ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Medical multimodal entity linking under modality missingness

Shiji Zang[1], Chenyang Bu[1]*, Yunpeng Hong[1], He Ren[2], and Weiping Ding[3]

[1]Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China), School of Computer Science and Engineering, Hefei University of Technology, Hefei, Anhui, China

[2]Department of Data Science and Big Data Technology, Faculty of Medical Instrumentation, Shanghai University of Medicine & Health Sciences, Shanghai, China

[3]Department of Computer Science, School of Artificial Intelligence and Computer Science, Nantong University, Nantong, Jiangsu, China

Abstract

Multimodal entity linking aligns real-world multimodal mentions with corresponding entities in a knowledge base and is increasingly important for applications such as medical knowledge grounding and retrieval-augmented generation. However, existing methods struggle in real-world medical settings, where missing modalities are common due to privacy restrictions, acquisition failures, and legacy data, leading to frequent missing visual information in mentions. Furthermore, medical knowledge bases typically suffer from sparse visual coverage and long-tail entities—a modality asymmetry that weakens multimodal matching and reduces entity disambiguation performance. To address this, we introduce a robust framework for medical multimodal entity linking under incomplete modality conditions, which leverages conditional generative models to reconstruct missing visual information with high semantic consistency. The framework incorporates a dynamic loss balancing strategy to adaptively coordinate modality-specific and cross-modal learning during training. Experiments on a medical multimodal entity linking benchmark covering 11 medical domains demonstrate that the proposed method achieves consistent performance gains and more stable training under modality-missing conditions.

Keywords: Knowledge base; Multimodal linking; Entity linking; Robust framework; Dynamic loss balancing

***Corresponding author:**
Chenyang Bu
(chenyangbu@hfut.edu.cn)

Citation: Zang S, Bu C, Hong Y, Ren H, Ding W. Medical multimodal entity linking under modality missingness. *Artif Intell Health.* 2026;3(4):026070015. doi: 10.36922/AIH026070015

Received: February 11, 2026

Revised: March 14, 2026

Accepted: March 30, 2026

Published online: May 19, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

1.1. Definitions and background

Multimodal entity linking aims to ground real-world mentions, expressed through textual and visual evidence, to canonical entities in a knowledge base (KB). In the medical domain, this task is important be cause the same clinical concept may appear across clinical notes, medical images, and patient-facing documents in different forms and varying details. Accurate grounding supports medical information retrieval, facilitates downstream applications such as clinical decision support and retrieval-augmented generation, and improves the interpretability and traceability of model outputs.[1-3]

In practice, however, medical multimodal entity linking is often affected by incomplete multimodal evidence. On the mention side, visual information may be unavailable because of privacy restrictions, acquisition failures, historical data limitations, or data-sharing constraints.[4] On the entity side, medical KBs often provide sparse and uneven visual coverage, and many entities lack associated images. As a result, the modalities available for mentions are often misaligned with those represented in the KB, making reliable disambiguation difficult.[5]

This issue is especially evident when textual expressions are broad, but clinically meaningful distinctions depend on fine-grained cues. For example, a mention such as "pigmented lesion" may correspond to entities with different medical implications. When an image is available, visual morphology can help link the mention specifically to melanoma. Without such evidence, the mention may instead be mapped to a broader entity, such as skin neoplasm or nevus. Figure 1 illustrates this problem and shows that missing visual evidence is a practical source of error in medical entity linking.

1.2. Literature review

Existing research on multimodal entity linking mainly follows two lines. One is based on embedding-oriented retrieval pipelines, which emphasize scalable candidate matching in large KBs.[6-8] The other relies on multimodal fusion or vision–language interaction models, which improve alignment by jointly modeling textual and visual information.[9-12] These approaches have achieved strong results in standard settings.

However, most existing methods implicitly assume that relevant modalities are available and informative during both training and inference. While this assumption is often acceptable in benchmark settings, it is difficult to maintain in real-world medical scenarios. Once visual evidence is absent on either the mention side or the entity side, the matching process becomes asymmetric, and methods designed for complete inputs may lose effectiveness.

A related limitation lies in optimization. Many existing methods combine textual, visual, and cross-modal alignment losses using fixed weighting schemes.[10-12] Such designs assume that the contribution of each modality remains relatively stable during training. In medical settings, however, one missing modality may cause the remaining modality to dominate optimization, while cross-modal supervision becomes weak or noisy. This can lead to biased representations and reduced robustness.

Although recent studies have improved multimodal matching performance, limited attention has been paid to medical multimodal entity linking under missing modality conditions. In particular, prior work has rarely considered how to compensate for missing modality evidence while maintaining stable optimization under imbalanced modality availability.

1.3. Purpose, motivation, research gap, and contributions

This study is motivated by the mismatch between existing multimodal entity linking methods and real-world medical data. In practice, complete multimodal evidence is often unavailable, and sparse visual coverage on the entity side further complicates grounding. In medical applications, an incorrect association between a malignant condition and a benign or overly broad entity may introduce clinically meaningful downstream risks. Robustness to missing modalities is therefore a practical requirement.

To address this gap, we propose generative incomplete-modality multimodal entity linking (GIMEL), a framework for medical multimodal entity linking under missing modality conditions. The framework is designed from two complementary perspectives. First, missing modality evidence is reconstructed using conditional generative modeling to recover part of the fine-grained information lost due to incomplete inputs. Second, modality-specific and cross-modal objectives are dynamically balanced during training to reduce the bias caused by asymmetric modality availability.

Through this design, GIMEL aims to improve medical entity linking when visual evidence is incomplete, sparse, or unreliable. Rather than assuming ideal multimodal conditions, it explicitly accounts for the practical asymmetry between mentions and KB entities. The main contributions of this work are summarized as follows:

(i) We investigate medical multimodal entity linking under missing modality conditions and highlight its practical importance for reliable medical grounding.
(ii) We propose GIMEL, a unified framework for medical multimodal entity linking under modality missingness, which integrates conditional generative completion to recover missing visual semantics and dynamic loss balancing to alleviate optimization bias caused by asymmetric modality availability.
(iii) We conduct extensive experiments on a domain-specific benchmark for medical multimodal entity linking and show that the proposed framework yields stronger robustness and more stable convergence across diverse medical domains.

Taken together, medical multimodal entity linking differs from standard multimodal grounding in that its challenge arises not only from semantic complexity, but also from the structural incompleteness of available

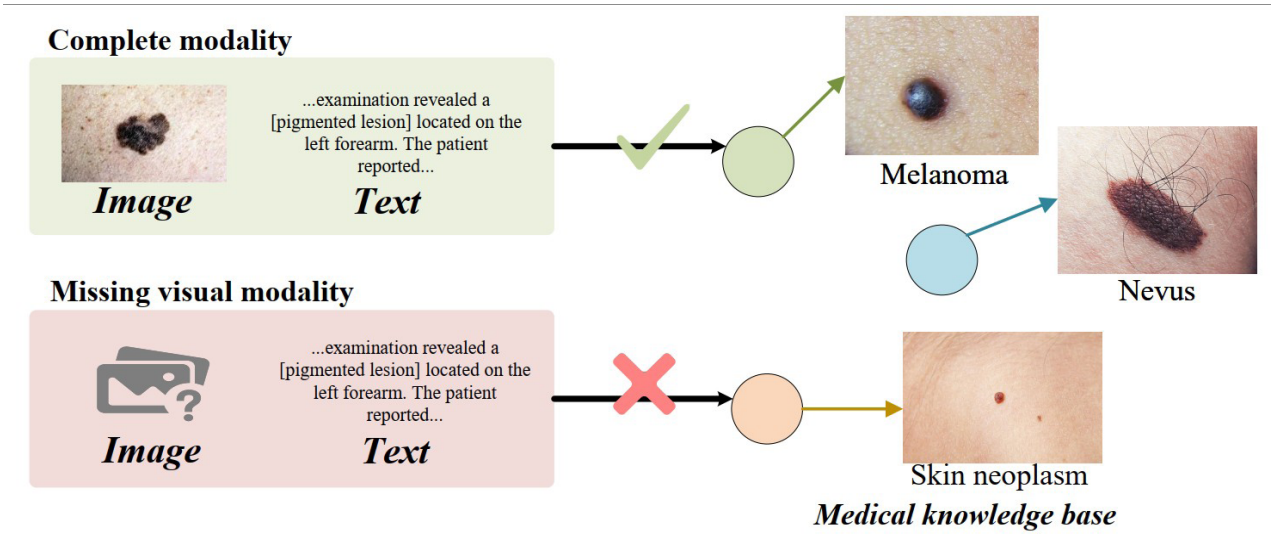


Figure 1. Illustration of medical entity linking errors caused by missing visual evidence. When visual cues are available, a pigmented lesion mention is grounded to melanoma. Without visual evidence, the same mention is often linked to coarse-grained or benign entities such as skin neoplasms or nevus.

