MedTriage-LM: Anatomically Grounded Visual Phenotype Synthesis for Interpretable ED Triage

Zhixiang Lu[1,4⋆], Xiwei Liu[2*], Jinfeng Wang[3], Mian Zhou[1], Anh Nguyen[4], Jionglong Su[1(✉)], Imran Razzak[2(✉)], and Sifan Song[1(✉)]

[1] Xi'an Jiaotong-Liverpool University, Suzhou, China
[2] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[3] Kunming University of Science and Technology, Kunming, China
[4] University of Liverpool, Liverpool, UK
zhixiang@liverpool.ac.uk

Abstract. Accurate emergency severity assessment fundamentally dictates patient survival and resource allocation. However, existing automated triage pipelines rely exclusively on tabular data and clinical text, neglecting the visual semantics of patient distress. A paramount barrier to anatomically informed reasoning is the absolute absence of visual modalities in public cohorts like the MIMIC-IV-ED dataset. To overcome this limitation, this study introduces MedTriage-LM, an anatomically grounded Multimodal Large Language Model (MLLM) that algorithmically injects visual priors without requiring real patient images. The architecture maps latent clinical representations onto a canonical human template via a weakly supervised Gaussian field generator, yielding continuous, interpretable Visual Phenotype Maps (VPMs). Synergizing these spatial representations with clinical embeddings via cross-attention shifts the paradigm from numerical risk scoring toward actionable three-class triage instruction prediction: Life-saving, High-Risk Assessment, and Resource Estimation. MedTriage-LM attains state-of-the-art performance among fine-tuned open-source models and approaches the capabilities of proprietary MLLMs, while additionally providing explicit, anatomically grounded interpretability. Crucially, explicit spatial grounding empowers an efficient 8B-parameter model to approach large proprietary multimodal foundation models, generating spatially grounded textual rationales that improve clinical trustworthiness.

Keywords: Multimodal Learning · Synthetic Anatomical Prior Generation · Multimodal Clinical Reasoning · Emergency Department Triage

1 Introduction

Emergency department (ED) triage is a time-critical and high-stakes decision process that directly affects patient outcomes and resource allocation [7]. The Emergency Severity Index is a widely adopted triage scale that stratifies patients

⋆ Equal contribution. Zhixiang Lu is the project leader.

by acuity and anticipated resource needs [9]. However, despite standardized protocols, real-world triage relies heavily on the subjective judgment of medical staff under extreme time pressure, often leading to inconsistencies and suboptimal resource allocation [24].

Large-scale electronic health record (EHR) datasets have catalyzed data-driven clinical decision support. The MIMIC-IV database provides comprehensive records spanning emergency medicine [16], and its dedicated module MIMIC-IV-ED [15] offers vital signs and triage information. While prior machine learning approaches utilizing these databases have demonstrated the value of automated triage support [21, 11], they are predominantly unimodal and restricted by the fact that MIMIC-IV-ED completely lacks visual records. To overcome this fundamental limitation, our study utilizes the MIMIC-IV-Ext Triage Instruction Corpus, known as the MIETIC dataset [23]. MIETIC provides structured triage-time variables, chief complaints, instruction-answer pairs, and expert-validated triage records. Since it does not provide native bedside visual data, we synthesize anatomically grounded visual priors from clinical variables.

The inclusion of visual cues is essential, as clinicians routinely rely on visual distress signals at the bedside. Furthermore, effective clinical decision-making must be highly actionable. Therefore, rather than predicting standard numerical grades, our primary classification objective aligns with the three-class Instruction Types defined in the MIETIC dataset: Life-saving, High-Risk Assessment, and Resource Estimation. This three-class target provides a more directed and practical triage directive. To enable anatomically grounded multimodal representation learning and address potential scenarios with missing visual information, we introduce MedTriage-LM, a hybrid learned-and-deterministic pipeline that generates an interpretable Visual Phenotype Map (VPM). The VPM projects clinical evidence onto a canonical human silhouette using a Gaussian field generator, providing an explicit spatial inductive bias. MedTriage-LM then integrates structured triage variables, clinical texts, and these visual maps within a unified Transformer-based architecture [26, 17]. By employing strong clinical text encoders [1, 10] and a vision encoder [6], our model deeply aligns cross-modal features to perform three-class triage instruction prediction. The main contributions of this work are as follows:

