate the final clinical impression.

modal transformer framework (ViT + T5) to generate structured clinical impressions. An overview is provided in Figure 1.

### 3.1 ROI-Guided Visual Context Extraction

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote a chest X-ray image with height $H$ and width $W$ and $\mathcal{V} \in \{\text{AP}, \text{PA}, \text{L}, \text{LL}\}$ indicate its projection view (Anteroposterior, Posteroanterior, Lateral, Left Lateral, respectively). We process each view independently to avoid view-specific confounding effects as shown in Figure 2.

**Ensemble Bounding-Box Prompting.** We utilize MedSAM2 model (Ma et al., 2025) with predictor $f_\theta$ to obtain binary lung segmentation masks. A reference bounding box $B_0 = [x_1, y_1, x_2, y_2]$ is defined in pixel coordinates using empirical statistics computed from a small set of manually annotated images. Its corners are expressed in normalized image coordinates as: $\left(\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}\right)$. Here $(x_1, y_1)$ and $(x_2, y_2)$ represent the top-left and bottom-right corners, respectively.

To increase robustness against anatomical variability, patient positioning differences, and model sensitivity to exact box placement, we create an ensemble $\mathcal{B} = \{B_k\}_{k=0}^{4}$ of $K = 5$ bounding boxes by applying small directional shifts of magnitude $s$ pixels: $\delta_k \in \{[0,0,0,0], [-s,0,-s,0], [+s,0,+s,0], [0,-s,0,-s], [0,+s+\Delta,+s+\Delta,0]\}$, where $\Delta$ is a small additional downward bias sometimes applied to better cover the lower lung fields (cardiophrenic angles). Each $\delta_k = [\Delta x_1, \Delta y_1, \Delta x_2, \Delta y_2]$ is added element-wise to $B_0$, after which the resulting box coordinates are clipped to lie within $[0, W] \times [0, H]$. For each box $B_k \in \mathcal{B}$, MedSAM2 produces a binary mask:

$$M_k = f_\theta(I,\ B_k)\ \in\ \{0,1\}^{H\times W}, \qquad (1)$$

where value 1 indicates a pixel predicted to belong to lung tissue. The final lung region-of-interest (ROI) mask is obtained by taking the pixel-wise logical union (inclusive OR) over the ensemble:

$$M = \bigvee_{k=0}^{4} M_k. \qquad (2)$$

The union improves robustness to box misalignment and under-segmentation, yielding a more inclusive yet well-localized lung mask.

The segmented lung image is then computed via element-wise multiplication (Hadamard product):

$$I_{\text{seg}} = I \odot M, \qquad (3)$$

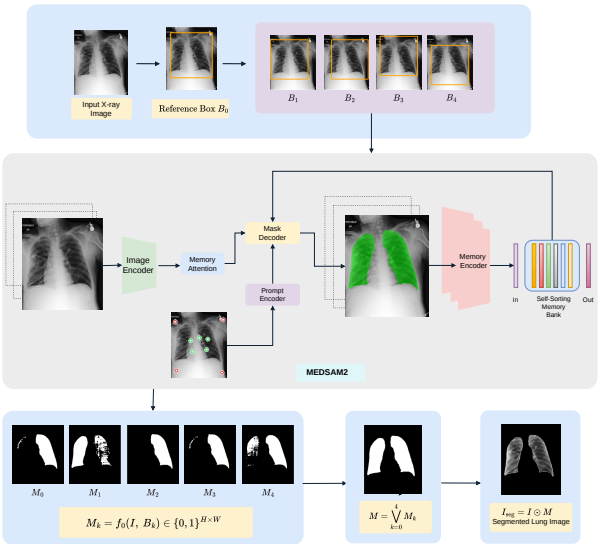


Figure 2: Ensemble-guided MedSAM2 lung segmentation pipeline. A reference box generates five shifted bounding boxes. MedSAM2 produces binary masks which are unioned into the final lung ROI mask. The segmented image retains only pulmonary fields.

which suppresses all non-lung regions (setting them to zero) and retains high-fidelity visual context from the pulmonary fields only.

### 3.2 Multi-View Feature Fusion with Bidirectional Spatial Cross-Attention

As shown in Figure 3, segmented chest X-ray views are grouped into two complementary categories: frontal (AP, PA) and lateral (L, LL). Lung-focused segmented images $I^{\mathrm{F}}_{\mathrm{seg}}, I^{\mathrm{L}}_{\mathrm{seg}} \in \mathbb{R}^{H\times W\times 3}$ are processed by a dual-backbone architecture based on Swin Transformer V2 (Liu et al., 2021). Early-stage encoders $\mathcal{E}_{\mathrm{F}}(\cdot), \mathcal{E}_{\mathrm{L}}(\cdot) : \mathbb{R}^{3\times 256\times 256} \to \mathbb{R}^{N\times D}$ extract spatial token sequences from frontal and lateral views, where $N = (256/p)^2$ is the number of patches and $D$ is the token dimension. To enable interaction between the two views while preserving view-specific representations, we perform mid-level fusion using bidirectional spatial cross-attention (Zhou et al., 2023). For fusion layer $\ell$, let $\mathbf{F}^{(\ell-1)}, \mathbf{L}^{(\ell-1)} \in \mathbb{R}^{B\times N\times D}$ denote frontal and lateral tokens, where $B$ is the batch size. Frontal-to-lateral attention is computed as

