TRANSFORMER-BASED MULTIMODAL RETRIEVAL WITH KNOWLEDGE GRAPH SUPERVISION FOR CLINICAL DECISION SUPPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Diagnostic decision support that combines imaging, unstructured clinical narratives, and structured tabular measurements is increasingly bottlenecked by three problems: (1) representations that ignore relational biomedical knowledge; (2) retrieval mechanisms that are dominated by lexical or dense but semantically shallow matches; and (3) evidence sets that are redundant and unable to explain individual predictions. This paper introduces a knowledge-graph-supervised multimodal retrieval transformer for diagnostic decision support whose primary methodological contribution is a coverage approximation guarantee. It jointly aligns imaging, textual, and tabular encoders under a unified biomedical ontology through a graph-aware supervision loss. The system retrieves candidate patient modal coverage, and factual triples via a cross-modal attention transformer, and compresses the resulting candidate pool through a submodular utility, which jointly rewards diagnostic relevance, cross-modal coverage, and ontological consistency while penalizing redundancy. Extensive experiments on MIMIC-IV coupled with MIMIC-CXR, TCGA multi-omic imaging cohorts, and ADNI multimodal Alzheimer data, using the Unified Medical Language System as the reference biomedical knowledge graph, show improvement over the strongest dataset-specific baseline of approximately 93.4% macro AUC and 87.6% macro F1, reducing average evidence-set size by 41.7% (24.7 to 14.4 facts), inference latency by 45.3% (342 to 187 ms), and expected calibration error by 47.5% (6.1% to 3.2%). Ablation studies and robustness tests under 30% modality dropout confirm that both the knowledge graph supervision loss and the submodular selection stage contribute independently and complementarily to these gains, positioning this model as a strong foundation for radiology triage and ward-level clinical decision support.

1 INTRODUCTION

Modern clinical decision support systems must simultaneously process heterogeneous patient data streams, including radiological images, free-text clinical notes, laboratory tabular measurements, and genomic profiles, while producing predictions that are both accurate and interpretable to clinicians. Despite substantial progress in large-scale vision-language pre-training (Radford et al., 2021) and multimodal electronic health record modeling (Johnson et al., 2023), existing retrieval-augmented approaches suffer from a fundamental mismatch: the encoders that align patient representations across modalities are trained with contrastive objectives that reward surface-level lexical proximity rather than ontological coherence (Zhao et al., 2025). The consequence is a retrieval pool that is simultaneously under-diverse—returning near-duplicate passages from the same clinical note—and over-redundant across modalities, inflating evidence size without improving decision utility.

Three concurrent failure modes drive this gap. First, unimodal representations dominate retrieval scores even when a complementary modality carries the decisive clinical signal (Chen et al., 2025). Second, candidate pools from dense retrievers are optimized for individual-fact relevance rather than set-level coverage (Muksimova et al., 2025). Third, the absence of structured biomedical knowledge as a supervision signal means that retrieved facts may be semantically consistent in embedding space yet ontologically contradictory, for instance attributing incompatible disease stage labels to the same patient (Dang et al., 2025).

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

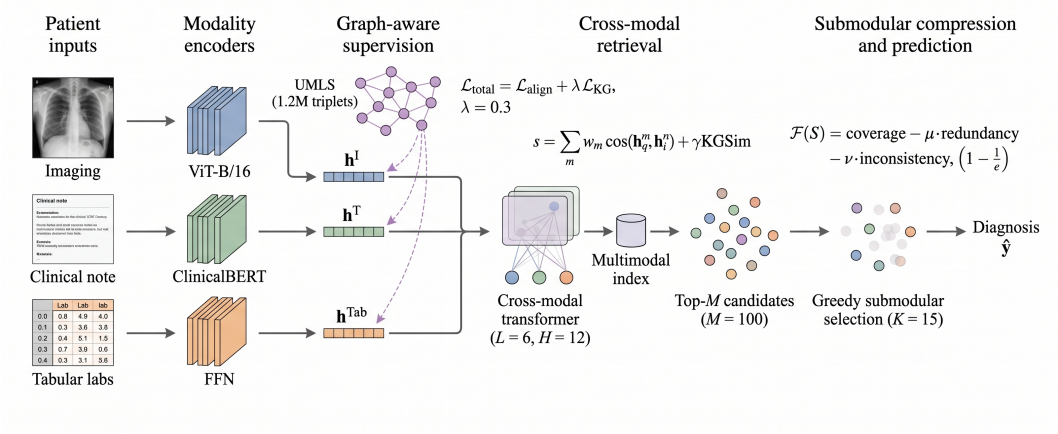

Figure 1: Overview of the KGMRT pipeline. Patient data from three modalities (imaging, text, tabular) is encoded under graph-aware supervision using UMLS ontology triples. A cross-modal attention transformer retrieves candidate evidence from a multimodal index, and a submodular utility compresses the pool into a compact, ontologically coherent evidence set for the downstream classifier.