- We propose a hybrid learned-and-deterministic Visual Phenotype Map algorithm for synthetic anatomical visual prior generation, effectively translating sparse clinical evidence into anatomically grounded visual priors to bridge the visual gap in public emergency datasets [29, 23].
- We introduce a cross-representation clinical reasoning architecture enhanced by attention mechanisms, which dynamically aligns clinical state embeddings with visual features to achieve competitive performance in the three-class triage instruction classification task [23, 27].
- We design a rationale generation framework that leverages synthesized visual features to generate clinically actionable text, providing transparent and trustworthy intermediate reasoning for triage decisions [18].

2 Method

We propose MedTriage-LM, a hybrid learned-and-deterministic, interpretable framework designed to bridge the gap between sparse clinical records and visual anatomical intuition. As illustrated in Fig. 1, our architecture comprises four tightly integrated modules: (1) Clinical State Encoder, (2) Anatomical Region Mapper, (3) Anatomical Field Generator and Heatmap Rendering, and (4) Multimodal Integration and Prediction.

2.1 Multimodal Clinical State Encoder

Given a patient encounter during triage, we denote the structured time-series and tabular data, such as vital signs and demographics, as $\mathbf{x}_s \in \mathbb{R}^{D_s}$. The unstructured clinical text, including chief complaints, is denoted as $\mathbf{x}_t$. To capture the complex physiological interactions within the structured variables, we employ an FT-Transformer [10], yielding a structured latent representation $\mathbf{h}_s = f_{\text{FT}}(\mathbf{x}_s) \in \mathbb{R}^{d_s}$. Concurrently, the unstructured clinical text is tokenized and encoded via a pre-trained Language Model (ClinicalBERT) [1, 13]. We extract the `[CLS]` token as the holistic textual embedding $\mathbf{h}_t = f_{\text{BERT}}(\mathbf{x}_t)_{[\text{CLS}]} \in \mathbb{R}^{d_t}$.

A dedicated Fusion Layer subsequently integrates these disparate modalities into a unified latent clinical state embedding $\mathbf{z} \in \mathbb{R}^d$:

$$\mathbf{z} = \sigma\left(\mathbf{W}_f[\mathbf{h}_s \parallel \mathbf{h}_t] + \mathbf{b}_f\right), \tag{1}$$

where $\parallel$ denotes concatenation, and $\sigma(\cdot)$ represents a non-linear activation. This comprehensive latent vector $\mathbf{z}$ encapsulates the patient's entire clinical profile, serving as the foundation for the subsequent anatomical reasoning.

2.2 Anatomical Region Mapper

To introduce an explicit anatomical inductive bias without relying on real photographs, we predefine K clinical anatomical regions. An Adaptive Region Mapping Layer, parameterized by a multi-layer perceptron (MLP), projects the clinical state $\mathbf{z}$ to a region intensity vector $\boldsymbol{\mu} \in [0, 1]^K$:

$$\boldsymbol{\mu} = \text{Sigmoid}(\mathbf{W}_r\mathbf{z} + \mathbf{b}_r), \tag{2}$$

where each scalar μ_k quantifies the predicted physiological distress level in region k. Given the inherent absence of ground-truth regional distress labels in public cohorts, we extract rule-based weak labels $\mathbf{y}^{\text{weak}} \in \{0, 1\}^K$ through Chief Complaint Mapping, effectively linking explicit complaints like chest pain to the thoracic region. We enforce spatial alignment by minimizing a Weak Supervision Loss using Binary Cross-Entropy (BCE):

$$\mathcal{L}_{\text{weak}} = \sum_{k=1}^{K} \text{BCE}(\mu_k, y_k^{\text{weak}}). \tag{3}$$