$$\mathbf{A}_{\mathrm{F}\to\mathrm{L}} = \mathrm{softmax}\Big(\frac{\mathbf{Q}_{\mathrm{F}}\mathbf{K}_{\mathrm{L}}^{\top}}{\sqrt{d_h}} \odot \mathbf{M}_{\mathrm{L}}\Big), \qquad (4)$$

where $h$ is the number of attention heads, $d_h = D/h$ is the head dimension, and $\mathbf{M}_{\mathrm{L}} \in \{0,1\}^{B\times 1\times 1\times N}$ indicates lateral-view availability. The frontal representation is updated as

$$\hat{\mathbf{F}}^{(\ell)} = \mathbf{F}^{(\ell-1)} + \mathrm{Proj}_{\mathrm{F}}(\mathbf{A}_{\mathrm{F}\to\mathrm{L}}\mathbf{V}_{\mathrm{L}}) \odot \mathbf{M}_{\mathrm{L}} \qquad (5)$$

and the lateral update is computed symmetrically. Each branch is refined using residual connections, layer normalization (LN), and a feed-forward network (FFN). To handle missing views, learnable missing tokens $\mathbf{m}_{\mathrm{F}}, \mathbf{m}_{\mathrm{L}} \in \mathbb{R}^{1\times 1\times D}$ replace unavailable inputs using view masks.

After $L$ fusion layers, the token sequences are passed through the remaining Swin stages to produce global descriptors $\mathbf{z}_{\mathrm{F}}, \mathbf{z}_{\mathrm{L}} \in \mathbb{R}^{B\times D_f}$, where $D_f$ is the final feature dimension.

A lightweight bidirectional cross-attention module produces the fused representation

$$\mathbf{z}_{fused} = \mathrm{Concat}\Big(\mathrm{MHA}_{\mathrm{L}\to\mathrm{F}}(\mathbf{z}_{\mathrm{F}}, \mathbf{z}_{\mathrm{L}}), \mathrm{MHA}_{\mathrm{F}\to\mathrm{L}}(\mathbf{z}_{\mathrm{L}}, \mathbf{z}_{\mathrm{F}})\Big). \qquad (6)$$

A small MLP classifier outputs pathology probabilities $\hat{\mathbf{y}} \in \mathbb{R}^{B\times 8}$ for Atelectasis, Cardiomegaly, Consolidation, Edema, No Finding, Pleural Effusion, Pneumothorax, and Support Devices. The model is trained using binary cross-entropy loss.

### 3.3 Attention-Driven Post-hoc Interpretability with Adaptive Patch Clustering

To interpret the classification output $\hat{\mathbf{y}}$, we extract spatial attention maps from the final fusion layer. Let $\mathbf{A}^{(L)}_{\mathrm{F}\to\mathrm{L}}, \mathbf{A}^{(L)}_{\mathrm{L}\to\mathrm{F}} \in \mathbb{R}^{B\times h\times N\times N}$ denote cross-view attention weights. Patch importance for the frontal view is computed by aggregating attention received from lateral tokens:

$$\mathcal{H}^{\mathrm{F}} = \mathrm{softmax}\Big(\sum_{h=1}^{h} \mathbf{A}^{(L)}_{\mathrm{L}\to\mathrm{F}}[b, h, :, :]\Big) \qquad (7)$$

which is reshaped to the image resolution to form the heatmap $\mathcal{H}^{\mathrm{F}} \in [0,1]^{H\times W}$. The lateral heatmap $\mathcal{H}^{\mathrm{L}}$ is computed analogously.

To quantify view contributions, we compute 2-player Shapley values (Lundberg and Lee, 2017). For pathology class $c$,

$$\phi^{(c)}_{\mathrm{F}} = \frac{1}{2}\Big(v(\{\mathrm{F}\}) - v(\emptyset) + v(\{\mathrm{F},\mathrm{L}\}) - v(\{\mathrm{L}\})\Big) \qquad (8)$$

and similarly for $\phi^{(c)}_{\mathrm{L}}$. These are converted to percentage contributions $\pi^{(c)}_{\mathrm{F}}, \pi^{(c)}_{\mathrm{L}}$.

High-attention regions are then grouped into clusters. The heatmap is divided into a $p \times p$ grid, producing patch scores $s^{v}_{i,j}$. The top $k\%$ patches are selected and clustered using DBSCAN (Ester

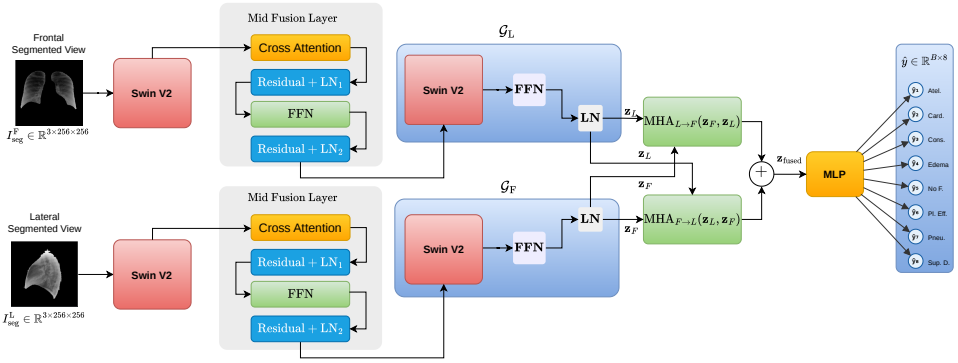


Figure 3: Dual Swin Transformer V2 with bidirectional mid-fusion cross-attention. Frontal and lateral views are processed by separate backbones, exchange information via cross-attention, and their global features are fused before classification.