evidence. This motivates the following review of related work and clarifies why robustness to missing modalities should be treated as a central design objective.

2. Related work

2.1. Multimodal medical retrieval and entity linking

Medical information systems increasingly rely on accurately grounding textual or multimodal mentions to relevant biomedical entities, documents, or visual evidence within large-scale medical knowledge repositories. This problem is closely related to traditional entity linking, which aims to associate entity mentions in natural language with corresponding entities in a KB.[13,14] Although single-modal entity linking methods are highly effective for purely textual data, they tend to falter when applied to medical contexts. This is largely due to challenges such as domain-specific terminology, semantic ambiguity, and the inherent sparsity of information in clinical texts.

To tackle these issues, recent work has explored multimodal medical retrieval[15,16] alongside entity linking methods[17]—combining cues from clinical text and medical images to distinguish ambiguous mentions and retrieve clinically meaningful entities or documents. These methods aim to align multimodal queries—consisting of textual descriptions and optional visual context—with target entities or passages in multimodal medical corpora. Existing approaches in this line of work can be broadly categorized into embedding-based retrieval methods and retrieval methods enhanced with stronger semantic modeling.

Recent studies have also explored stronger semantic modeling techniques as auxiliary components in medical retrieval pipelines, aiming to improve candidate discrimination after the initial retrieval stage.[18] This trend is also consistent with the broader use of large language models in healthcare, where stronger semantic understanding is increasingly adopted to support complex medical intelligence tasks.[19]

2.1.1. Embedding-based multimodal retrieval

The foundation of most multimodal medical retrieval systems lies in embedding-based methods, which project inputs from different modalities into a common vector space to enable similarity-based matching.[6–9] A widely adopted strategy is to first encode text and images with separate, modality-specific encoders, and then merge these representations into a joint embedding. For example, dual-branch architectures follow this pattern by keeping textual and visual streams independent until the fusion stage.[20] More recent approaches, however, have introduced graph neural networks to fuse multimodal features in a more flexible way, capturing fine-grained cross-modal relationships and structural dependencies.[21]

Contrastive learning has emerged as a dominant paradigm for training multimodal encoders, as it effectively enforces alignment between paired text and image representations while separating semantically unrelated samples. For example, contrastive objectives are employed to pre-train multimodal encoders such that textual descriptions and corresponding medical images are embedded closely in a shared latent space.[22] These techniques substantially improve retrieval precision

and scalability in large medical corpora. While current embedding-based methods are largely designed to maximize accuracy in settings where all modalities are available, they rarely account for the imperfect nature of real-world medical data. Factors such as missing modalities or imbalanced multimodal inputs are often neglected, despite their potential to severely compromise retrieval effectiveness.[23]

2.1.2. Retrieval methods with stronger semantic modeling

When semantic modeling is prioritized, retrieval systems typically begin with a broad candidate generation step before applying deeper contextual reasoning to refine the rankings.[1,2,24,25] In the medical domain, this refinement can take the form of entity descriptions, clinical context, or structured biomedical knowledge, all of which help the model grasp finer-grained semantic relationships. The result is often a meaningful improvement in ranking quality, especially when shallow similarity measures fall short of resolving entity ambiguity.

However, their effectiveness still depends heavily on the quality of the initial candidate set produced by the first-stage retriever. This limitation is particularly evident in medical settings, where the candidate space is large, heterogeneous, and frequently affected by incomplete multimodal evidence.[26] As a result, robust embedding-based multimodal retrieval remains the foundation of practical medical retrieval and entity linking systems. Under modality-missing scenarios, incomplete multimodal representations can reduce recall at the retrieval stage, and existing approaches still lack principled mechanisms to address this issue directly.

2.2. Multimodal learning under missing modalities

Missing or partially observed modalities constitute a critical challenge for multimodal medical retrieval systems.[27,28] In practice, a substantial portion of medical data lacks one or more modalities due to incomplete clinical records, privacy restrictions, or modality-specific acquisition constraints. A widely adopted baseline strategy is to represent missing modalities using zero vectors or placeholder embeddings.[29] However, such heuristics fail to exploit the semantic information available in observed modalities and often introduce noise, thereby limiting both retrieval accuracy and robustness.

To better address this issue, we reviewed representative techniques from the broader multimodal learning literature, focusing on two main categories: joint learning methods and cross-modal generation methods.

2.2.1. Joint learning methods

Joint learning methods sidestep the reconstruction problem altogether by learning representations that remain robust representations that hold up even when some modalities are missing. Many such approaches are built on multitask or multi-branch frameworks that jointly expose the model to complete and incomplete input combinations throughout training. The interplay between shared and modality-specific encoders is central to this design: it gives the model something to fall back on when particular modalities are absent, and has been shown to reduce performance drops in incomplete-modality scenarios.[30,31]

Knowledge distillation has similarly been applied to strengthen robustness in settings where inter-modality correlations are pronounced. A teacher model with access to complete multimodal data is used to supervise a student model restricted to incomplete observations, enabling the student to better compensate for missing information at test time.[32] Adaptive fusion mechanisms offer yet another perspective, modulating the extent to which each modality contributes to the final representation based on its availability or quality.[33] Multi-level architectures that blend early and late fusion strategies have also been proposed to strike a balance between cross-modal interaction and resilience to modality absence.[34]

That said, in the specific context of multimodal medical retrieval and entity linking, a large portion of existing work still relies on zero-vector masking to handle absent modalities, counting on joint training to recover any information that is lost.[29] This remains a problematic assumption—the strategy tends to yield unstable and ambiguous representations, particularly when diagnostically important modalities are missing or when the distribution across modalities is heavily skewed.

2.2.2. Cross-modal generation methods

Cross-modal generation methods take a more direct approach by explicitly reconstructing missing data, using correlations among available modalities as a guide—for instance, inferring visual representations from text or the reverse.[35] By learning these cross-modal dependencies, the model can estimate the latent content of absent modalities and, in doing so, produce more complete multimodal representations.

Existing work in this area can be broadly divided into data-level and feature-level generation approaches. One category of cross-modal generation operates at the data level, directly synthesizing raw modality inputs such as medical images from text descriptions.[36] The reconstructions produced can be strikingly realistic, but the approach is computationally demanding and does not

scale well—a serious constraint when working with high-resolution imagery or large-scale medical datasets. Feature-level generation sidesteps this bottleneck by reconstructing latent embeddings instead of raw inputs.[37] Conditional generative models are typically used here to estimate informative feature representations for absent modalities based on observed ones.[38] Generative adversarial network (GAN)-based methods, for instance, have been adapted for this purpose using contrastive objectives to align generated and observed representations within a shared latent space, boosting robustness in high-dimensional multimodal tasks.[39] The result is an approach that is both more computationally tractable and easier to slot into existing retrieval workflows.

