Toward a Unified Multimodal Framework for Chest X-ray Analysis: A Hybrid Convolutional–State Space Approach

Adjia Ndeye Penda Dieng*, Samawel Jaballi†, Henri Nicolas†, Mounir Zrigui*
*Univ. of Monastir, RLANTIS, Monastir, Tunisia
†Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, Talence, France

***Abstract*—Chest radiography remains central to first-line diagnosis, but interpretation is affected by workload, inter-observer variability, and the fragmented design of many AI pipelines. Most published systems treat classification, segmentation, and report generation independently, which limits multimodal consistency and weakens clinical integration. We present MAF-Net, a unified framework for chest X-ray triage, lung segmentation, and automated report generation. The core design combines a hybrid ResNet–VisionMamba visual encoder with BioClinicalBERT and asymmetric cross-attention. This common backbone is reused across three task-specific heads in order to preserve architectural coherence while controlling model complexity. On publicly available chest X-ray benchmarks, MAF-Net reaches 96.45% accuracy and 97.44% sensitivity for triage, Dice 0.9491 for lung segmentation, and up to 1.2181 CIDEr for report generation. These results show that a single multimodal design can remain competitive across heterogeneous tasks while remaining coherent across heterogeneous clinical objectives.**

***Index Terms*—Medical imaging, Deep learning, Multimodal systems, Image segmentation, Natural language generation**

I. Introduction

Chest radiography is among the most frequently prescribed imaging procedures worldwide, yet its interpretation remains sensitive to reader fatigue, case volume, and inter-observer variability [1], [5], [10], [15], [21]. For computer-aided triage, this creates a practical need for systems able to combine image evidence and report text rather than treating them as isolated signals [16].

Recent work has improved chest X-ray analysis with convolutional networks, transformers, and state-space models [9], [11], [17], [23]. However, many methods still optimize one task at a time: classification, segmentation, or report generation. This separation increases engineering cost and prevents architectural consistency across downstream tasks. In addition, pure attention models often remain less sample-efficient than convolutional designs on moderate medical datasets [14].

This work focuses on a central question in multimodal medical AI: can a single design support abnormality triage, lung segmentation, and report generation without sacrificing performance? A unified architecture matters because it reduces engineering fragmentation and keeps image–text representations consistent from screening to report drafting. Our main contributions are: 1) a unified three-head framework built on a shared hybrid ResNet–VisionMamba encoder and BioClinicalBERT [2]; 2) a controlled comparison of six visual encoder configurations for triage, followed by progressive text-encoder unfreezing; and 3) a compact empirical validation on public chest X-ray datasets showing competitive results for the three tasks.

II. Related Work

Visual encoders for chest X-rays. Convolutional networks remain strong baselines in thoracic imaging because their locality prior is well matched to moderate dataset sizes [14], [17]. Vision transformers improve long-range modeling but often require more data and heavier regularization [9]. More recently, state-space models such as Mamba and VisionMamba have offered a promising compromise between long-context processing and computational efficiency [11], [23].

Medical vision–language modeling. Biomedical multimodal pretraining has been explored with systems such as GLoRIA and BioViL-T, which align image and text features to improve downstream retrieval and classification [6], [12]. These models demonstrate the value of clinical text, but they typically target a narrow family of downstream tasks rather than an integrated triage-to-report pipeline [3], [4].

Segmentation and report generation. In segmentation, U-Net and its transformer or Mamba variants remain reference baselines [8], [18], [20]. In report generation, approaches such as R2Gen optimize textual quality from chest X-ray features [7], but most prior systems remain task-specific. Our difference is therefore not a new standalone module, but a shared multimodal backbone reused across classification, segmentation, and generation under one controlled protocol.

III. Method

A. Unified Architecture

MAF-Net receives chest radiographs and associated clinical text, then routes shared multimodal representations to three task-specific modules: C-MAF for triage classification, S-MAF for lung segmentation, and D-MAF for report generation. The visual backbone couples a ResNet-50 branch, which preserves local texture and boundary cues, with a VisionMamba branch, which captures longer-range dependencies with linear-time sequence modeling [11], [23]. Text features are extracted with BioClinicalBERT and fused with visual features through asymmetric cross-attention.