et al., 1996): $\mathcal{C}^v = \text{DBSCAN}(\mathcal{Q}^v, \varepsilon, m)$, where $\varepsilon$ is the neighborhood radius and $m$ is the minimum cluster size. Finally, the number of patches extracted from each view is dynamically allocated according to the Shapley percentages:

$$n_{\text{F}}^{(c)} = \left\lfloor N_{total} \frac{\pi_{\text{F}}^{(c)}}{100} \right\rfloor, \quad n_{\text{L}}^{(c)} = N_{total} - n_{\text{F}}^{(c)}. \quad (9)$$

This produces view-adaptive patch sets that highlight diagnostically relevant regions.

### 3.4 Hierarchical Cluster-Guided Visual Tokenization for Summarization Task

Although attention heatmaps and clustered patches provide strong cues for pathology classification, directly concatenating these patches for report summarization introduces two limitations. First, independent patch extraction discards important structural information such as spatial position, cluster membership, and source view. Second, patches selected solely based on classification saliency may not be optimal for the downstream summarization task. To address these issues, we construct a hierarchical visual token representation that preserves structural context and dynamically supplements classification-driven patches with additional tokens relevant for text generation.

The attention heatmaps and clusters $\mathcal{C}^{\text{F}}, \mathcal{C}^{\text{L}}$ obtained from the interpretability stage are therefore used to construct structured visual tokens for the downstream T5-based report summarizer. Let $I^{\text{F}}, I^{\text{L}}$ denote the $224 \times 224$ frontal and lateral images. A ViT encoder (Dosovitskiy et al., 2020) produces $\mathbf{Z}_{img}^{(v)} = \mathcal{V}(I^{(v)}) \in \mathbb{R}^{B \times 197 \times D_{\text{vit}}}$, where 196 tokens correspond to the $14 \times 14$ spatial grid. These tokens are projected into the T5 embedding space $\mathbf{E}_{img}^{(v)} \in \mathbb{R}^{B \times 197 \times D_{\text{t5}}}$, $D_{\text{t5}} = 512$. For each view $v$, a hierarchical visual token sequence is constructed.

**Global token.** A global representation is obtained from the CLS token and augmented with learnable view and token-type embeddings:

$$\mathbf{t}_{global}^{(v)} = \text{CLS}^{(v)} + \mathbf{e}_{view}^{(v)} + \mathbf{e}_{type}^{(0)}. \quad (10)$$

**Cluster tokens.** For each cluster $c \in \mathcal{C}^{(v)}$, patch embeddings are aggregated using attention weights. Cluster geometry features (centroid, extent, size, and mean attention) are encoded through a small MLP to produce $\mathbf{e}_{geom}, c$. The cluster token is defined as

$$\mathbf{t}_c^{(v)} = \left( \sum_i w_i \mathbf{E}_{img,i}^{(v)} \right) + \mathbf{e}_{geom}, c + \mathbf{e}_{view}^{(v)} + \mathbf{e}_{type}^{(1)}. \quad (11)$$

This representation preserves cluster-level structure while providing intermediate semantic context between local patches and global image tokens.

**Dynamic extra patches.** To reduce over-reliance on classification-driven patches, remaining token capacity is filled using two lightweight scoring modules. An image-aware selector scores candidate patches using the global image token, while a text-aware selector additionally conditions on mean-pooled textual findings embeddings. These modules identify additional patches relevant to the summarization objective.

Each view produces a fixed set of n tokens, including global, cluster, and selected patch tokens. Frontal and lateral tokens are concatenated to form the visual sequence $\mathbf{E}_{vis} \in \mathbb{R}^{B \times 128 \times D_{\text{t5}}}$. Given textual findings embeddings $\mathbf{E}_{text}$, the multimodal representation becomes $\mathbf{E}_{fused} = [\mathbf{E}_{text}; \mathbf{E}_{vis}]$ with

| Model | B-4 | R-L | BScore | CBert | RadG | |
|---|---|---|---|---|---|---|
| **Text-Only Baselines** | | | | | | |
| BioBART | 19.71 | 43.68 | 90.66 | 58.84 | 36.68 | 45.33 |
| T5 | 21.83 | 44.61 | 91.45 | 38.49 | 38.26 | 46.79 |
| **Full-Image Multimodal Models (Single vs Multi-View)** | | | | | | |
| ViT + T5 (Frontal View) | 17.71 | 52.30 | 92.05 | 59.90 | 44.70 | 53.45 |
| ViT + T5 (Lateral View) | 20.30 | 56.42 | 92.74 | 59.06 | 49.29 | 56.89 |
| ViT + T5 (Frontal + Lateral Views) | 24.11 | 61.09 | 93.54 | 63.62 | 54.69 | 61.30 |
| **Region-Guided Multimodal Models (ROI-based Visual Context)** | | | | | | |
| ROI (Segmented Regions) | 25.13 | 58.04 | 92.97 | 55.61 | 51.08 | 56.81 |
| ROI Patches (Frontal View) | 24.95 | 57.45 | 91.64 | 60.57 | 53.92 | 57.68 |
| ROI Patches (Lateral View) | 25.79 | 58.36 | 91.63 | 62.57 | 54.92 | 57.08 |
| Random Patches (Frontal + Lateral Views) | 25.35 | 58.55 | 92.67 | 62.91 | 54.97 | 58.22 |
| **ViTAS (Selective Informative Patches, Ours)** | **29.25** | **69.83** | **95.61** | **70.57** | **63.92** | **61.68** |