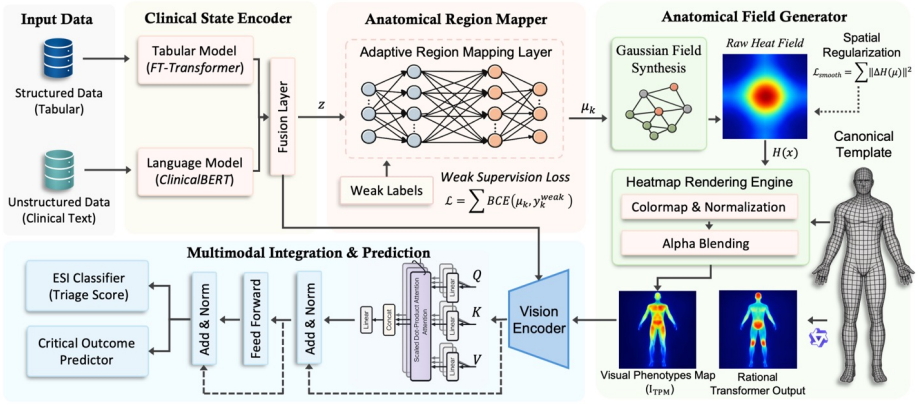


Fig. 1. Overview of the proposed MedTriage-LM framework.

2.3 Anatomical Field Generator and Heatmap Rendering

We algorithmically transform the discrete regional intensities $\boldsymbol{\mu}$ into a continuous two-dimensional visual manifold via Gaussian Field Synthesis. Assigning each region k a fixed spatial coordinate $\mathbf{c}_k$ and variance σ_k^2 on a canonical two-dimensional template Ω, we synthesize a raw heat field $H(\mathbf{u})$ for any pixel coordinate $\mathbf{u} \in \Omega$:

$$H(\mathbf{u}) = \sum_{k=1}^{K} \mu_k \exp\left(-\frac{\|\mathbf{u} - \mathbf{c}_k\|_2^2}{2\sigma_k^2}\right). \tag{4}$$

To ensure physiological coherence and strictly suppress high-frequency topological noise, we impose a spatial regularization term $\mathcal{L}_{\text{smooth}}$ over the gradient field:

$$\mathcal{L}_{\text{smooth}} = \sum_{\mathbf{u} \in \Omega} \|\nabla H(\mathbf{u})\|^2. \tag{5}$$

Finally, the Heatmap Rendering Engine applies Colormap and Normalization $\mathcal{C}(\cdot)$ to the raw field, followed by Alpha Blending with a canonical human silhouette $\mathbf{I}_{\text{canonical}}$. This produces the final Visual Phenotype Map I_{VPM}:

$$I_{\text{VPM}} = \alpha \cdot \mathcal{C}(H) + (1 - \alpha) \cdot \mathbf{I}_{\text{canonical}}. \tag{6}$$

2.4 Transformer Multimodal Architecture and Prediction

This pivotal module aligns the synthesized visual evidence with the clinical reasoning context. The rendered I_{VPM} is processed by a Vision Encoder [6] to extract a sequence of spatial visual tokens $\mathbf{V} \in \mathbb{R}^{N_v \times d_v}$.

Cross-Modal Vision-Language Alignment. To optimally fuse the synthesized visual perception with the clinical text and tabular context, we utilize a cross-attention mechanism within the Transformer architecture [17]. The clinical

embedding $\mathbf{z}$ is linearly projected to form the Query ($\mathbf{Q}$), while the visual tokens $\mathbf{V}$ are projected to form Keys ($\mathbf{K}$) and Values ($\mathbf{V}_{val}$):

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}_{val}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}_{val}. \tag{7}$$

This Scaled Dot-Product Alignment allows the clinical state to dynamically localize relevant regions within the visual phenotype, effectively bridging the semantic gap between modalities. Following standard Transformer topology [26], the attended output traverses an Add and Norm layer, a Feed-Forward Network, and a subsequent Add and Norm layer, yielding an enriched cross-modal representation $\mathbf{h}_{\text{multi}}$.