Cross-modal generation methods are particularly suitable for medical retrieval scenarios, where reconstructing full-resolution data may be unnecessary or impractical. Unlike joint learning approaches, which implicitly rely on shared optimization objectives, generation-based methods provide a direct mechanism to compensate for missing modality semantics. In this work, we adopted a feature-level generation strategy to address missing modality, balancing representational completeness and computational efficiency through lightweight generative architectures.

Overall, the existing literature suggests that current methods address only part of the problem considered in this study. Embedding-based and vision–language retrieval methods offer efficient large-scale matching, but their robustness often declines once modality completeness is no longer guaranteed. Large language model-based reranking methods provide stronger reasoning ability, yet they still depend on the recall quality of an upstream retriever and do not fundamentally resolve missing-modality issues. Meanwhile, missing-modality learning methods either emphasize robust joint optimization without explicitly restoring missing semantics or reconstruct missing information without sufficiently accounting for the imbalance in retrieval-oriented optimization. These observations indicate that medical multimodal entity linking under modality missingness requires a more integrated solution that jointly addresses semantic completion and optimization robustness.

3. Methods

The overall framework consists of two stages: a pre-training stage for the generative model and a joint multimodal training stage. The generative model was pre-trained independently as a plug-in module to generate pseudo-label vectors for the image modality based on textual information. In the joint multimodal training stage, the pre-trained generative module was used to supplement missing image modality data, while a dynamic loss balancing mechanism was employed to adjust the update rates of different modalities, thereby mitigating modality imbalance caused by missing data.

3.1. Problem formulation

The multimodal medical knowledge repository is presented by $\xi = \{E_i\}_{i=1}^{N}$, where each entry E_i corresponds to a biomedical unit (e.g., a paragraph-level passage or a medical case description) represented by a tuple (e_{t_i}, e_{v_i}) consisting of its textual information e_{t_i} and visual context e_{v_i} (e.g., a medical image associated with the passage).

Some entries, E_i, may have missing modalities, for instance, lacking an associated image e_{v_i} due to privacy restrictions or incomplete documentation. A query (or mention) X_j is represented as (X_{t_j}, X_{v_j}), where X_{t_j} denotes the textual component of the query and X_{v_j} is its optional associated visual component. Similarly, queries can also be incomplete, potentially missing the visual component X_{v_j}. The objective of the task is to retrieve and ground a given query X_j to the most relevant entry E_i within the repository ξ. This retrieval must be robust even when faced with missing modalities in either the queries or the repository entries, ensuring the model can accurately match biomedical semantics despite these imperfections.

In this work, we focused on improving the robustness of the embedding-based retrieval stage, while allowing optional inference-time refinement without modifying the learned representations.

3.2. Overview of generative incomplete-modality multimodal entity linking

Our proposed framework was designed around two core objectives aimed at enhancing model robustness: (i) objective 1 (O1): to develop a plug-and-play modality-missing completion module to alleviate the loss of semantic evidence caused by missing modalities, and (ii) objective 2 (O2): to design a modality loss adjustment mechanism to mitigate the modality imbalance exacerbated by missing modalities.

As illustrated in Figure 2, the framework consists of two key stages: (i) generative model pre-training, and (ii) multimodal training with dynamic loss balancing. The overall training process is shown in Algorithm 1, and the generative model pre-training process is described in Algorithm 2.

In this work, we kept the overall framework unchanged and only replaced the pre-trained multimodal encoder with a biomedical-domain variant. Specifically, $M_\phi(\cdot)$

was instantiated by Biomed contrastive language–image pre-training (BiomedCLIP), where the text encoder was initialized from PubMedBERT, and the vision encoder followed a ViT-B/16 architecture. This replacement improved domain alignment for medical terminology and medical visual concepts while preserving the same embedding-based retrieval pipeline.[40]

In Stage I, we focused primarily on O1 as shown in Lines 4 to 12 of Algorithm 1 and described in Section 3.3. We began by selecting samples from the training set that contained complete modality information. Our goal was to explore the ways to leverage the available modalities to complete the missing image modality by generating corresponding pseudo-label vectors (as shown in Algorithm 2). To this end, we assumed that an implicit mapping exists between the available modalities and the missing image modality. Based on this assumption, we employed lightweight generative models such as GANs and variational autoencoders (VAEs) to learn the mapping from the existing modalities to the pseudo-label vectors of the missing image modality. Moreover, emphasis was placed on how to effectively use the plug-and-play module to complete missing modalities. Specifically, during multimodal training, if the input data contained all modalities, it was used directly for training. However, when certain modalities were missing, the pre-trained generative models were invoked to synthesize the missing image tensors based on the available textual information. This design ensured the model could operate robustly even with incomplete multimodal data.

In the second stage, we focused on O2, primarily addressing the modality imbalance caused by missing modalities by designing a modality loss balancing method. Specifically, by introducing a dynamic loss balancing mechanism (as shown in Lines 26 to 31 of Algorithm 1 and Section 3.3), the contribution of each modality's loss to the total loss was dynamically adjusted based on the performance of different modalities during training, thereby optimizing the overall model performance. As shown in the bottom-right panel of Figure 2, compared to static loss weighting methods, this strategy significantly improved model performance, demonstrating its effectiveness.

In the following two subsections, we introduce the core modules of the framework: the pre-training of generative models for missing modality completion and the dynamic loss balancing strategy during multimodal data training, respectively. The rationale of this design is that missing-modality robustness in medical retrieval cannot be adequately addressed from a single perspective. If the model only relies on completion, the recovered representations may still be underutilized or overwhelmed during optimization. Conversely, if the model only adjusts loss contributions without restoring missing semantics, the training process remains constrained by incomplete evidence. By combining generative completion with dynamic loss balancing, the proposed framework aims to simultaneously improve representational completeness and training stability, particularly in medical settings characterized by KB-side visual sparsity.

3.3. Generative model pre-training

To obtain a more robust model, we designed a plug-and-play module to fill in missing image modalities. The training procedure is detailed in Algorithm 2. We assumed that an implicit mapping exists between the available modalities and the missing image modality. Based on this assumption, we first leveraged the complete modality information to establish a correspondence between the text and image modalities. When the image modality was missing, we used the text modality to generate pseudo-label vectors for the image modality, effectively filling in the missing information. Specifically, we adopted two generative models: VAE[41] and GAN[42]. These models were chosen for their ability to generate plausible representations of missing modalities while remaining computationally efficient compared to more complex alternatives, such as diffusion models.

In our setting, the goal of the generative module was not to perfectly reconstruct visually realistic images under arbitrary text–image mappings, but to preserve retrieval-oriented semantic consistency in the biomedical multimodal feature space. To this end, the generator was trained on paired complete-modality samples, so that text-conditioned generation was supervised by aligned text–image representations observed in the training data. As a result, the synthesized representation was encouraged to remain close to the target visual embedding that is useful for downstream retrieval, rather than to generate free-form visual content.

3.3.1. Variational autoencoders-based pre-training

We adapted the VAE framework proposed by Chen *et al.*[5] for the multimodal retrieval setting. In our implementation, the textual input T corresponds to the token ID sequence with a fixed maximum length of 256, and the visual input I corresponds to the preprocessed image tensor. To ensure a consistent input scale between training and inference, T can be optionally normalized by the tokenizer vocabulary size.