Table 1: Ablation study on MIMIC-CXR summarization evaluating (1) text-only models, (2) full-image multimodal inputs (single vs multi-view), (3) ROI-guided visual context, and selective patch-based visual evidence. Results show that selecting salient visual evidence is more effective than simply increasing the amount of visual input. The evaluation metrics include BLEU-4 (B-4), ROUGE-L (R-L), BERTScore (BScore), CheXbert score (CBert), and RadGraph score (RadG).

corresponding attention mask $\mathbf{M}_{fused}$. The T5 decoder then generates the final impression using teacher forcing and cross-entropy loss.

# 4 Experiment and Result

## 4.1 Experimental Setup

### 4.1.1 Dataset Description

The MIMIC-CXR dataset (Johnson et al., 2019) is a large-scale radiology report summarization benchmark comprising 377,110 chest X-ray images and 227,827 corresponding reports. We adopt an 8:1:1 split for training, validation, and testing.

### 4.1.2 Evaluation Metrics.

We evaluate generated summaries along two dimensions: similarity and factual correctness. For similarity, we use ROUGE-L F1 (Lin, 2004), along with BERTScore (Zhang et al., 2020), BLEU-4 (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005). For factual correctness, we employ CheXbert (Smit et al., 2020) to compute label-based F1 scores and RadGraph (Jain et al., 2021) to assess clinical entity and relation accuracy.

Refer to Appendix A for implementation details.

## 4.2 Rethinking Visual Input in Multimodal Radiology Summarization

We challenge two long-standing assumptions in multimodal medical report summarization. Prior radiology-report studies have repeatedly shown that multimodal models add little value when transforming richly detailed FINDINGS (already written by radiologists with all key image information) into the concise IMPRESSION section, attributing the gap to failed cross-modal interactions between text and images. At the same time, a large body of work in medical document understanding has operated under the heuristic that "the more visual input, the better," assuming that simply concatenating more text and images will monotonically improve performance. Our ablation study on MIMIC-CXR provides the first rigorous evidence that *both assumptions, while influential, do not fully account for the role of visual information in report summarization*. Table 1 directly tests these claims.

We first establish strong text-only baselines with BioBART (Yao et al., 2023) and T5 (Raffel et al., 2020). Introducing full-image visual context with ViT+T5 consistently improves performance across all evaluation metrics: BLEU-4, ROUGE-L, BERTScore, CheXbert, and RadGraph, with gains of 7.69% (frontal view alone), 11.81% (lateral view), and 16.48% when both views are combined-clearly demonstrating the value of multi-view radiology images over pure text and refuting the claim that multimodal models provide little benefit in the FINDINGS $\rightarrow$ IMPRESSION step.

However, simply increasing the amount of visual information does not guarantee further improvements, challenging the common "more is better" assumption. When using region-guided inputs, segmented ROIs provide a modest gain of BLEU-4 of 3.3% over text-only baselines, while ROI patches-frontal (+2.42%) and lateral (+2.12%)-and even random patches from both views (+2.52%) show little additional benefit compared to the full multi-view setting, including all metrics. The plateau highlights a key challenge in multimodal learning: uncurated or noisy visual regions can dilute cross-modal alignment rather than strengthen it, emphasizing that quality and relevance of visual input matter more than sheer volume.

Our proposed ViTAS model, which selectively extracts only the most informative patches via attention-guided selection, breaks this plateau decisively. It achieves the highest overall gain of 25.22% over the text-only baseline and outperforms the strongest full-image multi-view baseline, delivering substantially larger gains especially on the primary metrics. Critically, the improvements are most pronounced precisely in the FINDINGS $\rightarrow$ IMPRESSION transformation-the very stage where all previous multimodal systems had failed. Unlike prior multimodal approaches that fed full uncurated images (or naive ROI patches) directly into LVLMs and domain-specific fusion