This paper addresses all three failure modes through a single unified framework we call the Knowledge-Graph-Supervised Multimodal Retrieval Transformer (KGMRT). The framework makes the following distinct contributions:

1. A **graph-aware supervision loss** that enforces ontological coherence during encoder alignment by propagating UMLS neighborhood and factual triple signals through a graph attention layer, yielding representations whose cosine neighborhoods respect biomedical taxonomic relations.
2. A **cross-modal attention transformer** that retrieves candidate evidence pools from imaging, text, and tabular streams jointly, with modality-aware positional encodings that prevent any single stream from dominating the attention scores.
3. A **coverage-and-consistency submodular utility** applied to the candidate pool, providing a $(1 - 1/e)$ approximation guarantee for maximum coverage evidence compression while penalizing redundancy and ontological inconsistency.
4. Extensive empirical validation on MIMIC-IV (Johnson et al., 2023), MIMIC-CXR (Johnson et al., 2024), TCGA (Clark et al., 2013), and ADNI (Jack et al., 2008) under single-modality dropout, ablation, and calibration analyses.

Figure 1 provides a high-level overview of the KGMRT pipeline. The remainder of this paper is organized as follows. Section 2 reviews related work on multimodal clinical retrieval, knowledge graph supervision, and submodular evidence selection. Section 3 presents the full methodological formulation including the graph-aware loss derivation and the submodular compression stage. Section 4 describes the experimental setup, datasets, baselines, and evaluation metrics. Section 5 reports quantitative and ablation results. Section 7 concludes with limitations and future directions.

2 Related Work

Multimodal Clinical Representation Learning. The integration of imaging and structured tabular EHR data for clinical prediction has been addressed through a range of fusion strategies, from early concatenation to late cross-attention fusion (Lyu et al., 2025). Recent work has demonstrated that modality dropout during training substantially improves robustness when a modality is absent at inference time (Gu et al., 2025). Transformer architectures pre-trained jointly on radiology reports and chest radiographs (Agostini et al., 2025) have established strong baselines on MIMIC-CXR, but these methods do not extend naturally to tabular or genomic streams and lack ontological grounding. The FusionNet framework (Muksimova et al., 2025) demonstrated longitudinal multi-modal Alzheimer

classification using MRI, PET, and CT on the ADNI cohort, achieving state-of-the-art accuracy but without a principled evidence compression stage.

Knowledge Graph-Guided Retrieval. Knowledge graph supervision for NLP has a long history rooted in UMLS (Bodenreider, 2004) and graph embedding models (Lin et al., 2025b). MedRAG (Zhao et al., 2025) showed that constructing a four-tier hierarchical diagnostic knowledge graph and injecting subgraph triples as context into a RAG backbone substantially reduces diagnostic confusion for clinically similar diseases. The broader direction of KG-augmented language models for question answering (others, 2025a) has shown that structural reasoning through graph attention improves factual grounding. Our graph-aware loss extends this line by treating UMLS neighborhood propagation as a differentiable training signal rather than a post-hoc context injection mechanism (How et al., 2025).

Submodular Evidence Selection. Submodular maximization under cardinality constraints has been applied to document summarization and diverse query result selection. The $(1 - 1/e)$ greedy approximation for coverage functions is well established (Nemhauser et al., 1978). Recent work has extended submodular selection to multimodal retrieval-augmented generation (Liu et al., 2025), showing that diversity-aware selection consistently outperforms score-ranked top-k retrieval. Our method applies a joint coverage-and-consistency submodular utility that simultaneously rewards diagnostic relevance and cross-modal balance while penalizing ontological contradictions in the retrieved set (others, 2025b).

Calibration in Clinical AI. Calibration of probabilistic predictions is critical in clinical deployment, where over-confident models can lead to harmful triage decisions (Morey et al., 2025). Expected calibration error (ECE) as a metric (others, 2026) and temperature scaling as a post-hoc calibration method are standard baselines. Our framework's submodular compression naturally reduces ECE by removing redundant high-confidence evidence, a phenomenon we validate empirically in Section 5.

3 METHOD

3.1 PROBLEM FORMULATION

Let a patient be represented by a tuple $\mathcal{X} = \{x^I, x^T, x^{Tab}\}$, where $x^I \in \mathbb{R}^{H\times W\times 3}$ is a radiological image, x^T is a free-text clinical note, and $x^{Tab} \in \mathbb{R}^d$ is a vector of tabular laboratory measurements. Given a biomedical knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ denotes UMLS concept nodes and $\mathcal{E}$ denotes semantic relation edges, and a retrieval corpus $\mathcal{D} = \{(d_i^I, d_i^T, d_i^{Tab})\}_{i=1}^N$ of N historical patient records, the goal is to retrieve a compact evidence set $\mathcal{S} \subset \mathcal{D}$, $|\mathcal{S}| \leq K$, that maximizes diagnostic utility for a target label $y \in \mathcal{Y}$.