Rationale Generation via MLLM Backbone. To generate clinically actionable text rationales, we deploy a multimodal large language model backbone (Qwen3-VL-8B [2]). The cross-modal representation $\mathbf{h}_{\text{multi}}$ serves as a conditioning soft-prompt. Driven by the instruction-answer pairs from the MIETIC dataset [23], the backbone autoregressively decodes the rationale sequence $\mathbf{R} = \{w_1, w_2, \ldots, w_T\}$. The generative probability is formulated as:

$$p(\mathbf{R} \mid \mathbf{h}_{\text{multi}}) = \prod_{t=1}^{T} p(w_t \mid w_{<t}, \mathbf{h}_{\text{multi}}), \tag{8}$$

where w_t represents the generated token at time step t. This process is optimized using a standard causal language modeling loss $\mathcal{L}_{\text{gen}}$.

Prediction Heads and Optimization. Diverging from conventional basic severity scores, our primary classifier explicitly targets the three-class Triage Instructions defined by the MIETIC dataset: Life-saving, High-Risk Assessment, and Resource Estimation. We formulate the primary instruction probability distribution $\hat{\mathbf{y}}_{\text{inst}}$ and the secondary critical outcome probability $\hat{y}_{\text{crit}}$ as independent analytical blocks to ensure clarity:

$$\begin{aligned} \hat{\mathbf{y}}_{\text{inst}} &= \text{Softmax}(\text{MLP}_{\text{inst}}(\mathbf{h}_{\text{multi}})), \\ \hat{y}_{\text{crit}} &= \text{Sigmoid}(\text{MLP}_{\text{crit}}(\mathbf{h}_{\text{multi}})). \end{aligned} \tag{9}$$

The end-to-end framework is jointly optimized by minimizing the classification and generation losses alongside the anatomical regularization terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}}(\hat{\mathbf{y}}_{\text{inst}}, \mathbf{y}_{\text{inst}}) + \lambda_1 \mathcal{L}_{\text{BCE}}(\hat{y}_{\text{crit}}, y_{\text{crit}}) + \lambda_2 \mathcal{L}_{\text{weak}} + \lambda_3 \mathcal{L}_{\text{smooth}} + \lambda_4 \mathcal{L}_{\text{gen}}. \tag{10}$$

This multi-task objective encourages MedTriage-LM to learn anatomically aligned visual priors that support reliable clinical triage instructions and transparent textual rationales.

3 Experiments

3.1 Datasets and Evaluation Metrics

Datasets. MedTriage-LM is evaluated in the MIMIC-IV-Ext Triage Instruction Corpus, referred to as the MIETIC dataset [23]. Inputs are strictly constrained

Table 1. Quantitative comparison on the MIETIC dataset. Results are reported as Mean ± 95% Confidence Interval over five runs. Models are grouped by architectural paradigm and sorted chronologically. Best results are **bolded**, and second best results are underlined. Vis. indicates synthesized VPMs, not native patient images.