Formally, for frontal and lateral radiographs I_f and I_l, and report tokens $T = (t_1, \ldots, t_L)$, the framework defines a family of task mappings

$$f_\theta^{(m)} : (I_f, I_l, T) \to \mathcal{Y}^{(m)}, \quad m \in \{C, S, D\}, \tag{1}$$

where $\mathcal{Y}^{(C)} = \{0, 1\}$ for triage, $\mathcal{Y}^{(S)} = [0, 1]^{H \times W \times K}$ for segmentation, and $\mathcal{Y}^{(D)} = \Delta(\mathcal{V})^{L'}$ for report generation. This formulation highlights that the same multimodal input supports three complementary outputs with a shared encoder.

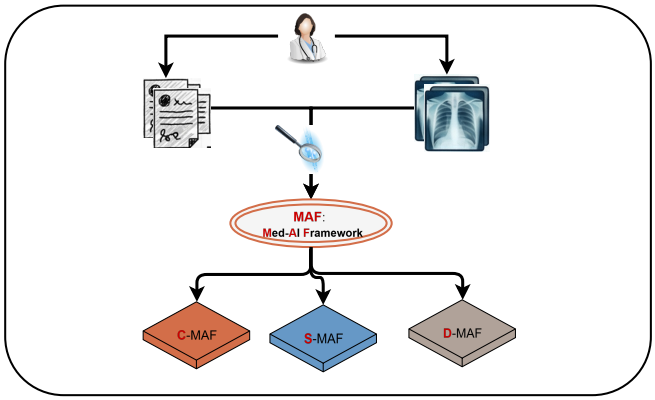

Fig. 1. Overview of the unified MAF-Net framework. A shared hybrid ResNet–VisionMamba visual encoder and a BioClinicalBERT text encoder support the three task-specific modules through asymmetric cross-attention.

B. Task Heads

Classification: C-MAF pools multimodal features and predicts normal vs. abnormal status. We first compare six visual encoder choices with the text encoder frozen, then fine-tune the best configuration by progressively unfreezing the last BioClinicalBERT layers. Let $\mathbf{F}_{\text{cat}}$ denote the concatenation of pooled visual features, text features, and the attended multimodal vector. The prediction is obtained by

$$\hat{y} = \text{softmax}\big(\mathbf{W}_2 \ \text{ReLU}(\mathbf{W}_1 \mathbf{F}_{\text{cat}} + \mathbf{b}_1) + \mathbf{b}_2\big), \tag{2}$$

optimized with weighted cross-entropy,

$$\mathcal{L}_{\text{C-MAF}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{c=1}^{2} w_c \, y_i^{(c)} \log \hat{y}_i^{(c)}. \tag{3}$$

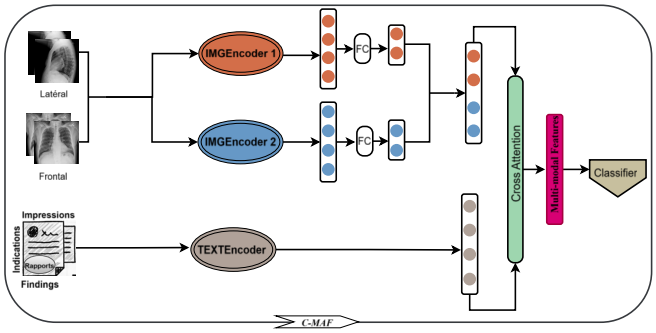

Fig. 2. C-MAF architecture. Frontal and lateral views are encoded, fused by cross-attention guided by BioClinicalBERT, and classified with a lightweight MLP head.

Segmentation: S-MAF uses the same hybrid encoding principle inside a U-shaped decoder. Cross-attention injects report-derived cues into bottleneck features, which is useful when reports mention laterality or localized opacities.