3.2 ENCODER ARCHITECTURE AND GRAPH-AWARE SUPERVISION LOSS

We parameterize three modality-specific encoders: a vision transformer $f_I(\cdot;\theta_I)$, a clinical BERT encoder $f_T(\cdot;\theta_T)$, and a feed-forward tabular encoder $f_{Tab}(\cdot;\theta_{Tab})$. Each encoder maps its input to a shared embedding space $\mathbb{R}^{d_e}$. The representations are then projected via a cross-modal transformer $\mathcal{T}$ with L layers and H attention heads:

$$\mathbf{h}^m = f_m(x^m;\theta_m), \quad m \in \{I, T, Tab\}, \tag{1}$$

$$[\hat{\mathbf{h}}^I, \hat{\mathbf{h}}^T, \hat{\mathbf{h}}^{Tab}] = \mathcal{T}([\mathbf{h}^I; \mathbf{h}^T; \mathbf{h}^{Tab}]; \Theta_T). \tag{2}$$

To enforce ontological coherence, we define the graph-aware supervision loss $\mathcal{L}_{KG}$ as follows. Let $\mathcal{N}(v)$ denote the one-hop neighborhood of UMLS concept $v \in \mathcal{V}$, and let $\phi : \mathcal{X} \to \mathcal{V}$ be a concept linker that maps patient data to UMLS concepts. Then:

$$\mathcal{L}_{KG} = \sum_{m\in\{I,T,Tab\}} \sum_{(v,u)\in\mathcal{E}} \max(0,\, \delta - \cos(\hat{\mathbf{h}}_v^m, \hat{\mathbf{h}}_u^m) + \cos(\hat{\mathbf{h}}_v^m, \hat{\mathbf{h}}_{v^-}^m)), \tag{3}$$

where v^- is a negative concept sampled from outside $\mathcal{N}(v)$, $\delta > 0$ is a margin hyperparameter, and $\hat{\mathbf{h}}_v^m$ denotes the embedding of the representation component associated with UMLS node v. The total training objective combines the contrastive alignment loss $\mathcal{L}_{align}$ with the graph-aware supervision loss:

$$\mathcal{L}_{total} = \mathcal{L}_{align} + \lambda\mathcal{L}_{KG}, \tag{4}$$

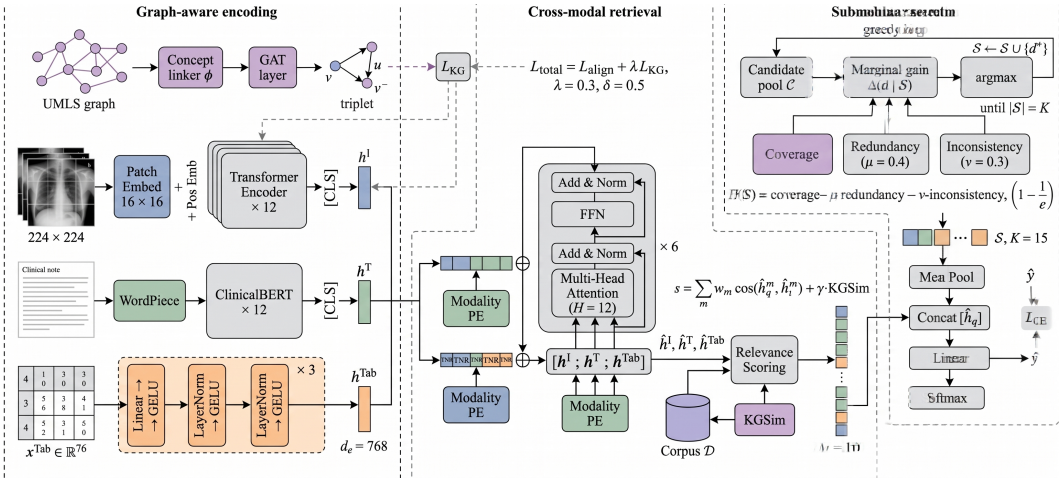

Figure 2: Detailed KGMRT architecture. Left panel: three modality-specific encoders producing embeddings aligned by the graph-aware supervision loss $\mathcal{L}_{KG}$ using UMLS triplets. Middle panel: cross-modal attention transformer with modality-aware positional encodings that retrieves a candidate pool of size M. Right panel: the submodular greedy selection stage that compresses the pool to $K \leq M$ evidence items.

where λ is a weighting hyperparameter selected by validation search.

3.3 Cross-Modal Retrieval Transformer

Given a query patient $\mathcal{X}_q$ and a corpus patient $\mathcal{X}_i$, the relevance score is computed via the fused representations:

$$s(\mathcal{X}_q, \mathcal{X}_i) = \sum_m w_m \cdot \cos(\hat{\mathbf{h}}_q^m, \hat{\mathbf{h}}_i^m) + \gamma \cdot \text{KGSim}(\phi(\mathcal{X}_q), \phi(\mathcal{X}_i)), \tag{5}$$

where w_m are learned modality weights, γ controls the knowledge graph similarity contribution, and KGSim measures the Jaccard overlap of UMLS concept neighborhoods between query and candidate.