The VAE encodes T into a latent space, producing the mean μ and log variance $\log\delta^2$ of a Gaussian distribution

Algorithm 1. Proposed multimodal entity linking framework for modality missingness

Input: Training dataset D_train = { (X_j, E_true,j) }_{j = 1..|D_train|}; Knowledge base KB = { E_i }_{i = 1..N}; Pre-trained generative model G_theta (Algorithm 2); Multimodal encoders M_phi; Interaction unit S_psi.
Output: Trained models M_phi*, S_psi*, varphi_C*, varphi_Total*.
Initialize loss balancing weights varphi_C and varphi_Total.
For each training epoch do
For each sample (X_j, E_true,j) in D_train do
Step 1: Handle missing visual modality.
If x_v,j is missing in X_j then
Generate visual content: x_v_hat,j <- G_theta(x_t,j).
Set X'_j <- (x_t,j, x_v_hat,j).
Else
Set X'_j <- X_j.
End if.
Step 2: Encode mention.
Compute mention representation: m_j <- M_phi(X'_j).
Step 3: Compute candidate scores.
Initialize candidate score set: S_j <- empty.
For each entity E_k = (e_t,k, e_v,k) in KB do
If e_v,k is missing in E_k then
Generate visual content: e_v_hat,k <- G_theta(e_t,k).
Set E'_k <- (e_t,k, e_v_hat,k).
Else
Set E'_k <- E_k.
End if.
Encode entity: e_k <- M_phi(E'_k).
Compute similarity score: s_j,k <- S_psi(m_j, e_k).
Update candidate scores: S_j <- S_j ∪ { (E_k, s_j,k) }.
End for.
Step 4: Update parameters.
Compute losses L_T, L_V, L_C based on S_j and E_true,j.
Compute total loss L according to the loss function equation: $\mathcal{L}_X = -\log \frac{\exp(U_X(\mathbf{M}, \mathbf{E}))}{\sum_i \exp(U_X(\mathbf{M}, \mathbf{E}'_i))}, \quad X \in \{T, V, C\}.$
Update parameters phi, psi, varphi_C, varphi_Total via gradient descent.
End for.
End for.
Return (M_phi*, S_psi*, varphi_C*, varphi_Total*).

(**Equation 1**):

$$\mu, \log \delta^2 = \text{Encoder}(T) \tag{1}$$

We then sampled a latent variable z via the reparameterization trick (**Equation 2**):

$$z = \mu + \epsilon \odot \delta, \epsilon \sim N(0, I) \tag{2}$$

where $\delta = \exp(1/2 \log \delta^2)$ and $\odot$ denotes element-wise multiplication.

The decoder reconstructed the multimodal targets from the sampled latent variable. Specifically, we reshaped z to

Algorithm 2. Generative model pre-training

Input: Complete-modality dataset D_complete = { (T_i, I_i) }_{i = 1..N}; Method choice method in {VAE, GAN}.
Output: Pre-trained generative model G_theta.
If method = VAE then
Initialize VAE parameters theta_VAE (encoder + decoder; decoder is denoted as G_theta).
For each training epoch do
For each batch (T_batch, I_batch) in D_complete do
Extract text representations: t_batch <- Encode(T_batch).
Encode to latent space: (mu, sigma) <- Encoder(t_batch).
Sample latent variable: z ~ N(mu, sigma^2).
Reconstruct: I_hat_batch <- G_theta(z); and t_hat_batch <- Decoder(z).
Compute VAE loss L_VAE according to Equation 4.
Update theta_VAE via gradient descent.
End for.
End for.
Set G_theta <- decoder part of the trained VAE.
Else if method = GAN then
Initialize generator G_theta and discriminator D_omega.
For each training epoch, do
For each batch (T_batch, I_batch) in D_complete do
Extract text representations: t_batch <- Encode(T_batch).
Generate images: I_hat_batch <- G_theta(t_batch).
Compute discriminator loss L_D and update omega according to Equation 13.
Generate images (again): I_hat_batch <- G_theta(t_batch).
Compute generator loss L_G and update theta according to Equation 12.
End for.
End for.
Set G_theta <- trained generator.
End if.
Return G_theta.

obtain an image pseudo-label tensor and decoded z to reconstruct the textual representation. The reconstructed image pseudo-label tensor $\hat{I}$ and textual embedding $\hat{T}$ were generated as follows (**Equation 3**):

$$\begin{cases} \hat{I} = \psi(\text{reshape}(z)), \\ \hat{T} = \text{Decoder}(z) \end{cases} \tag{3}$$

where reshape$(\cdot)$ maps z into a tensor of shape $3 \times 224 \times 224$, and $\psi(\cdot)$ is an optional output stabilization function (e.g., $\psi(x) = s \cdot \tanh(x)$ with scale s).

In the optimization of the VAE, three loss functions were designed to ensure the model's effectiveness in generating and reconstructing multimodal data. The overall VAE loss was composed of the following three components (**Equation 4**):

$$L_{VAE} = w_T L_T + w_I L_I + \beta L_{KL} \tag{4}$$

where L_T represents the text alignment loss, L_I is the image reconstruction loss, and L_{KL} denotes the Kullback–Leibler (KL) divergence loss. w_T and w_I are the loss weights. β is the KL coefficient, which can be linearly warmed up during early epochs.

The text alignment loss ensures semantic consistency between the input and reconstructed text representations. This loss is computed by measuring the difference between

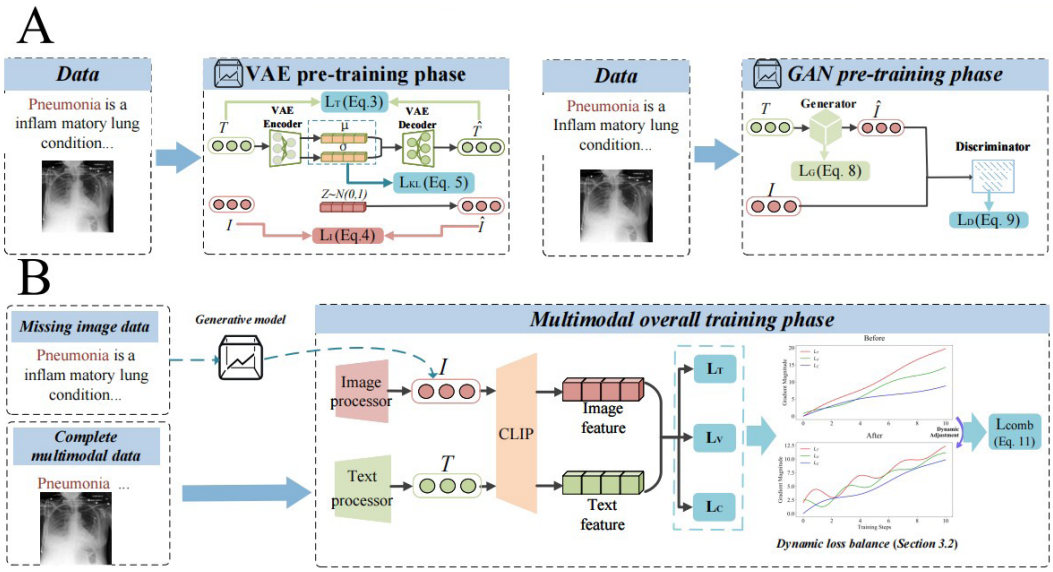


Figure 2. The overall framework consists of two stages: (A) a pre-training stage for the generative model and (B) a joint multimodal training stage. The generative model is pre-trained independently as a plug-in module to generate pseudo-label vectors for the image modality based on textual information. In the joint multimodal training stage, the pre-trained generative module is used to supplement missing image modality data, while a dynamic loss balancing mechanism adjusts the update rates of different modalities, thereby mitigating modality imbalance caused by missing data.
Abbreviations: CLIP: Contrastive language–image pre-training; GAN: Generative adversarial network; VAE: Variational autoencoders.

the input text vector T_i and the reconstructed text vector $\hat{T}_i$ using cosine similarity, as expressed by **Equation 5**:

$$L_T = \frac{1}{N}\sum_{i=1}^{N}\left(1-\cos(T_i,\hat{T}_i)\right) \quad (5)$$

where N is the number of samples, and T and $\hat{T}_i$ represent the i-th input and reconstructed text vectors, respectively.