Multi-scale hybrid encoding: at each encoder scale ℓ, ResNet and VisionMamba branches run in parallel and are fused by concatenation:

$$\mathbf{F}_{\text{res}}^{\ell} = \text{ResNetBlock}(\mathbf{F}^{\ell-1}) \in \mathbb{R}^{C_\ell \times H_\ell \times W_\ell}, \tag{4}$$

$$\mathbf{F}_{\text{mamba}}^{\ell} = \text{VMambaBlock}(\mathbf{F}^{\ell-1}) \in \mathbb{R}^{C_\ell \times H_\ell \times W_\ell}, \tag{5}$$

$$\mathbf{F}_{\text{fused}}^{\ell} = [\mathbf{F}_{\text{res}}^{\ell} \oplus \mathbf{F}_{\text{mamba}}^{\ell}] \in \mathbb{R}^{2C_\ell \times H_\ell \times W_\ell}. \tag{6}$$

Bottleneck textual fusion: at the bottleneck, multi-head cross-attention ($h = 4$ heads) integrates BioClinicalBERT token representations:

$$\mathbf{V}_{\text{ctx}}, \boldsymbol{\alpha} = \text{MultiHeadAttn}(\mathbf{V}_{\text{seq}}, \mathbf{T}_{\text{emb}}, \mathbf{T}_{\text{emb}}) \in \mathbb{R}^{B \times 3136 \times 128}. \tag{7}$$

Progressive up-sampling yields a binary lung mask:

$$\hat{M} = \sigma\Big(\text{Conv}_{1\times1}(\mathbf{D}_{\text{up}}^2)\Big) \in [0, 1]^{1 \times 224 \times 224}. \tag{8}$$

The model is trained with the composite objective

$$\mathcal{L}_{\text{S-MAF}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}. \tag{9}$$

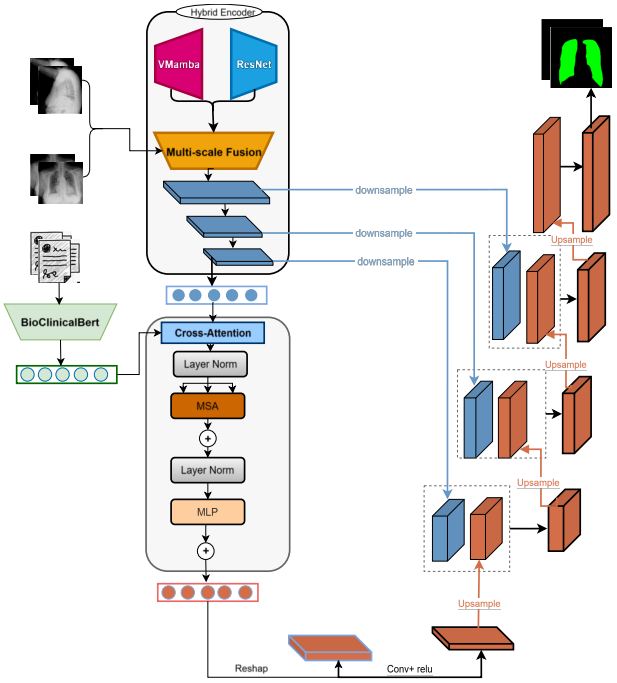

Fig. 3. S-MAF architecture. A hybrid U-Net combines ResNet and VisionMamba encoders, textual fusion at the bottleneck, and a progressive decoder with skip connections.

Report generation: D-MAF encodes frontal and lateral views together with text context and decodes a report from three memory streams. We study two encoder assignments, α-D-MAF and β-D-MAF, which swap the ResNet and VisionMamba branches across views while keeping the rest of the architecture unchanged.

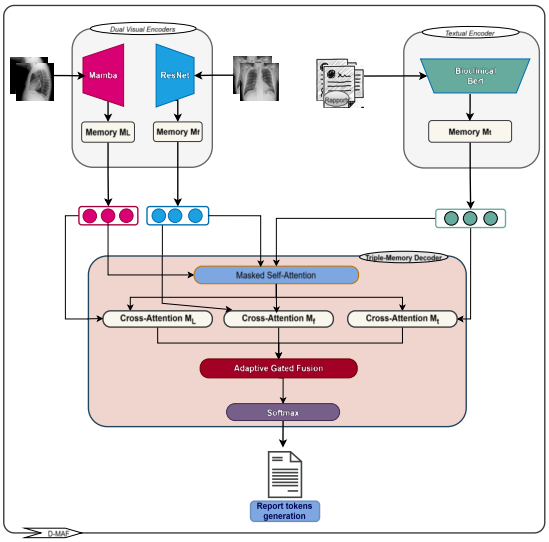

Fig. 4. D-MAF architecture. A triple-memory decoder with constrained adaptive gating is evaluated under two encoder assignments, which swap ResNet-50 and VisionMamba across the frontal and lateral views.