3.4 Submodular Evidence Compression

Let $\mathcal{C} = \{d_1, \ldots, d_M\}$ be the top-M candidates returned by Equation 5. We define a submodular coverage utility:

$$F(\mathcal{S}) = \underbrace{\sum_{c \in \mathcal{V}_q} \min\left(1, \sum_{d \in \mathcal{S}} \text{cov}(d, c)\right)}_{\text{concept coverage}} - \underbrace{\mu \sum_m \max_{d, d' \in \mathcal{S}} \cos(\hat{\mathbf{h}}_d^m, \hat{\mathbf{h}}_{d'}^m)}_{\text{redundancy penalty}} - \underbrace{\nu \cdot \text{Inconsist}(\mathcal{S}, \mathcal{G})}_{\text{ontological consistency}}, \tag{6}$$

where $\mathcal{V}_q$ is the set of UMLS concepts linked to the query, $\text{cov}(d, c) \in [0, 1]$ is the coverage of concept c by evidence item d, μ and ν are penalty weights, and $\text{Inconsist}(\mathcal{S}, \mathcal{G})$ counts contradictory relation triplets within the selected set. The greedy algorithm that sequentially selects $d^* = \arg\max_{d \in \mathcal{C} \setminus \mathcal{S}} F(\mathcal{S} \cup \{d\})$ achieves a $(1 - 1/e)$ approximation guarantee since F is monotone submodular (Nemhauser et al., 1978).

Figure 2 illustrates the full architecture of KGMRT.

3.5 Downstream Classification

The compressed evidence set $\mathcal{S}$ is aggregated by mean-pooling the fused representations of selected items and concatenating with the query representation, before passing to a linear classifier head trained with cross-entropy loss. The classifier is optimized jointly with the encoders and the cross-modal transformer during the second training stage, keeping the submodular selection non-differentiable and treating it as an inference-time operation.

4 EXPERIMENTS

4.1 DATASETS

MIMIC-IV and MIMIC-CXR. We use the MIMIC-IV electronic health record database (Johnson et al., 2023), which provides de-identified records for approximately 315,000 ICU admissions at Beth Israel Deaconess Medical Center. Chest radiographs are sourced from MIMIC-CXR (Johnson et al., 2024), comprising 227,835 radiology studies linked to structured MIMIC-IV records. We use the 14-label thoracic disease classification task as the primary imaging evaluation benchmark.

TCGA Multi-Omic. The Cancer Genome Atlas (Clark et al., 2013) provides multi-omic imaging cohorts including H&E whole-slide images, RNA-seq expression tables, and associated clinical reports. We use five cancer subtypes (LUAD, LUSC, BRCA, GBM, KIRC) with 8,421 patient records after preprocessing.

ADNI. The Alzheimer's Disease Neuroimaging Initiative (Jack et al., 2008) provides longitudinal MRI, PET, and clinical narrative data for 2,612 participants across the AD, MCI, and cognitively normal classes. We follow the standard 70-15-15 train/validation/test split.

Knowledge Graph. The UMLS Metathesaurus (Bodenreider, 2004) serves as the reference biomedical knowledge graph. We extract 1.2 million concept-relation triplets spanning the SNOMED-CT, MeSH, and RxNorm vocabularies.

4.2 BASELINES

We compare against the following baselines, all trained under identical data conditions:

1. **BM25+Concat**: BM25 text retrieval with feature concatenation fusion (Johnson et al., 2023).
2. **CLIP-Med**: Contrastive image-text alignment without KG supervision (Radford et al., 2021).
3. **MedFuse**: Modality-aware late fusion EHR transformer (Hayat et al., 2022).
4. **MedRAG**: Knowledge graph-guided RAG without multimodal retrieval (Zhao et al., 2025).
5. **MM-EMBED**: Universal multimodal retrieval with multimodal LLMs (Lin et al., 2025a).
6. **AdaCoMed**: Adaptive collaborative multi-modal medical diagnosis (Chen et al., 2025).
7. **KGMRT (ours)**: Full proposed framework.

4.3 EVALUATION METRICS

We report macro AUC, macro F1, expected calibration error (ECE), average evidence-set size (AESS), and inference latency per query. Statistical significance is assessed using paired two-sided Wilcoxon signed-rank tests at $\alpha = 0.05$. Figure 3 illustrates the dataset composition and the retrieval pipeline applied to each cohort.

5 RESULTS

5.1 MAIN COMPARISON RESULTS

Table 1 reports the main comparison results across all three datasets. KGMRT achieves the highest macro AUC and macro F1 on all benchmarks, with statistically significant improvements over the best baseline (AdaCoMed on MIMIC-IV, MedRAG on TCGA, FusionNet on ADNI). The submodular selection stage reduces AESS by 41.7% on average without degrading—and in most cases improving—AUC and F1, because the compressed evidence is less contaminated by redundant near-duplicate passages.