| Model | B-4 | R-L | BScore | CBert | RadG |
|---|---|---|---|---|---|
| **Text-Only Baselines** | | | | | |
| PG (Sim et al., 2025) | – | 36.84 | – | 54.44 | 26.72 |
| BART (Sim et al., 2025) | – | 47.00 | – | 65.39 | 40.90 |
| GSum (Sim et al., 2025) | – | 42.98 | – | 59.60 | 35.18 |
| WGSum (Sim et al., 2025) | – | 41.55 | – | 58.92 | 31.17 |
| Llama-3.1-8B (Zhao et al., 2025) | 11.67 | 32.46 | 54.56 | 69.03 | 29.16 |
| Mistral-7B (Zhao et al., 2025) | 8.67 | 29.81 | 51.58 | 66.33 | 24.91 |
| Gemma-2-9b-it (Zhao et al., 2025) | 10.16 | 30.71 | 52.95 | 68.42 | 27.44 |
| **Multimodal Baselines** | | | | | |
| Flan-T5 + ViT (Wang et al., 2023b) | 25.87 | 47.86 | 64.74 | 77.93 | 51.84 |
| OFA + ResNet (Kim et al., 2023b) | 25.58 | 47.75 | 64.80 | 76.29 | 50.96 |
| BLOOMZ-7b (Karn et al., 2023) | 25.32 | 47.48 | 63.61 | 74.34 | 49.00 |
| BioBART (Wu et al., 2023) | 22.97 | 46.15 | 63.43 | 75.14 | 48.04 |
| BERT + CvT-21 (Nicolson et al., 2023b) | 17.97 | 44.14 | 61.47 | 71.67 | 44.95 |
| VG-BART (MHA) (Sim et al., 2025) | – | 46.89 | – | 65.33 | 41.10 |
| VG-BART (Dot) (Sim et al., 2025) | – | 46.68 | – | 65.59 | 40.92 |
| LLaVA-1.5 (Sim et al., 2025) | – | 44.32 | – | 62.14 | 38.81 |
| Qwen-VL (Sim et al., 2025) | – | 44.30 | – | 63.08 | 38.83 |
| DeepSeek-VL (Sim et al., 2025) | – | 43.12 | – | 62.95 | 37.79 |
| LLaVA-Med (Sim et al., 2025) | – | 44.53 | – | 62.98 | 38.81 |
| **ViTAS (Ours)** | **29.25** | **69.83** | **95.61** | | |

Table 2: Performance comparison of SOTA models on the MIMIC-CXR test set. Text-only results are shown at the top, followed by multimodal models.

layers-thereby suffering from visual noise and failing to learn meaningful cross-modal interactions during summarization. Our framework first isolates lung ROIs with ensemble MedSAM2, performs bidirectional spatial cross-attention across complementary views, and then dynamically selects only the highest-Shapley-attention patches. This eliminates irrelevant regions, enforces clinically grounded text-image alignment, and is what finally enables multimodal models to outperform strong text-only baselines and produce accurate IMPRESSIONS where every previous system plateaued.

### 4.3 Comparison with State-of-the-Art Models

Table 2 presents the performance of text-only and multimodal models on the MIMIC-CXR test set. Text-only baselines achieve moderate performance, with BART obtaining the highest ROUGE-L (47.00%) and Llama-3.1-8B achieving highest CheXbert score (69.03%), highlighting the limitations of using text alone. Multimodal models consistently outperform text-only baselines. For example, Flan-T5 + ViT reaches 25.87% BLEU-4 and 77.93% CheXbert score, confirming that visual information enhances report summarization. Our ViTAS model achieves the best results with 29.25% BLEU-4 and 69.83% ROUGE-L, demonstrating that selectively leveraging salient visual evidence across multi-view images effectively guides medical report summarization.

### 4.4 Qualitative Analysis

Table 3 demonstrates that the Swin Transformer V2 backbone, which generates the attention heatmaps, achieves markedly higher AUROC on lateral views than on frontal views, with the largest gains on edema, pleural effusion, consolidation, and cardiomegaly. This confirms that lateral projections supply complementary diagnostic cues, directly motivating the bidirectional spatial cross-attention and Shapley-value view weighting in our multi-view fusion stage.

Figure 4 shows the full attention-driven pipeline. MedSAM2 ensemble segmentation isolates the pulmonary fields. The Swin Transformer V2 produces $32\times32$ heatmaps (downsampled from internal $64\times64$ features), which are overlaid on the lung ROI. DBSCAN then forms spatially coherent clusters. Finally, heatmap scores are projected onto the $14\times14$ ViT grid via overlap averaging, yielding the selected high-importance clusters. These steps eliminate background noise, convert scattered attention into anatomically meaningful groups, and ensure only the highest-Shapley-contribution patches reach the multimodal transformer.

As shown in Figure 5, the raw radiologist findings describe low lung volumes, stable cardiac contours, multiple round nodular densities (more numerous in the left upper lobe), and an additional hazy opacity in the left mid-upper lung suggestive of coinciding pneumonia, while ruling out effusion or pneumothorax. The full-image multimodal baseline correctly flags pneumonia but misses the metastatic nodular context entirely. The ROI-only model improves localization yet remains vague on the chronic-metastatic link. In contrast, our ViTAS approach (ROI patches selected via attention clustering and Shapley weighting) synthesizes the hazy opacity with the documented nodules into a precise impression. This closely matches the ground-truth expert interpretation while staying concise and actionable. ViTAS therefore outperforms the ablated variants because it feeds the T5 decoder exclusively with tokens from pathology-dominant, view-adaptive clusters, removing noise that dilutes both full-image and full-ROI baselines. The resulting impression is more faithful to the radiographic findings and clinically more useful-exactly the objective of our multi-stage framework.