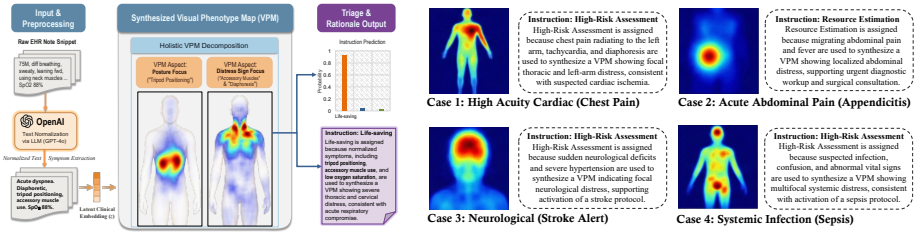
Model	Modalities			Metrics (Mean ± 95% CI)				
	Tab.	Txt.	Vis.	Acc. ↑	Macro F1 ↑	AUROC ↑	AUPRC ↑	BERTScore ↑
Traditional and Deep Tabular Models								
MLP (1986) [22]	√	×	×	0.502 ± 0.035	0.498 ± 0.042	0.821 ± 0.028	0.791 ± 0.033	N/A
Random Forest (2001) [3]	√	×	×	0.463 ± 0.024	0.458 ± 0.031	0.817 ± 0.018	0.785 ± 0.022	N/A
AutoML (2015) [8]	√	×	×	0.694 ± 0.011	0.553 ± 0.014	0.838 ± 0.012	0.811 ± 0.015	N/A
XGBoost (2016) [4]	√	×	×	0.687 ± 0.015	0.546 ± 0.019	0.824 ± 0.014	0.793 ± 0.017	N/A
TabTransformer (2020) [14]	√	×	×	0.713 ± 0.009	0.579 ± 0.012	0.849 ± 0.010	0.816 ± 0.011	N/A
Specialized Triage Pipelines								
Med-UniC (2023) [27]	×	√	√	0.728 ± 0.014	0.667 ± 0.016	0.852 ± 0.012	0.824 ± 0.015	N/A
ED-Copilot (2024) [25]	√	√	√	0.746 ± 0.011	0.683 ± 0.013	0.864 ± 0.010	0.837 ± 0.012	N/A
SDTA (2025) [23]	×	√	√	0.761 ± 0.008	0.712 ± 0.009	0.879 ± 0.007	0.846 ± 0.008	N/A
Zero-Shot Multimodal Large Language Models (with CoT [28])								
GPT-4o (2024) [19]	√	√	√	0.879 ± 0.006	0.882 ± 0.007	0.912 ± 0.005	0.887 ± 0.006	0.894 ± 0.021
GPT-5.2 (2025) [20]	√	√	√	**0.912 ± 0.003**	**0.910 ± 0.004**	**0.961 ± 0.003**	**0.942 ± 0.004**	**0.925 ± 0.013**
GLM-5 (2026) [31]	√	√	√	0.824 ± 0.008	0.831 ± 0.009	0.886 ± 0.006	0.862 ± 0.007	0.873 ± 0.025
Fine-Tuned Foundation Models								
LLaMA3.1-8B (2024) [12]	√	√	×	0.836 ± 0.007	0.843 ± 0.008	0.876 ± 0.006	0.852 ± 0.007	0.864 ± 0.036
Gemma3-4B (2025) [5]	√	√	×	0.821 ± 0.009	0.829 ± 0.011	0.864 ± 0.008	0.841 ± 0.010	0.859 ± 0.037
Qwen3-VL-8B (2025) [2]	√	√	√	0.867 ± 0.005	0.871 ± 0.006	0.906 ± 0.005	0.883 ± 0.006	0.886 ± 0.034
MedTriage-LM (Ours)	√	√	√	0.898 ± 0.002	0.895 ± 0.001	0.948 ± 0.001	0.925 ± 0.003	0.908 ± 0.009

to triage time variables, comprising structured hemodynamics and unstructured clinical complaints. To align with practical clinical workflows, our primary task is the three-class triage instruction classification: Life Saving, High Risk Assessment, and Resource Estimation.

Evaluation Metrics. Classification performance is evaluated via Accuracy and Macro F1. For critical risk stratification, specifically predicting intensive care unit admission, the Area Under the Receiver Operating Characteristic (AUROC) and Precision Recall Curve (AUPRC) are utilized, with AUPRC being essential for imbalanced clinical datasets. Rationale semantic quality is objectively measured via BERTScore [30]. All quantitative results are reported as the mean and 95% Confidence Interval (95% CI) across five independent runs.

3.2 Implementation Details

The MedTriage-LM architecture is implemented in PyTorch. The vision encoder is initialized with pretrained ViT weights, while the multimodal rationale generation backbone is driven by Qwen3-VL-8B. End-to-end training is executed on a single NVIDIA A100 GPU employing mixed precision. Optimization is performed via the AdamW optimizer, configured with a base learning rate of 2×10^{-5} and regulated by a cosine annealing scheduler. The loss weighting hyperparameters are empirically established through a comprehensive grid search on the validation set, yielding optimal values of $\lambda_1 = 0.5$, $\lambda_2 = 1.0$, $\lambda_3 = 0.5$, and $\lambda_4 = 1.0$.