The image reconstruction loss evaluates the quality of the reconstructed image tensors. This loss is calculated as the mean squared error between the real image tensor I_i and the reconstructed image tensor $\hat{I}_i$, with a small l_2 regularization term to stabilize the output magnitude (**Equation 6**):

$$L_I = \frac{1}{N}\sum_{i=1}^{N}\| I_i - \hat{I}_i \|_2^2 + \lambda \| I_i \|_2^2 \quad (6)$$

where λ is a small constant.

The KL divergence loss regularizes the latent space to prevent overfitting by encouraging the latent distribution to be close to a standard normal distribution. It is expressed as **Equation 7**:

$$L_{KL} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{d}\left(\exp(+\mu_{ij}^2 - 1 - \log\delta_{ij}^2)\log\delta_{ij}^2\right) \quad (7)$$

where d is the latent dimensionality, and μ_{ij} and $\log\delta_{ij}^2$ denote the mean and log variance for the j-th latent dimension of the i-th sample, respectively.

Together, these three objectives were designed to encourage semantic consistency between the text-conditioned latent representation and the paired visual target. Specifically, the text alignment term preserves the semantic content of the conditioning text, the image reconstruction term anchors the generated output to the paired visual feature space, and the KL regularization stabilizes the latent structure to avoid overfitting or arbitrary feature drift. Therefore, the VAE variant was optimized to recover task-relevant visual semantics for retrieval rather than visually realistic image content.

3.3.2. Generative adversarial network-based pre-training

In our GAN-based pre-training, the generator was conditioned on the token ID sequence T and aimed to synthesize a visual tensor that matches the paired real image representation I. Different from the original draft that assumes a probability form discriminator output, our implementation followed the standard logit-based adversarial training and performed adversarial learning directly in the image vector space, so that real and generated samples are compared under the same normalization. In our setting, we intentionally adopted

the standard logit-based binary cross-entropy objective rather than more sophisticated adversarial losses such as hinge loss or Wasserstein-style losses. The main reason is that the discriminator in our framework plays a relatively constrained role: it only needs to distinguish generated visual feature embeddings from real target embeddings in a biomedical feature space, rather than support unrestricted raw image generation. Moreover, adversarial learning was not the sole optimization signal in our GAN variant, since the generator was simultaneously regularized by a stronger reconstruction objective. In this retrieval-oriented feature-completion setting, standard binary cross-entropy (BCE) provided a simple and sufficiently stable formulation.

(a) Generator–discriminator interaction

Given T, the generator produces a raw image tensor in [-1,1] (**Equation 8**):

$$\tilde{I} = G(T) \tag{8}$$

To align with the real image preprocessing, we mapped $\tilde{I}$ into the CLIP-normalized image vector space via a deterministic transformation $\phi(\cdot)$ (i.e., $[-1,1] \rightarrow [0,1]$ followed by mean–standard normalization) (**Equation 9**):

$$\hat{I} = \phi(\tilde{I}) \tag{9}$$

so that $\hat{I}$ and I lie in the same space. The discriminator $D(\cdot)$ takes $X \in \{I, \hat{I}\}$ and outputs a logit validity score $D(X)$.

(b) Generator objective

We trained the generator with a reconstruction-dominant objective plus an adversarial term. Specifically, the reconstruction loss enforces fidelity in the image vector space (**Equation 10**):

$$L_{recon} = E\left[\left\|\hat{I} - I\right\|_2^2\right] \tag{10}$$

The adversarial loss encourages the generated sample to be judged as real by the discriminator under a logit-based binary cross-entropy (**Equation 11**):

$$L_{adv}^{G} = E\left[\text{BCEWithLogits}(D(\hat{I}), y_{real})\right] \tag{11}$$

where y_{real} can use label smoothing (e.g., 0.9) for stability. The final generator loss is shown as **Equation 12**:

$$L_G = \lambda_{recon} L_{recon} + \lambda_{adv} L_{adv}^{G} \tag{12}$$

In our setting, λ_{recon} was set larger than λ_{adv}, reflecting that pixel-level reconstruction is the primary objective, while adversarial learning improved realism. This design was intentional. In our task, the goal of the generator was not to produce visually realistic outputs for human inspection, but to recover feature representations that remain close to the paired biomedical target embedding and can serve as reliable substitutes in the downstream retrieval pipeline. If the adversarial term were weighted too heavily, training could favor distribution-level realism over sample-level semantic alignment, which would be less suitable for retrieval-oriented completion. Therefore, we treated adversarial learning as an auxiliary regularizer, while keeping reconstruction as the dominant optimization signal to improve both stability and retrieval utility.

(c) Discriminator objective

The discriminator was optimized to separate real and generated samples in the same normalized space. Using the same logit-based binary cross-entropy, the discriminator loss is shown as **Equation 13**:

$$L_D = \frac{1}{2}\left(\begin{array}{l} E\left[\text{BCEWithLogits}(D(I), y_{real})\right] \\ +E\left[\text{BCEWithLogits}(D(\hat{I}), y_{fake})\right] \end{array}\right) \tag{13}$$

where y_{fake}. When updating D, the generated $\hat{I}$ WAS detached to avoid gradient propagation into G. We noted that, although alternative adversarial objectives may offer stronger stability in some high-dimensional generation tasks, we used the standard logit-based BCE loss because the present discriminator operates in a constrained biomedical feature space and serves mainly as an auxiliary regularizer for retrieval-oriented completion. Since the generator was already anchored by a stronger reconstruction term, BCE was sufficient to provide stable adversarial supervision without introducing additional optimization complexity.

From a methodological perspective, the role of the generative module was not to produce visually realistic images for human inspection, but to recover task-relevant visual semantics in a representation space suitable for retrieval. This distinction is important in the present setting. Since the downstream objective is entity ranking rather than image generation itself, feature-level recovery is more aligned with the needs of medical multimodal retrieval than full data-level synthesis. It offers a practical trade-off between semantic restoration, computational efficiency, and integration with existing embedding-based pipelines.

In the GAN variant, this semantic consistency was enforced jointly by reconstruction toward the paired target embedding and adversarial regularization in the same normalized biomedical feature space. Therefore, even under highly non-linear text-image relationships, the generated feature was constrained to remain close to

the target visual–semantic manifold relevant to retrieval, rather than being optimized for unrestricted realism.

3.4. Dynamic loss balancing

In multimodal learning, balancing the losses of different modalities is a critical challenge during the optimization process[43]. When the semantic signal of a modality is missing, the optimization dynamics across modalities can become highly unbalanced, leading to modality imbalance. As a result, the model may overemphasize the loss associated with certain modalities during training, producing overly rapid gradient updates and further exacerbating the imbalance between modalities. Traditional approaches typically address this by assigning fixed weights (φ_i) to modality-specific losses (L_i), formulated as **Equation 14**, where L_i represents the loss of the i-th modality, and φ_i is the corresponding fixed weight factor.

$$L_{comb} = \sum_{i=1}^{n} L_i \varphi_i \tag{14}$$

However, this static weighting approach has significant limitations, such as difficulty in manual weight adjustment, inconsistent optimization rates across modalities, and the risk of trivial solutions or imbalance.

In our implementation, the dynamic balancing mechanism does not manually prescribe epoch-wise update rates for each modality. Instead, it introduces learnable scalar parameters for the selected loss terms and optimizes them jointly with the retrieval model through backpropagation. Concretely, for the two losses dynamically balanced in the main multimodal matching stage, namely the image–text interaction loss and the overall matching loss, the combined dynamic objective is defined as **Equation 15**:

$$L_{comb} = \sum_i \frac{1}{2\delta_i^2} L_i + \log(1+\delta^2) \tag{15}$$

where L_i denotes each task loss and δ_i is the corresponding learnable weighting parameter. Under this formulation, the effective contribution of each selected loss term was adjusted automatically according to the optimization dynamics learned from the data, rather than by a fixed heuristic schedule. In the current implementation, the image–text loss and the overall matching loss were dynamically weighted, while the text-only loss and image-only loss were retained as direct additive supervision terms.