The fusion stage follows an asymmetric cross-attention design in which textual features query the visual token set:

$$\mathbf{Q} = \mathbf{W}^{Q}\mathbf{F}_t, \tag{10}$$

$$\mathbf{K} = \mathbf{W}^{K}\mathbf{V}, \tag{11}$$

$$\boldsymbol{\alpha} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right), \tag{12}$$

$$\mathbf{F}_{\mathrm{fused}} = \boldsymbol{\alpha}\,\mathbf{W}^{V}\mathbf{V}. \tag{13}$$

This choice is computationally lighter than full joint self-attention and reflects a clinical prior: report semantics should guide the search for relevant image evidence. In practice, we use a two-phase optimization protocol. Phase 1 freezes BioClinicalBERT to isolate the effect of the visual encoder, whereas Phase 2 progressively unfreezes the last language layers to improve visuo-semantic alignment.

IV. Experimental Setup

We evaluate C-MAF and D-MAF on the Indiana University Chest X-ray collection [13]. After filtering, classification uses 3,376 complete cases (56% abnormal), while report generation uses 3,592 paired samples with an 80/10/10 split and $\times 4$ augmentation for training. S-MAF uses 566 image-mask-report triplets with an 85/15 train/validation split. Metrics are accuracy, balanced accuracy, sensitivity, specificity, AUC, and macro-F1 for classification; Dice and IoU for segmentation; and BLEU, METEOR, ROUGE-L, CIDEr, and embedding-based semantic metrics for generation.

All models are implemented in PyTorch on a single 15 GB GPU. Images are resized to 224×224. Unless stated otherwise, optimization uses AdamW with learning rate 10^{-4} and weight decay 10^{-5}. For D-MAF, the decoder dimension is 512 with 8 heads, 4 layers, and beam size 4.

For classification, class weights are computed from the training split to limit bias toward the majority class. For segmentation, optimization combines binary cross-entropy and Dice-sensitive supervision. For generation, we monitor both lexical-overlap and semantic metrics because short n-gram scores alone do not fully reflect clinical adequacy. This multi-metric evaluation is especially important in radiology, where two semantically equivalent reports can differ substantially in surface wording.

TABLE I
Main experimental settings used across the three modules.

Item	C-MAF	S-MAF	D-MAF
Input size	224^2	224^2	224^2
Batch size	16	4	8
Epochs	5+5	15	5
Text encoder	BioClinicalBERT	BioClinicalBERT	BioClinicalBERT
Main LR	10^{-4}	10^{-4}	3×10^{-4}
BERT LR	2×10^{-5}	—	2×10^{-5}
Key output	Normal/abnormal	Lung mask	Report tokens

To keep the comparison controlled, all six classification encoders are trained under identical optimization settings in Phase 1. This makes the visual backbone the only changing factor and allows performance differences to be interpreted as architectural effects rather than side effects of tuning. For report generation, both α-D-MAF and β-D-MAF use the same decoder size, training budget, and text encoder, so the comparison isolates the assignment of ResNet and VisionMamba to frontal and lateral views.

V. Results

A. Triage Classification

Table II shows that the best-performing triage configuration is the hybrid ResNet+VisionMamba model after Phase 2 unfreezing. Although ResNet-50 attains slightly higher raw accuracy in Phase 1, the hybrid model is preferable because it reaches the highest balanced accuracy and sensitivity, which is more appropriate for minimizing false negatives in screening. Progressive unfreezing of BioClinicalBERT further improves all metrics and yields the best final classification model.

TABLE II
C-MAF ablation and final performance.

Configuration	Acc.	Bal.Acc.	AUC	$F1_M$	Sens.
ResNet-50	**95.71**	94.25	0.983	**0.954**	96.97
VisionMamba	86.39	85.69	0.941	0.856	86.95
ViT	86.39	83.18	0.947	0.846	95.10
CheXNet	84.76	84.90	0.935	0.840	84.38
ViT+Mamba	82.25	80.61	0.934	—	—
ResNet+Mamba (Ph.1)	95.41	94.70	0.983	0.950	96.99
ResNet+Mamba (Ph.2)	**96.45**	**95.52**	**0.993**	—	**97.44**

The ablation reveals a clear hierarchy. Encoders retaining a convolutional prior (ResNet-50 and ResNet+VisionMamba) outperform pure transformer- or SSM-dominant alternatives on this moderate-sized dataset, which is consistent with prior observations on sample efficiency in medical imaging [14].