(a) Inference pipeline **(b)** Diverse clinical scenarios

Fig. 2. MedTriage-LM architecture and clinical generalization. (a) The end-to-end inference pipeline, illustrated with an acute respiratory distress case. **(b)** Synthesized Visual Phenotype Maps (VPMs) accurately reflecting distinct anatomical activations across diverse high-acuity scenarios.

3.3 Main Quantitative Results

The proposed framework is benchmarked against a comprehensive suite of models. Baselines span traditional machine learning, advanced tabular architectures, specialized triage pipelines, state-of-the-art large language models, and fine-tuned foundation models, including LLaMA3.1-8B [12], Gemma3-4B [5], and Qwen3-VL-8B [2]. To ensure fair evaluation, all zero-shot models employ a Zero-Shot Chain of Thought prompting strategy [28], with complete prompt templates detailed in the supplementary material.

As demonstrated in Table 1, pure tabular methods achieve moderate AUROC but inherently lack the capacity for semantic rationale generation. Conversely, advanced zero-shot multimodal models like GPT-5.2 [20] exhibit strong reasoning capabilities with Macro F1 scores exceeding 0.9. However, relying solely on massive parameter scaling often obscures explicit clinical spatial physiology and incurs prohibitive computational costs. MedTriage-LM addresses this efficiency and interpretability gap. By synthesizing the intermediate VPM to anchor the fine-tuned Qwen3-VL-8B backbone, the specialized 8B framework achieves performance highly competitive with the large proprietary model, GPT-5.2. Notably, although advanced baselines such as GPT-4o [19] and the Qwen3-VL-8B also receive synthesized VPM inputs, MedTriage-LM outperforms them. Ultimately, MedTriage-LM closely approaches the strong GPT-5.2 baseline across all predictive metrics (trailing by merely 0.014 in Accuracy and 0.017 in AUPRC) while generating interpretable rationales (BERTScore 0.908) with stable performance. This highlights that directly providing synthetic visual inputs to general-purpose multimodal models is clinically insufficient. Instead, explicit anatomical grounding acts as a powerful inductive bias, enabling an efficient architecture to approach the performance of large proprietary models while retaining explicit anatomical interpretability.

3.4 Ablation Study

Progressive Architecture and Loss Ablation. Reconstructing MedTriage-LM (Table 2, upper) demonstrates that while text integration improves rationale

Table 2. Comprehensive ablation study integrating modality progression and hyperparameter sensitivity on the MIETIC dataset. The upper section demonstrates the sequential addition of modalities and optimization constraints (λ_1–λ_4). The lower section isolates the sensitivity of the spatial smoothness regularization (λ_3) to demonstrate its qualitative and quantitative impact. All metrics are reported as Mean ± 95% CI over five independent runs. **Bold** denotes the optimal full configuration.