In essence, dynamic loss balancing serves as an optimization-level complement to generative completion. While the completion module reduces the semantic gap caused by missing modalities, dynamic balancing regulates how strongly each modality influences parameter updates during training. This is especially valuable in asymmetric medical data, where a dominant modality can otherwise overshadow weaker but still clinically informative signals. Therefore, the balancing mechanism is expected not only to stabilize convergence but also to improve the effective use of recovered multimodal evidence.

3.5. Datasets and evaluation metrics

To evaluate the effectiveness of our framework on domain-specific medical data, we constructed a medical multimodal entity linking benchmark by processing multiple sub-corpora into a unified mention–entity ranking format. The resulting benchmark spanned 11 medical domains and exhibited substantial variation in both KB scale and sample distribution (Table 1). Overall statistics are summarized in Table 2.

Our raw data was derived from M3Retrieve, a large-scale multimodal medical retrieval benchmark designed for clinically grounded retrieval tasks.[44] In particular, we used the M3Retrieve_VisualContextRetrieval subset, which targets the setting of retrieving the most relevant medical passage given a query consisting of an image and its short textual context (e.g., a caption).[44] We further reorganized this retrieval-style data into a multimodal entity linking formulation, where each query was treated as a mention, and each retrievable passage was treated as a KB entity, enabling evaluation under the standard mention-to-entity ranking protocol.

For KB construction, we retained paragraph-level entries and performed entity deduplication to obtain a clean set of medical entities. Each KB entity was represented by its canonical name, descriptive paragraph text, and an optional associated image when available. To support reliable supervision, gold labels were obtained by aligning query-side gold references to KB entities via text-based matching, yielding a set of valid samples with gold annotations; samples without matched gold labels were retained in the constructed pool but excluded from supervised evaluation.

The only additional normalization we applied was removing leading figure–number prefixes (e.g., "Figure x.x") from mention descriptions to avoid spurious lexical cues. No other text cleaning or normalization was performed. This step was applied consistently to the training, validation, and test splits.

After gold filtering, we split valid samples into train/dev/test sets with a 70/10/20 ratio using a fixed random seed for reproducibility. As reported in Table 1, the final KB contained 16,509 deduplicated entities, and we constructed 1,878 total samples, among which 1,850 were valid with

gold labels. The resulting splits contained 1,295/185/370 samples for train/dev/test, respectively.

Representative medical cases illustrating modality missingness and modality imbalance in real-world medical multimodal data were retrieved. We compared against MM_Embed,[45] a multimodal embedding retrieval baseline that maps mention/entity text and images (when available) into a shared embedding space and ranks candidates by similarity.

3.6. Comparison of methods

We compared our framework with a diverse set of strong baselines covering general-purpose vision–language alignment, biomedical vision–language modeling, modern multimodal embedding retrieval, and text-only retrieval. Specifically, we included: (i) CLIP,[6] a general-domain contrastive vision–language model; (ii) BiomedCLIP,[40] a biomedical vision–language model pre-trained on large-scale biomedical image–text pairs; (iii) MM_Embed,[45] a multimodal embedding model for unified multimodal retrieval; (iv) NV_v2,[46] a strong text embedding model for dense retrieval; (v) BM25,[47] a classical lexical matching baseline; and (vi) fine-grained late-interaction multimodal retrieval (FLMR),[48] a fine-grained late-interaction multimodal retrieval model originally proposed for retrieval-augmented visual question answering. In addition to these baselines, we evaluated two variants of our method, GIMEL-V and GIMEL-G, which instantiated our framework with a VAE-based or a GAN-based conditional generative module, respectively.

We reported two standard ranking metrics: Hits@k and mean reciprocal rank (MRR). Hits@k measured whether the gold entity appeared in the top-k ranked candidates, while MRR averaged the reciprocal rank of the first correct entity, providing a balanced evaluation of ranking quality. Unless otherwise stated, we reported Hits@{1,3,5} and MRR on the test split.

3.7. Ablation study

To evaluate the contribution of each key component in the proposed framework, we conducted an ablation study. Specifically, six model variants were compared. The base model performed retrieval without using the generative completion module or the dynamic loss balancing strategy. The dynamic-loss-only model incorporated the dynamic loss balancing strategy but did not use generative completion. The VAE-only and GAN-only models used VAE-based and GAN-based generative completion modules, respectively, to synthesize missing visual representations, but did not apply dynamic loss balancing. Finally, GIMEL-V and GIMEL-G represented the complete versions of the proposed framework, incorporating the dynamic loss balancing strategy along with the VAE-based and GAN-based completion modules, respectively.

All model variants were trained and evaluated under the same experimental settings, including the same dataset split, training protocol, hyperparameter configuration, and evaluation procedure. The performance of each variant was assessed using the same retrieval metrics, including Hits@1, Hits@3, Hits@5, and MRR. This ablation analysis was designed to examine the individual and combined effects of the generative completion module and the dynamic loss balancing strategy on retrieval performance.

4. Results and discussion

Before examining each research question in detail, it is useful to note a general pattern in Tables 3 and 4. Performance differences across methods were not random; rather, they closely followed the degree of modality completeness and the structural characteristics of each domain. Domains with richer visual evidence and clearer visual–text correspondence tend to benefit more from multimodal modeling, whereas domains with sparse KB-side images or limited sample sizes exposed clear weaknesses in existing methods. This overall trend indicates that robustness in medical multimodal retrieval is shaped not only by model architecture, but also by the distribution and asymmetry of modality availability in the underlying data.

Research question 1: Do modality missingness and modality imbalance widely exist in real-world medical multimodal retrieval, and to what extent do they undermine the robustness of existing methods?

To answer this research question, we conducted a systematic analysis of modality missingness and modality imbalance in real-world medical multimodal retrieval from both quantitative and qualitative perspectives. Table 3 reports the performance of different methods across the overall dataset and individual medical domains, while Table 4 summarizes the availability of visual modalities on both the query side and the KB side. In addition, Figure 3 presents representative examples that illustrate typical failure modes caused by missing modalities.

4.1. Overall quantitative analysis

From the overall results in Table 3 (Overall [Global]), we observed substantial performance gaps among existing multimodal and text-based retrieval methods on real medical data. Vision–language models such as CLIP and BiomedCLIP exhibited relatively weak performance on both Hits@1 and MRR, while text-based methods achieved moderate improvements but still suffered from limited ranking quality. These results indicate that, under realistic

Table 1. Domain-level statistics of the dataset

Domain	Knowledge base entities	Constructed	Validated	Trained	Developed	Tested
Anatomy & physio	2,264	565	565	392	62	111
Cardio	2,253	63	62	46	7	9
Cell bio	3,314	322	296	227	22	47
Human repro	365	13	12	8	1	3
Microbiology	1,638	429	429	294	47	88
Neuroscience	111	21	21	13	3	5
Pulmonary	1,446	75	75	56	8	11
Pulmo2	2,130	89	89	62	10	17
Terms	1,366	109	109	71	10	28
Ultrasound	388	138	138	94	8	36
Undergradimage	1,234	54	54	32	7	15
Total	16,509	1,878	1,850	1,295	185	370

Table 2. Overall statistics of the dataset

Item	Count
Knowledge base entities (deduplicated)	16,509
Total samples (including no-gold)	1,878
Valid samples (with gold)	1,850
Training set (70%)	1,295
Validation set (10%)	185
Test set (20%)	370

medical settings with imperfect multimodal information, the expected advantages of multimodal models are not consistently realized. When visual information is missing, unavailable, or semantically weak, multimodal systems often degenerate into relying on a single dominant modality, leading to unstable ranking behavior and degraded robustness.[49] After introducing FLMR as an additional recent multimodal ranking baseline, we observed a stronger and more informative comparison. As shown in Table 3, FLMR clearly outperformed earlier baselines such as CLIP, BiomedCLIP, MM_Embed, NV_v2, and BM25 on the global benchmark, achieving 0.5351 on Hits@1 and 0.6885 on MRR. This indicates that fine-grained late-interaction retrieval is indeed a stronger baseline than conventional shared-embedding or text-only ranking methods in medical multimodal retrieval. However, both variants of our framework still achieved better overall performance than FLMR. In particular, GIMEL-V reached 0.6243 on Hits@1 and 0.7561 on MRR, while GIMEL-G achieved 0.6189 on Hits@1 and 0.7576 on MRR. These results show that the improvements of GIMEL were not limited to comparisons with conventional baselines but remained valid even when compared with a stronger recent ranking model.