The hybrid configuration is selected because it offers the best safety-oriented operating point: its sensitivity is highest, hence the number of missed abnormal exams is minimized.

TABLE III
PHASE 1 VS. PHASE 2 FOR THE SELECTED HYBRID C-MAF MODEL.

Phase	Acc.	Bal.Acc.	Sens.	Spec.	AUC
Phase 1 (frozen text)	95.41	94.70	96.99	91.90	0.983
Phase 2 (unfrozen text)	**96.45**	**95.52**	**97.44**	**94.74**	**0.993**
Gain	+1.04	+0.82	+0.45	+2.84	+0.010

The Phase 1 to Phase 2 comparison of Table III confirms that the language encoder was already well aligned with radiology text, since gains after unfreezing remain moderate but systematic. The most notable improvement is specificity, which means fewer normal examinations are escalated unnecessarily. This is important in triage workflows, where false positives increase workload while false negatives affect safety.

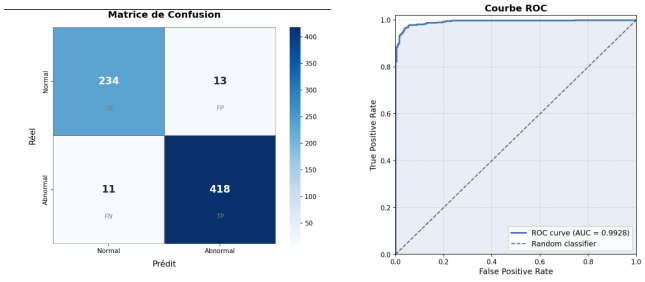

Fig. 5. ROC curve of the final C-MAF model after Phase 2 fine-tuning. The curve illustrates the discrimination ability of the selected hybrid classifier and confirms that the gain obtained after progressive BioClinicalBERT unfreezing is maintained across decision thresholds.

B. Cross-Task Performance

Table IV condenses the most important outcomes across the three modules. S-MAF improves substantially over the visual-only segmentation baseline (+0.053 Dice), suggesting that report-guided fusion helps localize clinically relevant regions. For report generation, the key difference between α-D-MAF and β-D-MAF is the assignment of ResNet and VisionMamba to frontal versus lateral views; β-D-MAF is stronger on lexical-overlap metrics, whereas α-D-MAF is marginally better on some semantic metrics. The shared conclusion is that the unified multimodal design remains effective despite task heterogeneity.

TABLE IV
MAIN RESULTS ACROSS THE THREE MODULES.

Module	Setting	Metric 1	Metric 2	Metric 3
C-MAF	Final hybrid model	Acc. 96.45%	Sens. 97.44%	AUC 0.993
S-MAF	Visual-only baseline	Dice 0.8961	IoU 0.8318	—
S-MAF	Full multimodal model	**Dice 0.9491**	**IoU 0.9035**	Epoch 13 best
α-D-MAF	Report generation	BLEU-4 0.1325	CIDEr 1.1885	VExt 0.6271
β-D-MAF	Report generation	**BLEU-4 0.1466**	**CIDEr 1.2181**	Emb. Avg. 0.9652

These results indicate that the hybrid CNN–SSM visual representation is robust across tasks with very different objectives.

The classification gain is clinically relevant because it improves sensitivity while preserving specificity; the segmentation gain shows that textual context is not only descriptive but spatially informative; and the generation results suggest that swapping the visual encoders between frontal and lateral views changes the balance between lexical precision and semantic agreement, without destabilizing the architecture.

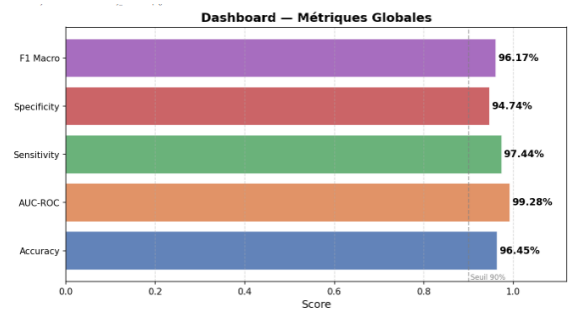

Fig. 6. Phase 2 global classification metrics for the selected C5 hybrid configuration (ResNet+VisionMamba). The figure summarizes the final screening performance after progressive BioClinicalBERT unfreezing, including accuracy, balanced accuracy, AUC, and sensitivity.