**Human Evaluation** We conducted expert evaluations to correlate model-generated summaries with human judgment. A total of 100 randomly selected samples from the dataset were reviewed. Five volunteers participated, including one MBBS doctor, two final-year medical students, two radiology researchers, and two clinical radiologists. Each

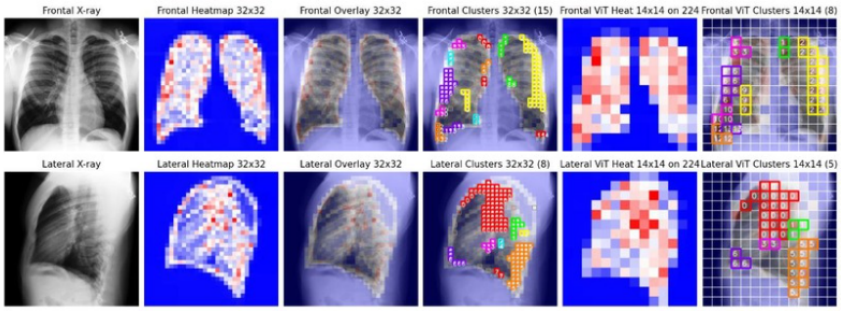


Figure 4: Attention-driven patch selection pipeline (frontal top row, lateral bottom row): original X-ray, Swin V2 32×32 heatmap (downsampled from 64×64), overlay on MedSAM2-segmented lungs, DBSCAN clusters (15 frontal, 8 lateral), projected ViT 14×14 heatmap, and final selected clusters (8 frontal, 5 lateral). Only these pathology-dominant patches feed the multimodal T5 decoder.

| Class | AP | PA | L | LL |
|---|---|---|---|---|
| Atelectasis | 0.7191 | 0.7227 | 0.8340 | 0.8128 |
| Cardiomegaly | 0.8252 | 0.8162 | 0.9035 | 0.8507 |
| Consolidation | 0.7509 | 0.7452 | 0.8485 | 0.8506 |
| Edema | 0.8689 | 0.8713 | 0.9348 | 0.9128 |
| No Finding | 0.7854 | 0.7857 | 0.8308 | 0.8241 |
| Pleural Effusion | 0.8666 | 0.8650 | 0.9315 | 0.9568 |
| Pneumothorax | 0.8447 | 0.8236 | 0.8956 | 0.7953 |
| Support Devices | 0.7930 | 0.7888 | 0.7860 | 0.6941 |
| **Avg. AUROC** | **0.8067** | **0.8023** | **0.8706** | **0.8372** |

Table 3: Per-class AUROC across different chest X-ray projection views. AP, PA, L, and LL indicate Anteroposterior, Posteroanterior, Lateral, and Left Lateral views.

Table 4: Human evaluation results based on summarization quality. Higher scores indicate better performance. Scores are reported as floats with two decimal points.

| Model | Readability | Factual Correctness | Informativeness | Redundancy | Completeness |
|---|---|---|---|---|---|
| Full-Image Multimodal Model | 4.11 | 4.22 | 3.78 | 4.00 | 3.67 |
| ROI (Segmented Regions) | 4.15 | 4.18 | 3.85 | 4.05 | 3.72 |
| ViTAS (ROI Patches) | 3.92 | 5.00 | 4.95 | 3.88 | 5.00 |

participant rated the summaries on a scale from 1 (very poor) to 5 (very good) across five dimensions: readability, factual correctness, informativeness, redundancy, and completeness. Table 4 shows that all three models performed closely, with the ViTAS model using ROI patches achieving slightly higher scores in most categories. Upon inspection, we found that using ROI patches allowed the model to better capture key clinical details, and maintain factual correctness, resulting in improved alignment with expert judgment.

## 5 Conclusion

We presented ViTAS, a selective-attention multimodal framework that challenges the conventional "more is better" paradigm in radiology report summarization. By systematically removing visual noise through ensemble segmentation, bidirectional cross-view fusion, and pathology-adaptive patch

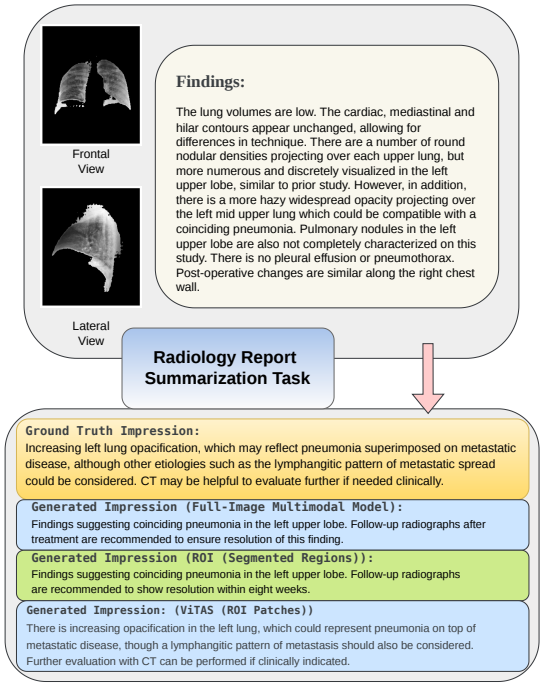


Figure 5: Qualitative analysis of left lung opacification. Ground truth shows increasing opacification. Full-image and ROI models detect pneumonia with suggested follow-up, while the ViTAS ROI-patch model closely matches ground truth, capturing pneumonia and possible lymphangitic spread with suggestions.

selection, ViTAS achieves SOTA performance on MIMIC-CXR while outperforming both full-image multimodal and strong text-only baselines. Qualitative analysis and human evaluation further confirm superior factual alignment, reduced hallucination, and better clinical utility. These results suggest that future multimodal medical summarization systems should prioritize relevance over volume in visual input. This work would inspire more interpretability-driven and noise-aware approaches in clinical NLP and medical vision-language modeling.