This comparison is important because it strengthens the claim that our gains did not arise merely from using a better encoder or a favorable evaluation setting. Rather, the superiority of GIMEL over FLMR suggests that explicitly addressing missing modality semantics and modality imbalance is more effective for this task than relying on stronger interaction-based ranking alone. In other words, fine-grained interaction helps, but it does not by itself resolve the structural incompleteness of medical multimodal evidence.

4.2. Modality missingness and imbalance in data distribution

To further quantify modality missingness in real-world data, we analyzed the availability of image modalities in the training, validation, and test splits, as well as in the KB. As shown in Table 4, image availability on the query side was generally high, with only a negligible proportion of missing or corrupted files across all splits. In contrast, the KB exhibited severe visual modality missingness: more than 75% of KB entities lacked valid images, and only 4,048 entities had usable visual information. This striking asymmetry reveals a critical characteristic of real medical retrieval scenarios, namely, a structural imbalance in modality availability between the query side and the KB side. Under such conditions, even when a query contains complete multimodal information, the retrieval process must compare it against a large number of entities with missing visual modalities.[50] This imbalance severely

hampers cross-modal semantic alignment, destabilizes similarity estimation, and ultimately degrades ranking reliability.[51] These observations provide a data-driven explanation for the pronounced performance fluctuations of multimodal methods observed in Table 3.

4.3. Domain-level performance analysis

A closer look at domain-specific results reveals that the impact of modality missingness is far from uniform across medical subdomains. In visually rich, structurally well-defined areas—such as Anatomy & physio, Ultrasound, and Undergradimage—multimodal methods generally performed well, yet persistent gaps in Hits@1 and MRR suggest that semantic inconsistency across modalities can still obstruct precise entity matching, even when visual information is plentiful. The findings were considerably worse in data-sparse domains with frequent image absence, including Cardio, Pulmonary, Pulmo2, and Neuroscience, where most baseline methods recorded near-zero Hits@1 scores. This makes it clear that existing approaches struggle to produce stable, discriminative representations when critical modalities are missing, leading to a sharp deterioration in robustness. Text-dominant domains such as Terms and Cell bio exhibited a distinct pattern: while text-based models performed comparably on certain metrics, their limitations became apparent in Top-3 and Top-5 rankings, highlighting the inadequacy of single-modality representations for distinguishing among semantically complex medical entities. The comparison with FLMR also revealed an informative domain-level pattern. In visually rich domains such as Anatomy & physio, Ultrasound, and Undergradimage, FLMR was already highly competitive, achieving 0.7838/0.8702, 0.8333/0.8806, and 0.9333/0.9667 on Hits@1/MRR, respectively. This suggests that fine-grained late interaction can effectively exploit strong visual–text correspondences when multimodal evidence is relatively complete. Even so, GIMEL remained competitive or superior in these domains, indicating that the proposed framework does not sacrifice performance in well-formed multimodal settings.

More importantly, FLMR showed clear instability in domains where modality sparsity and KB-side incompleteness were more severe, such as Cardio, Cell bio, Pulmonary, and Pulmo2. For example, in Cell bio, FLMR achieved only 0.0213 on Hits@1 and 0.2154 on MRR, which were substantially lower than those of both GIMEL variants. Similar gaps can be observed in Pulmo2 and Pulmonary. These results further support our central argument: although stronger ranking interactions can improve retrieval quality, robust medical multimodal entity linking still depends critically on recovering missing modality semantics and mitigating optimization bias under asymmetric modality availability.

4.4. Case study

Figure 3 shows a representative example from the Anatomy & physio domain with high multimodal dependency and notable domain-wise variability. The query mention was "mediastinum," accompanied by a short radiology-oriented description (in the context of computed tomography pulmonary embolism), while the KB-side visual evidence was partially unavailable, leading to an asymmetric modality setting in which the model had to rely heavily on incomplete cross-modal matching signals. As illustrated, the baseline retriever (MM_Embed) tends to return candidates that are superficially related to the query text (e.g., generic attributions or broader anatomical/educational entries such as functional anatomy), which are topically relevant but conceptually mismatched in terms of category or granularity. This reflects a typical failure mode under KB-side image missingness: once the visual grounding signal is weakened, ranking becomes dominated by textual overlap and coarse semantic proximity, making the system prone to drifting toward generic or loosely related concepts.[52] In contrast, our method produced a more coherent candidate set, promoting domain-consistent anatomy concepts (e.g., mediastinal contours and mediastinum) to the top ranks. This case highlights that modality missingness not only reduces ranking accuracy but also amplifies ambiguity among semantically close medical candidates; recovering missing modality semantics and stabilizing modality contributions are therefore crucial for robust medical multimodal retrieval in realistic KB-imperfect settings.

Based on the above quantitative results, data distribution analysis, and case studies, it is evident that modality semantic missingness and modality imbalance are pervasive and intrinsic challenges in real-world medical multimodal retrieval. Notably, these issues primarily stem from large-scale visual modality absence on the KB-side rather than from query-side incompleteness.[50] These structural characteristics significantly undermine the robustness and generalization ability of existing methods across different medical domains. More importantly, they highlight the need for principled solutions to recover missing modality semantics and mitigate optimization bias induced by modality imbalance, which we addressed in the subsequent research questions.

4.5. Effectiveness of generative completion for missing modality semantics

Research question 2: Is our framework effective in addressing missing modality semantics under real-world medical multimodal entity linking settings?

Table 3. Quantitative comparison of different retrieval models across the overall dataset and individual medical domains