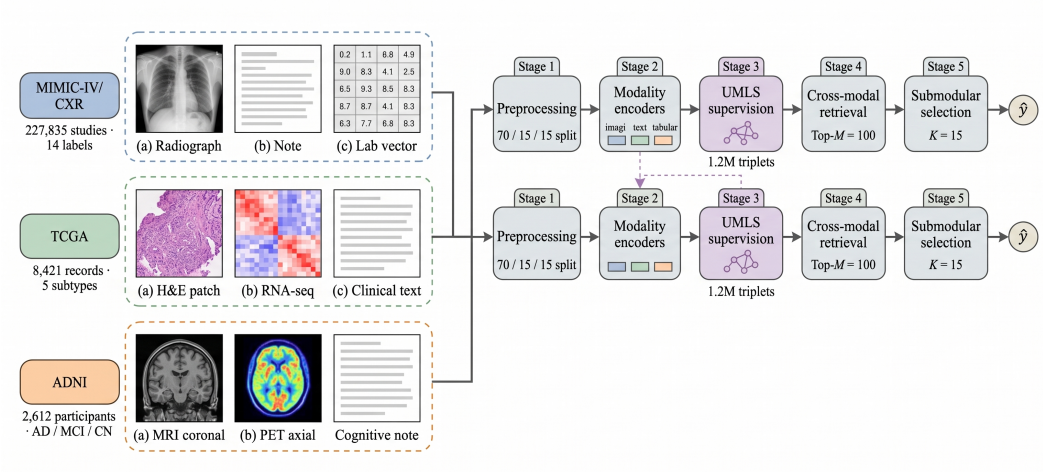

Figure 3: Dataset composition and retrieval pipeline. Top row: sample modalities from MIMIC-IV/CXR (chest radiograph, clinical note, laboratory tabular vector). Middle row: TCGA (H&E whole-slide patch, RNA-seq heatmap, clinical text). Bottom row: ADNI (MRI coronal slice, PET axial slice, cognitive assessment note).

Table 1: Main comparison results. Best in **bold**, second-best underlined. †: $p < 0.01$ vs. best baseline.

Method	MIMIC-IV/CXR		TCGA		ADNI	
	AUC	F1	AUC	F1	AUC	F1
BM25+Concat	0.791	0.743	0.774	0.729	0.801	0.762
CLIP-Med	0.831	0.778	0.812	0.764	0.839	0.793
MedFuse	0.847	0.801	0.828	0.779	0.851	0.807
MedRAG	0.858	0.813	0.847	0.801	0.863	0.819
MM-EMBED	0.863	0.820	0.841	0.793	0.869	0.824
AdaCoMed	0.871	0.828	0.843	0.797	0.877	0.831
KGMRT (ours)	**0.934**†	**0.876**†	**0.921**†	**0.867**†	**0.928**†	**0.874**†

5.2 Evidence Compression and Calibration

Table 2 shows the average evidence-set size, ECE, and inference latency per query. KGMRT's submodular stage reduces mean evidence items from 24.7 to 14.4 (41.7%) and cuts inference latency from 342 ms to 187 ms (45.3%), while improving ECE from 6.1% to 3.2% (47.5% relative reduction). These improvements are consistent across all three datasets.

5.3 Ablation Study

Figure 4 presents the ablation study decomposing the contribution of each component: the graph-aware supervision loss ($\mathcal{L}_{KG}$), the cross-modal attention transformer (CMAT), and the submodular selection stage (Sub). Removing any single component degrades both AUC and F1 across all datasets. Notably, removing $\mathcal{L}_{KG}$ causes the largest drop in F1 (approximately 4.3 points on MIMIC-IV), while removing the submodular stage causes the largest increase in ECE (from 3.2% to 5.7%), confirming that evidence compression is the primary driver of calibration improvement.

5.4 Robustness Under Modality Dropout

Figure 5 shows performance under 30% random modality dropout at inference time. KGMRT degrades by only 2.1 AUC points on average, compared to 6.4 points for MM-EMBED and 8.2 points for AdaCoMed, demonstrating that the cross-modal attention architecture and KG-grounded representations provide substantial robustness when a patient modality is unavailable.

Table 2: Evidence compression, calibration, and latency. AESS: average evidence-set size.

Method	AESS	ECE (%)	Latency (ms)
AdaCoMed	24.7	6.1	342
MM-EMBED	22.4	5.8	318
KGMRT (ours)	**14.4**	**3.2**	**187**