For segmentation, the multimodal model improves Dice by 0.053 over the visual-only baseline, which is a substantial gain for a mature lung-segmentation setting. This result suggests that report text provides clinically meaningful spatial priors that help the model localize lung boundaries more accurately, especially when radiographic appearance alone is ambiguous, rather than serving only as global semantic context.

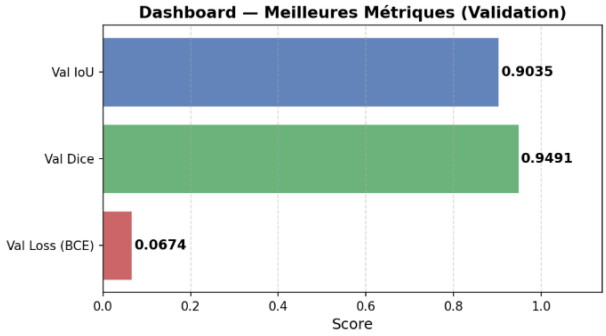

Fig. 7. Dashboard of the best segmentation metrics achieved by S-MAF. The figure summarizes the strongest validation performance of the multimodal model, highlighting its best overlap-based results for lung-mask prediction.

For report generation, the two D-MAF variants are complementary: β-D-MAF obtains the strongest lexical overlap, while α-D-MAF remains slightly stronger on semantic agreement. This pattern suggests that encoder assignment affects how the model trades local wording fidelity against global clinical coherence, but not the underlying viability of the architecture.

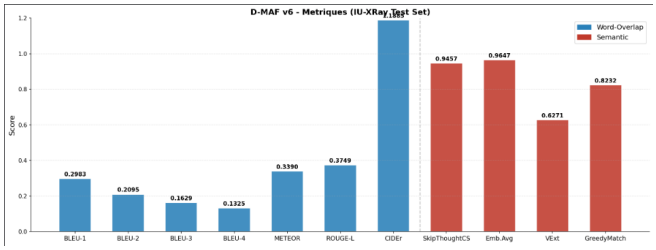

Fig. 8. Test-set performance of α-D-MAF. This encoder assignment is reported to illustrate its behavior on report-generation evaluation, where semantic agreement is slightly stronger than in the alternative assignment.

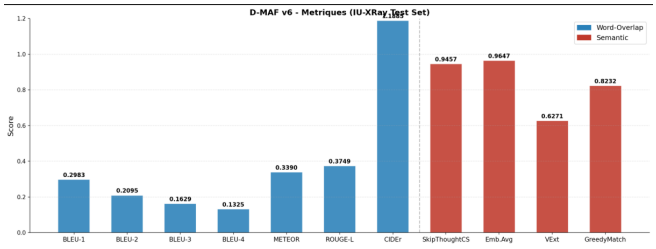

Fig. 9. Test-set performance of β-D-MAF. This second assignment highlights the complementary behavior of the model, with stronger lexical-overlap scores while remaining competitive on semantic metrics.

TABLE V
DETAILED REPORT-GENERATION RESULTS ON IU-XRAY.

Metric	α-D-MAF	β-D-MAF
BLEU-1	0.2983	**0.3174**
BLEU-2	0.2095	**0.2255**
BLEU-3	0.1629	**0.1775**
BLEU-4	0.1325	**0.1466**
METEOR	0.3390	**0.3446**
ROUGE-L	0.3749	**0.3781**
CIDEr	1.1885	**1.2181**
SkipThought	**0.9457**	0.9447
Emb. Average	0.9647	**0.9652**
VExt	**0.6271**	0.6232
Greedy Match	**0.8232**	0.8231

For report generation, Table V makes the complementarity of the two D-MAF assignments explicit: β-D-MAF dominates lexical overlap across all BLEU levels and CIDEr, whereas α-D-MAF remains marginally stronger on several embedding-based semantic criteria.

C. Comparison with Prior Work

Table VI positions the proposed method against representative prior work on Indiana University chest X-ray tasks. The comparison should be read as indicative because tasks, splits, and supervision settings are not always identical across papers. Still, it shows that the classification and segmentation branches remain competitive, while report generation is deliberately optimized for a data-limited multimodal setting rather than for very large-scale language modeling.