## Limitations

Our approach relies on accurate lung segmentation and attention heatmaps from a frozen Swin Transformer V2 classifier; errors in either stage may propagate to patch selection. The current implementation is evaluated only on MIMIC-CXR (single-institution, chest X-rays); generalization to multi-modal, multi-organ, or out-of-domain datasets remains untested.


## Acknowledgement

We used generative AI only to enhance the quality of English, with all outputs carefully reviewed and verified by the authors.


## References


Yash Kumar Atri, Shraman Pramanick, Vikram Goyal, and Tanmoy Chakraborty. 2021. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. *Knowledge-Based Systems*, 227:107152.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*.

Zhihong Chen, Maya Varma, Xiang Wan, Curtis Langlotz, and Jean-Benoit Delbrouck. 2023. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 469–484, Toronto, Canada. Association for Computational Linguistics.

J. DeGrave, J. D. Janizek, and S. Lee. 2021. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619.

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the RadSum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 478–482, Toronto, Canada. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Cassie Zhang, and Daniel Rubin. 2021. QIAI at MEDIQA 2021: Multimodal radiology report summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 285–290, Online. Association for Computational Linguistics.

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. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. Abs/2010.11929.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Akash Ghosh, Mohit Tomar, Abhisek Tiwari, Sriparna Saha, Jatin Salve, and Setu Sinha. 2024. From sights to insights: Towards summarization of multimodal clinical documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13117–13129, Bangkok, Thailand. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.

Jinpeng Hu, Zhihong Chen, Yang Liu, Xiang Wan, and Tsung-Hui Chang. 2023. Improving radiology summarization with radiograph and anatomy prompts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12066–12080, Toronto, Canada. Association for Computational Linguistics.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. 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)*, pages 388–403, Online. Association for Computational Linguistics.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Duong N. Duong, Tuan D. Bui, Andrew Y. Ng, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *NeurIPS*.

Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammed Hasanuzzaman. 2021. Multi-modal supplementary-complementary summarization using multi-objective optimization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 818–828, New York, NY, USA. Association for Computing Machinery.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathan R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Devvrat Joshi and Islem Rekik. 2025. NEURAL: Attention-guided pruning for unified multimodal resource-constrained clinical evaluation. *arXiv preprint arXiv:2508.09715*.

Sanjeev Kumar Karn, Rikhiya Ghosh, Kusuma P, and Oladimeji Farri. 2023. shs-nlp at radsum23: Domain-adaptive pre-training of instruction-tuned llms for radiology report impression generation. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 550–556, Toronto, Canada. Association for Computational Linguistics.

Gangwoo Kim, Hajung Kim, Lei Ji, Seongsu Bae, Chanhwi Kim, Mujeen Sung, Hyunjae Kim, Kun Yan, Eric Chang, and Jaewoo Kang. 2023a. KU-DMIS-MSRA at RadSum23: Pre-trained vision-language model for radiology report summarization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 567–573, Toronto, Canada. Association for Computational Linguistics.

Gangwoo Kim, Hajung Kim, Lei Ji, Seongsu Bae, Chanhwi Kim, Mujeen Sung, Hyunjae Kim, Kun Yan, Eric Chang, and Jaewoo Kang. 2023b. Ku-dmis-msra at radsum23: Pre-trained vision-language model for radiology report summarization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 567–573, Toronto, Canada. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 4152–4158. International Joint Conferences on Artificial Intelligence Organization.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2023a. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36.

Paul Pu Liang, Yun Cheng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2023b. Multimodal fusion interactions: A study of human and automatic quantification. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*, pages 425–435, New York, NY, USA. Association for Computing Machinery.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2023c. Summary-oriented vision modeling for multimodal abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2934–2951, Toronto, Canada. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*.

Nayu Liu, Kaiwen Wei, Yong Yang, Jianhua Tao, Xian Sun, Fanglong Yao, Hongfeng Yu, Li Jin, Zhao Lv, and Cunhang Fan. 2024. Multimodal cross-lingual summarization for videos: A revisit in knowledge distillation induced triple-stage training method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10697–10714.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2021. Swin transformer v2: Scaling up capacity and resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 4765–4774. SHAP: SHapley Additive exPlanations.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. 2025. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*.

Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised video summarization with adversarial LSTM networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 202–211.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. 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*, pages 5288–5304, Online. Association for Computational Linguistics.

Sourajit Mukherjee, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. Topic-aware multimodal summarization. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 387–398, Online only. Association for Computational Linguistics.

Mst. Fahmida Sultana Naznin, Adnan Ibney Faruq, Mostafa Rifat Tazwar, Md Jobayer, Md. Mehedi Hasan Shawon, and Md Rakibul Hasan. 2025. CSTRL: Context-driven sequential transfer learning for abstractive radiology report summarization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26526–26537, Vienna, Austria. Association for Computational Linguistics.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023a. e-health CSIRO at RadSum23: Adapting a chest X-ray report generator to multimodal radiology report summarisation. In *Proceedings of the*

*22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 545–549, Toronto, Canada. Association for Computational Linguistics.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023b. e-health csiro at radsum23: Adapting a chest x-ray report generator to multimodal radiology report summarisation. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 545–549, Toronto, Canada. Association for Computational Linguistics.