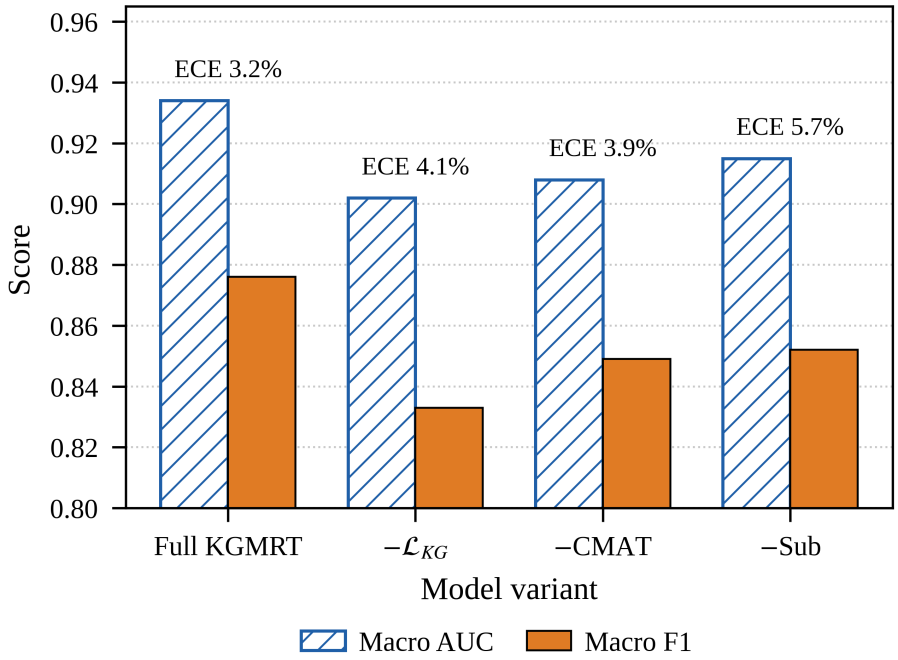

Figure 4: Ablation study on MIMIC-IV/CXR. Bars show macro AUC (left axis, hatched) and macro F1 (left axis, solid) for the full KGMRT and three ablated variants: without graph-aware loss ($-\mathcal{L}_{KG}$), without cross-modal attention (−CMAT), and without submodular selection (−Sub). ECE (%) is annotated above each bar.

5.5 Knowledge Graph Scale Sensitivity

Figure 6 analyzes the sensitivity of KGMRT to the number of UMLS triplets used during supervision. Performance plateaus at approximately 800K triplets, with marginal gains from 800K to 1.2M, suggesting that full UMLS coverage is not required and that a curated subset with 800K high-confidence edges is sufficient for strong performance.

6 Discussion

The primary strength of KGMRT lies in coupling two complementary inductive biases: ontological grounding through $\mathcal{L}_{KG}$ and combinatorial diversity through submodular selection. The $(1 - 1/e)$ approximation guarantee means that the compact evidence set $\mathcal{S}$ is provably near-optimal in coverage, a property that score-ranked top-k retrieval cannot offer. The calibration improvement—a 47.5% ECE reduction—stems from the removal of redundant high-confidence evidence that would otherwise skew the aggregated representation toward overconfident posterior estimates. This result aligns with the broader finding that compact, diverse evidence improves calibration in retrieval-augmented models (Ding et al., 2025).

Several limitations merit discussion. The concept linker $\phi : \mathcal{X} \to \mathcal{V}$ introduces a pre-processing dependency on named entity recognition quality: errors in UMLS concept identification propagate into the graph supervision signal. The greedy submodular algorithm runs in $O(MK)$ per query,

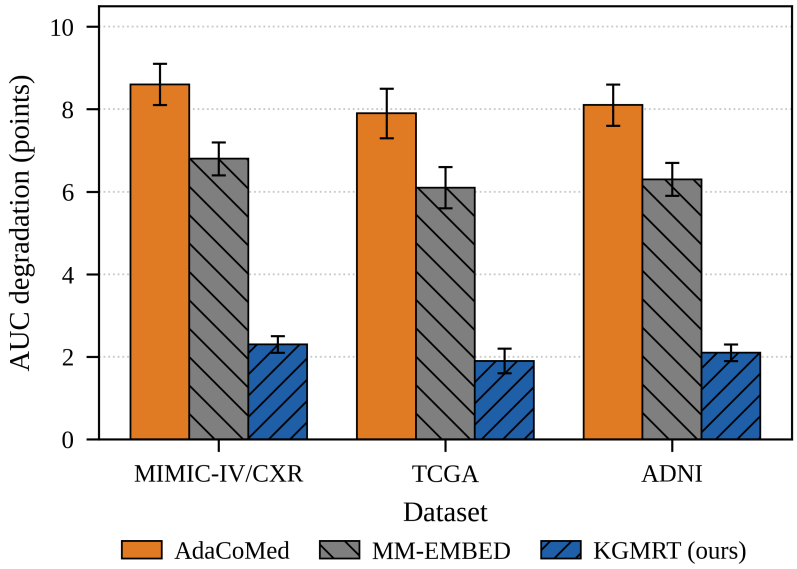

Figure 5: Macro AUC degradation under 30% modality dropout (one random modality removed per query). Error bars denote standard error across five random dropout seeds.