TABLE VI
INDICATIVE COMPARISON WITH REPRESENTATIVE PRIOR WORK.

Method	Task	Metric	Score	Year
GLoRIA [12]	Classification	Acc.	93.0	2021
BioViL-T [6]	Classification	Acc.	94.2	2023
C-MAF (ours)	Classification	Acc.	**96.45**	2026
U-Net [18]	Segmentation	Dice	0.870	2015
TransUNet [8]	Segmentation	Dice	0.897	2021
VM-UNet [20]	Segmentation	Dice	0.912	2024
S-MAF (ours)	Segmentation	Dice	**0.9491**	2026
R2Gen [7]	Generation	BLEU-4	**0.165**	2020
Wijerathna et al. [22]	Generation	BLEU-4	0.169	2022
β-D-MAF (ours)	Generation	BLEU-4	0.1466	2026
β-D-MAF (ours)	Generation	CIDEr	**1.2181**	2026

D. Discussion and Limitations

The main practical outcome is that one shared hybrid encoder remains effective across three tasks with distinct objectives. For deployment, this reduces architectural fragmentation and simplifies maintenance because triage, localization, and report drafting rely on compatible representations rather than disconnected systems. The model also behaves coherently across modules: the same encoder that improves abnormality screening supports better masks and more grounded reports, which could enable simple consistency checks between outputs. The main limitations remain dataset size, single-split evaluation, and the absence of external multi-center validation; future work will therefore target larger cohorts such as MIMIC-CXR [13], multi-site testing, and tighter multi-task co-training.

E. Qualitative and Efficiency Analysis

Figure 10 illustrates representative generated reports. In most examples, the model captures the overall clinical impression and the dominant radiographic pattern, although subtle findings and nuanced wording remain more difficult. This qualitative behavior is consistent with the quantitative results: CIDEr stays strong, while BLEU-4 remains more conservative.

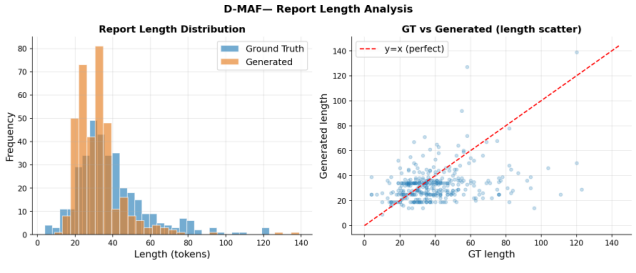

Fig. 10. Qualitative analysis of reports generated by D-MAF.

From a systems perspective, the unified approach remains computationally manageable on a single 15 GB GPU. The hybrid encoder increases memory usage compared with a

pure ResNet baseline, but the overhead stays compatible with practical research workflows. The main cost increase appears during report generation, where the decoder and multimodal memories dominate runtime.

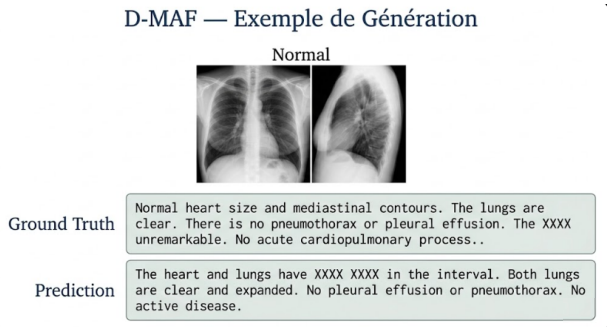

Fig. 11. Example of a report generated by D-MAF.

TABLE VII
INDICATIVE COMPUTATIONAL COST PER MODULE.

Module	Time / epoch	GPU memory	Params. (M)
C-MAF ResNet	5m 10s	11 GB	160.8
C-MAF Hybrid	5m 34s	13 GB	209.2
C-MAF Hybrid + unfreezing	8m 12s	15 GB	317.5
S-MAF	8m 15s	10 GB	145.3
α-D-MAF / β-D-MAF	15m	14 GB	~195

VI. Conclusion

This work presents a unified multimodal framework for chest X-ray triage, lung segmentation, and automated report generation. MAF-Net unifies triage classification, lung segmentation, and report generation around a shared hybrid ResNet–VisionMamba + BioClinicalBERT design. The hybrid encoder is the best choice for triage under limited data, multimodal fusion improves segmentation over a visual-only baseline, and both D-MAF configurations generate clinically coherent reports with CIDEr above 1. Future work will study larger-scale validation and tighter multi-task co-training.