J. H. Niehoff, J. Woeltjen, T. Weikert, B. Stieltjes, and G. Sommer. 2023. Evaluation of the clinical performance of an ai-based application for the automated analysis of chest x-rays. *Scientific Reports*, 13(1):3680.

Keighley Overbay, Jaewoo Ahn, Fatemeh Pesaran zadeh, Joonsuk Park, and Gunhee Kim. 2023. mRedditSum: A multimodal abstractive summarization dataset of Reddit threads with images. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4117–4132, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Letitia Parcalabescu and Anette Frank. 2023. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020. Abs/2103.00020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

S. Rajaraman, Z. Xue, and S. K. Antani. 2024. Noise-induced modality-specific pretext learning for pediatric chest x-ray image classification. *Frontiers in Artificial Intelligence*, 7:1419638.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. ArXiv:1811.00347.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, Biaoyan Fang, Sarbin Ranjitkar, Arjun Burlakoti, Jamie Taylor, and Haojie Zhuang. 2025. The more, the better? a critical study of multimodal context in radiology report summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19116–19131, Suzhou, China. Association for Computational Linguistics.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling. *arXiv preprint arXiv:2004.09167*.

Omkar Chakradhar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. 2024. XrayGPT: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 440–448, Bangkok, Thailand. Association for Computational Linguistics.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Zambrano Chaves, Curtis Langlotz, Akshay Chaudhari, and John Pauly. 2023. Radadapt: Radiology report summarization via lightweight domain adaptation of large language models. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 637–654, Toronto, Canada. Association for Computational Linguistics.

Tongnian Wang, Xingmeng Zhao, and Anthony Rios. 2023a. UTSA-NLP at RadSum23: Multi-modal retrieval-based chest X-ray report summarization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 557–566, Toronto, Canada. Association for Computational Linguistics.

Tongnian Wang, Xingmeng Zhao, and Anthony Rios. 2023b. Utsa-nlp at radsum23: Multi-modal retrieval-based chest x-ray report summarization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 557–566, Toronto, Canada. Association for Computational Linguistics.

Jinge Wu, Daqian Shi, Abul Hasan, and Honghan Wu. 2023. Knowlab at radsum23: comparing pre-trained language models in radiology report summarization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 535–540, Toronto, Canada. Association for Computational Linguistics.

Yuhao Yao, Jian Peng, Kevin Zhang, Nanyun Peng, Ruiqi Wang, Thomas Searle, Mostofa Patwary, Shashi Narayan, Lance M. Jones, and Cho-Jui Hsieh. 2023. Biobart: Pretraining and evaluation of a

biomedical generative language model. In *Proceedings of the Thirty-Seventh Conference on Artificial Intelligence (AAAI)*.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Yanghai Zhang, Ye Liu, Shiwei Wu, Kai Zhang, Xukai Liu, Qi Liu, and Enhong Chen. 2024. Leveraging entity information for cross-modality correlation learning: The entity-guided multimodal summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9851–9862.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213, Brussels, Belgium. Association for Computational Linguistics.

Ruoqing Zhao, Xi Wang, Hongliang Dai, Pan Gao, and Piji Li. 2023. Medical report generation based on segment-enhanced contrastive representation learning. *arXiv preprint arXiv:2312.15869*.

Xingmeng Zhao, Tongnian Wang, and Anthony Rios. 2025. Simplified rewriting improves expert summarization. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3075–3097.

Hong-Yu Zhou, Yizhou Yu, Cheng Wang, Shu Zhang, Yutong Xiao, Yizhou Sun, Jianxin Liu, Qianni Zhou, Lei Yu, Xiang Bai, and 1 others. 2023. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature Biomedical Engineering*, 7(6):743–755.

## A Implementations and Model Details.

We implement our model in PyTorch on an NVIDIA A100 GPU. The lung ROI extraction uses MedSAM2 with five bounding box prompts (one reference and four shifts of 32-48 pixels), followed by morphological closing with a $5 \times 5$ kernel. For the dual-view classifier, images are resized to $256 \times 256$ and processed using two Swin Transformer V2-Small backbones, with bidirectional cross-attention inserted after stage 3 (8 heads, embedding dimension 768). The model is trained for 100 epochs using the Ranger optimizer (learning rate $3 \times 10^{-4}$, weight decay 0.05), cosine annealing with 100 warmup epochs, and batch size 32. For interpretability, Grad-CAM-style backpropagation through the final cross-attention layer produces $32 \times 32$ attention maps, upsampled to full resolution, while Shapley estimation requires 4 forward passes per sample. DBSCAN clustering is applied with $\varepsilon = 1.8$ and minPts = 4 on the top 25% attention regions. In the summarization stage, ViT-B/16 encoders ($224 \times 224$) generate 64 tokens per view (128 total), projected to 512 dimensions and fused with text embeddings for a T5-small decoder. The model is trained for 100 epochs using Adafactor (learning rate $1 \times 10^{-3}$), batch size 32, label smoothing 0.1, and maximum target length 256.