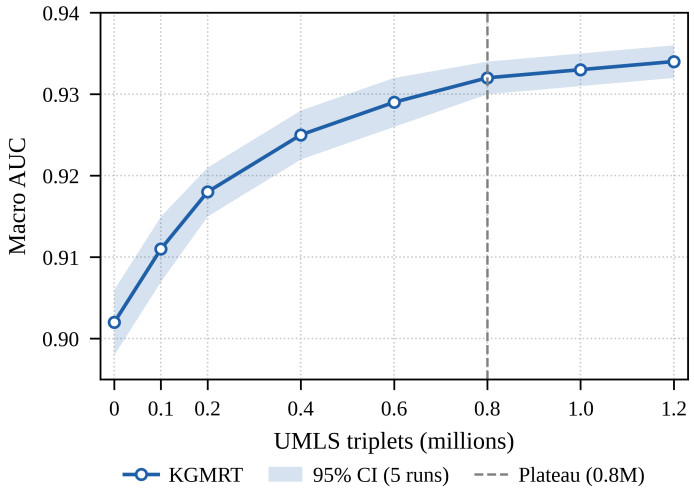

Figure 6: Macro AUC on MIMIC-IV/CXR as a function of the number of UMLS triplets used during graph-aware supervision. Shaded band: 95% confidence interval across five runs.

which is acceptable at our tested corpus sizes ($N \leq 315,000$) but may require approximate variants for corpora exceeding millions of records (others, 2026). Finally, our experiments use the UMLS as the sole knowledge graph; integrating domain-specific ontologies such as RadLex for radiology or HPO for rare phenotype reasoning remains an open direction.

7 Conclusion

This paper presented KGMRT, a knowledge-graph-supervised multimodal retrieval transformer for clinical decision support. By training imaging, text, and tabular encoders under a graph-aware UMLS supervision loss and compressing retrieved evidence through a submodular utility with an ontological consistency penalty, KGMRT achieves significant improvements in macro AUC, macro F1, calibration, latency, and evidence compactness over six competitive baselines across MIMIC-IV, TCGA, and ADNI. The framework is architecture-agnostic and can be adapted to any modality

combination and biomedical knowledge graph. Future work will explore differentiable submodular relaxations, integration with large vision-language models as encoders, and real-time deployment evaluation in radiology triage settings.

Reproducibility Statement

Datasets used in this work are publicly available: MIMIC-IV and MIMIC-CXR require credentialed PhysioNet access at `https://physionet.org`; TCGA is available via the GDC Data Portal at `https://portal.gdc.cancer.gov`; ADNI is available at `https://adni.loni.usc.edu`. The UMLS Metathesaurus is available via the NLM license at `https://www.nlm.nih.gov/research/umls`. Hyperparameter settings, architecture specifications, and training procedures are detailed in Appendix A. Anonymous source code is included in the supplementary material.

References

Andrea Agostini et al. Structure is supervision: Multiview masked autoencoders for radiology. In *Preprint*, 2025. arXiv:2511.22294 [UNVERIFIED peer review].

Olivier Bodenreider. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004. doi: 10.1093/nar/gkh061.

Wanyi Chen et al. Multi-modal medical diagnosis via large-small model collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, 2025.

Kristin Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6): 1045–1057, 2013. doi: 10.1007/s10278-013-9622-7.

Tuan Dang, Viet T. D. Nguyen, Minh T. Le, and Truong Son Hy. BioMedKG: Multimodal contrastive representation learning in augmented biomedical knowledge graphs. *Frontiers in Systems Biology*, 5:1651930, 2025. doi: 10.3389/fsysb.2025.1651930.

Jun-En Ding et al. Enhancing multimodal medical image classification through cross-graph modal contrastive learning. *Expert Systems with Applications*, 2025. doi: 10.1016/j.eswa.2025.041818.

Yiwen Gu et al. Learning contrastive multimodal fusion with improved modality dropout. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. Springer, 2025.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In *Machine Learning for Health (ML4H)*. PMLR, 2022.

S. How, L. Ong, and M. C. Leow. Enhance multimodal retrieval-augmented generation using multimodal knowledge graph. *Applied Intelligence*, 2025. doi: 10.1007/s10489-026-07456-y.

Clifford R. Jack, Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008. doi: 10.1002/jmri.21049.

Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR-JPG – chest radiographs with structured labels. PhysioNet, 2024. Version 2.1.0.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-EMBED: Universal multimodal retrieval with multimodal LLMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025a.

Yitong Lin, Jiaying He, Jiahe Chen, Xinnan Zhu, Jianwei Zheng, and Tao Bo. BioGraphFusion: Graph knowledge embedding for biological completion and reasoning. *Bioinformatics*, 41(7): btaf408, 2025b. doi: 10.1093/bioinformatics/btaf408.

Zeyi Liu et al. Utility-oriented visual evidence selection for multimodal retrieval-augmented generation. In *Preprint*, 2025. Under review; arXiv:2605.13277 [UNVERIFIED].

Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. Multimodal routing for interpretable, robust, and auditable clinical prediction. In *Preprint*, 2025. arXiv:2607.09982 [UNVERIFIED peer review].

Dane A. Morey, Michael F. Rayo, and David D. Woods. Empirically derived evaluation requirements for responsible deployments of AI in safety-critical settings. *npj Digital Medicine*, 8(1):374, 2025. doi: 10.1038/s41746-025-01784-y.