References

[1] N. Akhter et al., "Trends in global diagnostic imaging utilization," *Radiography*, vol. 29, no. 1, pp. 45–52, 2023.

[2] E. Alsentzer et al., "Publicly available clinical BERT embeddings," in *Proc. 2nd Clinical NLP Workshop*, pp. 72–78, 2019.

[3] S. Jaballi, M. J. Hazar, S. Zrigui, H. Nicolas, and M. Zrigui, "ResNet-based pandemic keyword spotting in continuous multilingual speech: A study in UNESCO's audio messages for rapid health response," in *International Conference on Computational Collective Intelligence*, Springer, pp. 76–91, 2025.

[4] S. Jaballi, S. Zrigui, M. J. Hazar, H. Nicolas, and M. Zrigui, "Exploring unsupervised word representations models and neural networks for informal multilingual text against COVID-19 social media content," in *Asian Conference on Intelligent Information and Database Systems*, Springer, pp. 347–359, 2024.

[5] M. A. Bruno, E. A. Walker, and H. H. Abujudeh, "Understanding and confronting our mistakes: the epidemiology of error in radiology," *RadioGraphics*, vol. 35, no. 6, pp. 1668–1676, 2015.

[6] S. Bannur et al., "Learning to exploit temporal structure for biomedical vision-language processing," *arXiv:2301.04558*, 2023.

[7] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. EMNLP*, pp. 1875–1885, 2020.

[8] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv:2102.04306*, 2021.

[9] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[10] L. H. Garland, "Studies on the accuracy of diagnostic procedures," *Am. J. Roentgenology*, vol. 82, no. 1, pp. 25–38, 1959.

[11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *ICLR*, 2024.

[12] S.-C. Huang et al., "GLoRIA: A multimodal global-local representation learning framework," in *Proc. ICCV*, pp. 3942–3951, 2021.

[13] A. E. W. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs," *Scientific Data*, vol. 6, p. 317, 2019.

[14] C. Matsoukas et al., "Is it worth it? Comparing six deep and classical methods for unsupervised representation learning in medical imaging," *arXiv:2101.09881*, 2021.

[15] M. Nishino and P. M. Boiselle, "Radiologist workload and reading efficiency," *Am. J. Roentgenology*, vol. 217, no. 2, pp. 295–302, 2021.

[16] S. Jaballi, S. Zrigui, M. A. Sghaier, D. Berchech, and M. Zrigui, "Sentiment analysis of Tunisian users on social networks: overcoming the challenge of multilingual comments in the Tunisian dialect," in *International Conference on Computational Collective Intelligence*, Springer, pp. 176–192, 2022.

[17] P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays," in *Proc. NeurIPS*, 2017.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, pp. 234–241, 2015.

[19] S. Jaballi, M. J. Hazar, S. Zrigui, A. Mahjoubi, H. Nicolas, and M. Zrigui, "Decoding multilingual topic dynamics and trend identification through ARIMA time series analysis on social networks: a novel data translation framework enhanced by LDA/HDP models," *Journal of Electrical and Computer Engineering*, vol. 2024, no. 1, p. 6669491, 2024.

[20] J. Ruan and S. Xiang, "VM-UNet: Vision Mamba UNet for medical image segmentation," *arXiv:2402.02491*, 2024.

[21] S. Jaballi, M. J. Hazar, S. Zrigui, H. Nicolas, and M. Zrigui, "Deep bidirectional LSTM network learning-based sentiment analysis for Tunisian dialectical Facebook content during the spread of the coronavirus pandemic," in *International Conference on Computational Collective Intelligence*, Springer, pp. 96–109, 2023.

[22] T. Wijerathna et al., "Automated radiology report generation using CheXNet features with multi-scale dense transformer," in *Proc. ICIAfS*, 2022.

[23] L. Zhu et al., "Vision Mamba: Efficient visual representation learning with bidirectional state space model," *arXiv:2401.13088*, 2024.