Ablation Variant	Modalities			Loss Weights				Metrics (Mean ± 95% CI)		
	Tab.	Txt.	Vis.	λ_1	λ_2	λ_3	λ_4	Macro F1	AUROC	BERTScore
Progressive Architecture and Joint Optimization										
Tabular Features Only	√	×	×	0	0	0	0	0.753 ± 0.009	0.874 ± 0.008	0.791 ± 0.024
+ Clinical Text Features	√	√	×	0	0	0	0	0.820 ± 0.006	0.892 ± 0.005	0.866 ± 0.028
+ Visual Map (Unconstrained)	√	√	√	0	0	0	0	0.851 ± 0.005	0.910 ± 0.004	0.880 ± 0.019
+ Auxiliary Outcome Prediction Task (λ_1)	√	√	√	0.5	0	0	0	0.868 ± 0.004	0.922 ± 0.004	0.883 ± 0.013
+ Weak Anatomical Supervision (λ_2)	√	√	√	0.5	1.0	0	0	0.877 ± 0.004	0.931 ± 0.003	0.892 ± 0.021
+ Spatial Smoothness Regularization (λ_3)	√	√	√	0.5	1.0	0.5	0	0.883 ± 0.003	0.941 ± 0.003	0.899 ± 0.016
Full Joint Opt. (+ Gen Loss λ_4)	√	√	√	0.5	1.0	0.5	1.0	**0.895 ± 0.001**	**0.948 ± 0.001**	**0.908 ± 0.009**
Spatial Regularization (λ_3) Sensitivity Analysis										
Insufficient Smoothness ($\lambda_3 = 0.1$)	√	√	√	0.5	1.0	0.1	1.0	0.862 ± 0.004	0.915 ± 0.004	0.882 ± 0.017
Minor Blurring ($\lambda_3 = 1.0$)	√	√	√	0.5	1.0	1.0	1.0	0.872 ± 0.003	0.926 ± 0.003	0.890 ± 0.015
Over Smoothness ($\lambda_3 = 5.0$)	√	√	√	0.5	1.0	5.0	1.0	0.811 ± 0.008	0.863 ± 0.007	0.844 ± 0.022

quality, VPM synthesis provides a notable discriminative gain. Sequentially, the auxiliary outcome loss (λ_1) enhances high-risk feature learning, weak supervision (λ_2) anchors anatomical activations, and spatial smoothness (λ_3) suppresses topological noise. Joint optimization with generation loss (λ_4) achieves the best performance within the ablation suite while preserving rationale integrity.

Spatial Regularization Sensitivity. Unlike standard cross entropy objectives ($\lambda_1, \lambda_2, \lambda_4$) which remain resilient to scalar shifts, the spatial smoothness weight (λ_3) strictly dictates visual manifold topology (Table 2, lower). Insufficient weighting ($\lambda_3 = 0.1$) collapses the field into high-frequency artifacts, degrading Macro F1 to 0.860 and AUROC to 0.915. Conversely, excessive penalization ($\lambda_3 = 5.0$) forces distributions toward a uniform mean, washing out focal clinical signals and degrading rationale quality. The optimum ($\lambda_3 = 0.5$) suggests that explicit anatomical grounding is important for accurate prediction.

Clinical Interpretability. Beyond quantitative metrics, the inference pipeline (Fig. 2a) provides exceptional transparency. The synthesized VPMs qualitatively reflect distinct anatomical activation patterns across diverse high-acuity scenarios (Fig. 2b). For example, respiratory distress consistently triggers thoracic activation, whereas systemic infections exhibit biologically plausible multifocal heat distributions. This explicit spatial alignment improves the transparency of the triage decision support prior to critical triage interventions.

4 Conclusion

This study presented MedTriage-LM†, a hybrid learned-and-deterministic, anatomically interpretable multimodal framework for emergency severity assessment.