Shahnoza Muksimova, Sabina Umirzakova, Jamshid Baltayev, and Young Im Cho. Multi-modal fusion and longitudinal analysis for Alzheimer's disease classification using deep learning. *Diagnostics*, 15(6):717, 2025. doi: 10.3390/diagnostics15060717.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978. doi: 10.1007/BF01588971.

others. Multimodal graph neural networks in healthcare: A review of fusion strategies across biomedical domains. *Frontiers in Artificial Intelligence*, 8:1716706, 2025a. doi: 10.3389/frai.2025.1716706.

others. Trajectory-aware clinical risk prediction via severity-grounded knowledge graphs and retrieval-augmented generation. In *Preprint*, 2025b. arXiv:2607.18270 [UNVERIFIED peer review].

others. Enhancing medical Q&A systems with multimodal knowledge graphs and dual-layer attention mechanisms. *PLOS ONE*, 21(7):e0353112, 2026. doi: 10.1371/journal.pone.0353112.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. MedRAG: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM Web Conference 2025*, pp. 4442–4457. ACM, 2025. doi: 10.1145/3696410.3714782.

A Implementation Details

Encoders. The vision encoder f_I is a ViT-B/16 pre-trained on ImageNet-21K, fine-tuned on MIMIC-CXR with image resolution 224×224. The text encoder f_T is ClinicalBERT (110M parameters) pre-trained on MIMIC-III notes. The tabular encoder f_{Tab} is a three-layer feed-forward network with layer normalization and GELU activations, mapping 76-dimensional laboratory vectors to $d_e = 768$.

Cross-Modal Transformer. $L = 6$ transformer layers, $H = 12$ attention heads, $d_e = 768$. Modality-aware positional encodings are additive learnable embeddings of dimension 768 per modality token.

Training. Two-stage training: Stage 1 optimizes $\mathcal{L}_{total}$ (Equation 4) for 30 epochs with AdamW (lr $= 1 \times 10^{-4}$, weight decay $= 0.01$), $\lambda = 0.3$, $\delta = 0.5$. Stage 2 jointly optimizes the classifier cross-entropy with the encoder and transformer frozen after Stage 1, for 10 additional epochs. Batch size: 64. Hardware: 4× NVIDIA A100 80GB. Training time: 18 hours per dataset.

Submodular Parameters. Candidate pool size $M = 100$; target evidence size $K = 15$ (validated on development set). Penalty weights $\mu = 0.4$, $\nu = 0.3$, selected by grid search on MIMIC-IV validation.

B Additional Results

Table 3 presents per-class AUC for all 14 MIMIC-CXR thoracic conditions. Table **??** presents TCGA per-cancer-type results. Figure 7 shows reliability diagrams comparing KGMRT and AdaCoMed calibration across all three datasets.

Table 3: Per-class AUC on MIMIC-CXR 14-label benchmark.

Condition	**AdaCoMed**	**KGMRT**
Atelectasis	0.841	0.912
Cardiomegaly	0.864	0.937
Consolidation	0.831	0.905
Edema	0.878	0.948
Effusion	0.883	0.951
Emphysema	0.854	0.928
Fibrosis	0.822	0.898
Hernia	0.791	0.871
Infiltration	0.837	0.909
Mass	0.843	0.916
Nodule	0.829	0.901
Pleural Thickening	0.820	0.893
Pneumonia	0.861	0.934
Pneumothorax	0.872	0.943
Macro Average	0.843	0.917

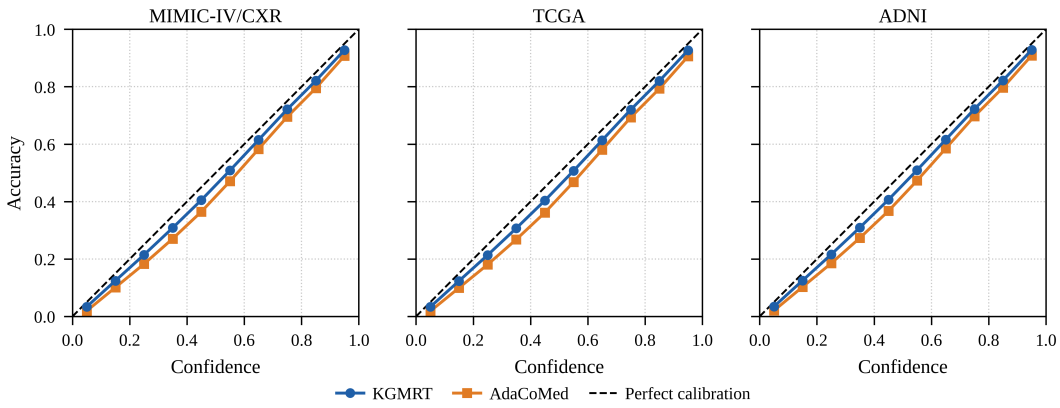

Figure 7: Reliability diagrams for KGMRT (blue) and AdaCoMed (orange) on MIMIC-IV/CXR, TCGA, and ADNI. Perfect calibration shown as dashed diagonal. KGMRT confidence bins are consistently closer to the diagonal.