† The code is available at https://github.com/Leo1998-Lu/MedTriage-LM

The introduction of the Visual Phenotype Map algorithm successfully bridges the critical visual modality gap prevalent in public electronic health record cohorts. By synergizing synthesized anatomical priors with tabular and textual clinical states via a cross-attention architecture, the framework achieves strong performance among efficient fine-tuned models and narrows the gap to proprietary MLLMs for three-class triage instruction prediction on the MIETIC dataset. Crucially, explicit anatomical grounding empowers a highly efficient 8B parameter model to achieve predictive accuracy approaching large proprietary models, while generating transparent textual rationales grounded in the synthesized anatomical representation. Future work will focus on validating the generalized robustness of these visual maps across diverse multi-institutional clinical environments and integrating dynamic temporal sequences for continuous patient monitoring.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop pp. 72–78 (2019). https://doi.org/10.18653/v1/W19-1909
2. Bai, S., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
4. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785
5. DeepMind: Gemma 3 technical report (2025), https://arxiv.org/abs/2503.19786
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
7. Eitel, D.R., et al.: The emergency severity index triage algorithm version 2 is reliable and valid. Academic Emergency Medicine **10**(10), 1070–1080 (2003). https://doi.org/10.1111/j.1553-2712.2003.tb00577.x
8. Feurer, M., et al.: Efficient and robust automated machine learning. In: Advances in Neural Information Processing Systems (2015)
9. Gilboy, N., Tanabe, P., Travers, D., Rosenau, A.: Emergency severity index (esi): A triage tool for emergency department care, version 4: Implementation handbook. Agency for Healthcare Research and Quality (AHRQ) (2011), aHRQ Publication No. 12-0014
10. Gorishniy, Y., et al.: Revisiting deep learning models for tabular data. In: Advances in Neural Information Processing Systems (2021)
11. Goto, T., Camargo Jr, C.A., Faridi, M.K., Freishtat, R.J., Hasegawa, K.: Machine learning–based prediction of clinical outcomes for children during emergency department triage. JAMA Network Open **2**(1), e186937 (2019). https://doi.org/10.1001/jamanetworkopen.2018.6937

12. Grattafiori, A., et al.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
13. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 (2019)
14. Huang, X., Khetan, A., Cvitkovic, M., Karnin, Z.: Tabtransformer: Tabular data modeling using contextual embeddings (2020), https://arxiv.org/abs/2012.06678
15. Johnson, A., Pollard, T., Mark, R.: Mimic-iv-ed (emergency department) database. PhysioNet (2022), https://physionet.org/content/mimic-iv-ed/
16. Johnson, A.E.W., et al.: Mimic-iv, a freely accessible electronic health record dataset. Scientific Data **10**, 1–15 (2023). https://doi.org/10.1038/s41597-022-01899-x
17. Lu, Z., et al.: Deepgb-tb: A risk-balanced cross-attention gradient-boosted convolutional network for rapid, interpretable tuberculosis screening. Proceedings of the AAAI Conference on Artificial Intelligence **40**(46), 38989–38997 (Mar 2026). https://doi.org/10.1609/aaai.v40i46.41245
18. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (2017)
19. OpenAI: Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/ (2024)
20. OpenAI: Openai gpt-5 system card (2025), https://arxiv.org/abs/2601.03267
21. Raita, Y., Goto, T., Faridi, M.K., Brown, D.F.M., Camargo Jr, C.A., Hasegawa, K.: Emergency department triage prediction of clinical outcomes using machine learning models. Critical Care **23**(1), 64 (2019). https://doi.org/10.1186/s13054-019-2351-7
22. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533–536 (1986)
23. Shen, Q., Zhang, X., Ren, H., Guo, Q., Yi, Z.: Knowledge-embedded large language models for emergency triage. Knowledge-Based Systems **318**, 113431 (2025). https://doi.org/10.1016/j.knosys.2025.113431
24. Souza-Pereira, L., Ouhbi, S., Pombo, N.: A process model for quality in use evaluation of clinical decision support systems. J. of Biomedical Informatics **123**(C) (Nov 2021). https://doi.org/10.1016/j.jbi.2021.103917
25. Sun, L., et al.: Ed-copilot: reduce emergency department wait time with language model diagnostic assistance. In: International Conference on Machine Learning (2024)
26. Vaswani, A., et al.: Attention is all you need. In: International Conference on Neural Information Processing Systems. p. 6000–6010 (2017)
27. Wan, Z., et al.: Med-unic: unifying cross-lingual medical vision-language pre-training by diminishing bias. In: International Conference on Neural Information Processing Systems (2023)
28. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 24824–24837 (2022)
29. Xie, F., et al.: Benchmarking emergency department prediction models with machine learning and public electronic health records. Scientific Data **9**, 1–11 (2022). https://doi.org/10.1038/s41597-022-01782-9
30. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SkeHuCVFDr
31. ZhipuAI: Glm-5: from vibe coding to agentic engineering (2026), https://arxiv.org/abs/2602.15763