Domain	Metric	Models							
		CLIP	BiomedCLIP	MM_embed	FLMR	NV_v2	BM25	GIMEL-V	GIMEL-G
Overall (Global)	Hits@1	0.2108	0.3135	0.5081	0.5351	0.4216	0.4135	0.6243[a]	0.6189[b]
	Hits@3	0.3838	0.5000	0.7919	0.8108[b]	0.6568	0.6459	0.8703[a]	0.8703[a]
	Hits@5	0.4973	0.6216	0.8676	0.8730	0.7378	0.7405	0.9108[b]	0.9189[a]
	MRR	0.3278	0.4372	0.6678	0.6685	0.5562	0.5377	0.7561[b]	0.7576[a]
Anatomy & physio	Hits@1	0.2432	0.3874	0.7748	0.7838[b]	0.5586	0.4685	0.7838[a]	0.7658
	Hits@3	0.3784	0.5811	0.8604	0.9459	0.7208	0.6352	0.9820[a]	0.9730[b]
	Hits@5	0.5135	0.7748	0.9459	0.9640	0.8829	0.8018	0.9910[b]	1.0000[a]
	MRR	0.3619	0.5428	0.8591	0.8702	0.6878	0.5947	0.8759[a]	0.8727[b]
Cardio	Hits@1	0.0000[b]	0.0000[b]	0.0000[b]	0.0000[b]	0.0000[b]	0.0000[b]	0.1111[a]	0.0000[b]
	Hits@3	0.1111[b]	0.0556	0.1111[b]	0.0000	0.0556	0.1111[b]	0.2222[a]	0.1111[b]
	Hits@5	0.2222	0.1111	0.2222	0.1111	0.1111	0.2222	0.5556[a]	0.4444[b]
	MRR	0.0500	0.0679	0.1074	0.1049	0.0494	0.0500	0.2784[a]	0.1740
Cell bio	Hits@1	0.0213	0.0851	0.0000	0.0213	0.0851	0.1277[b]	0.1277[b]	0.1915[a]
	Hits@3	0.1277	0.1808	0.2021	0.2766	0.1489	0.2234	0.4681[a]	0.4468[b]
	Hits@5	0.2340	0.2766	0.4043	0.4468	0.2128	0.3191	0.7021[a]	0.6809[b]
	MRR	0.1077	0.1619	0.1900	0.2154	0.1364	0.1865	0.3507[b]	0.3956[a]
Human repro	Hits@1	0.3333	0.6667[b]	1.0000[a]	1.0000[a]	1.0000[a]	0.6667[b]	1.0000[a]	1.0000[a]
	Hits@3	0.6667[b]	0.6667[b]	1.0000[a]	1.0000[a]	1.0000[a]	0.6667[b]	1.0000[a]	1.0000[a]
	Hits@5	1.0000[a]	0.6667[b]	1.0000[a]	1.0000[a]	1.0000[a]	0.6667[b]	1.0000[a]	1.0000[a]
	MRR	0.6667	0.7143[b]	1.0000[a]	1.0000[a]	1.0000[a]	0.6667	1.0000[a]	1.0000[a]
Microbiology	Hits@1	0.3182	0.4318	0.4091	0.4773	0.5114	0.4773	0.5568[a]	0.5227[b]
	Hits@3	0.4716	0.5681	0.7046	0.9773	0.7430	0.6160	0.977[b]3	0.9886[a]
	Hits@5	0.6250	0.7045	1.0000[a]	1.0000[a]	0.9545	0.9545	1.0000[a]	0.9886[b]
	MRR	0.4374	0.5411	0.6922	0.7254	0.7162	0.6708	0.7670[a]	0.7519[b]
Neuroscience	Hits@1	0.0000	0.0000	0.4000[a]	0.2000[a]	0.0000	0.2000[b]	0.4000[a]	0.2000[b]
	Hits@3	0.2000	0.2000	0.6000	0.8000[b]	0.2000	0.4000	1.0000[a]	1.0000[a]
	Hits@5	0.4000	0.4000	0.8000[b]	0.8000[b]	0.4000	0.6000	1.0000[a]	1.0000[a]
	MRR	0.1639	0.1022	0.5917	0.4583	0.2250	0.3067	0.6667[a]	0.5333[b]
Pulmo2	Hits@1	0.0000	0.0000	0.0000	0.0588	0.1176	0.1765[b]	0.4118[a]	0.4118[a]
	Hits@3	0.0294	0.0882	0.4118	0.7059	0.2353	0.2942	0.7059[b]	0.7647[a]
	Hits@5	0.0588	0.1765	0.8235[b]	0.8235[b]	0.3529	0.4118	0.7647	0.8824[a]
	MRR	0.0359	0.0662	0.2822	0.3714	0.2236	0.2696	0.5850	0.6113[a]
Pulmonary	Hits@1	0.0000	0.0000	0.0000	0.0909	0.0909	0.0909	0.2727[a]	0.1818[b]
	Hits@3	0.1818	0.2273	0.4545[b]	0.2727	0.3636	0.3182	0.6364[a]	0.6364[a]
	Hits@5	0.3636	0.4545	0.9091[a]	0.8182	0.6364	0.5455	0.8182[b]	0.7273
	MRR	0.1212	0.1776	0.2500	0.2939	0.2864	0.2227	0.4887[a]	0.4515[b]

(*Cont'd...*)

Table 2.(*Continued*)

Domain	Metric	Models							
		CLIP	BiomedCLIP	MM_embed	FLMR	NV_v2	BM25	GIMEL-V	GIMEL-G
Terms	Hits@1	0.3214	0.3214	0.5714	0.6429	0.4286	0.3929	0.7143[a]	0.6786[b]
	Hits@3	0.5000	0.4821	0.7680	0.9286[b]	0.6071	0.5536	0.9286[b]	1.0000[a]
	Hits@5	0.6786	0.6429	0.9643[b]	0.9643[b]	0.7857	0.7143	0.9286	1.0000[a]
	MRR	0.4691	0.4539	0.7440	0.7938	0.5753	0.5327	0.8286	0.8333[a]
Ultrasound	Hits@1	0.2222	0.4444	0.9167[b]	0.8333	0.5833	0.6667	0.9722[a]	0.9722[a]
	Hits@3	0.4306	0.6250	0.9306[b]	0.9167	0.7083	0.7917	1.0000[a]	1.0000[a]
	Hits@5	0.6389	0.8056	0.9444[b]	0.9444[b]	0.8333	0.9167	1.0000[a]	1.0000[a]
	MRR	0.3894	0.5733	0.9283	0.8806	0.6883	0.7671	0.9815[b]	0.9861[a]
Undergradimage	Hits@1	0.2667	0.2667	0.8000	0.9333[b]	0.4000	0.7333	0.9333[b]	1.0000[a]
	Hits@3	0.3667	0.4333	0.9000[b]	1.0000[a]	0.5333	0.8000	1.0000[a]	1.0000[a]
	Hits@5	0.4667	0.6000	1.0000[a]	1.0000[a]	0.6667	0.8667[b]	1.0000[a]	1.0000[a]
	MRR	0.3472	0.4407	0.8778	0.9667[b]	0.5067	0.7722	0.9667[b]	1.0000[a]

Note: [a] represents the best values, while [b] represents the second-best values.
Abbreviations: CLIP: Contrastive language–image pre-training; FLMR: Fine-grained late-interaction multi-modal retrieval; GIMEL: Generative incomplete-modality multimodal entity linking; MRR: Mean reciprocal rank.

Table 4. Statistics of image availability

Split	Total	Null	Miss.	Corr.	Valid	Rate (%)
Train	1,295	0	0	4	1,291	0.31
Validated	185	0	0	0	185	0.00
Test	370	0	0	2	368	0.54
Knowledge base	16,509	12,431	24	6	4,048	75.48

Note: Null: No path provided; Miss.: File missing; Corr.: File corrupt.Rate (%) represents error rate (Corr./Total) for dataset splits, and Null/Total for the knowledge base.

Missing modalities in real medical multimodal entity linking are often structural rather than occasional noise. As shown in Table 4, query-side images were largely available across the train/val/test splits, whereas the KB exhibited severe visual missingness: more than 75% of KB entities lacked valid images. This induces a pronounced query–KB modality asymmetry, where the model must rank a large number of candidates whose visual semantics are absent, making robust multimodal matching substantially harder.

To assess whether our generative completion module can effectively recover missing modality semantics, we conducted an ablation study, summarized in Figure 4. Specifically, we compared (i) the base model, which performed retrieval without generative completion; (ii) the VAE-only and GAN-only models, which enabled conditional generation to synthesize missing visual representations but did not apply any dynamic loss balancing; and (iii) our complete GIMEL-V and GIMEL-G models, which instantiated the proposed framework with a VAE-based or GAN-based completion module, respectively.

4.6. Impact of generative completion

Figure 4 clearly demonstrates the benefit of generative completion: every variant outperformed the base model, with nontrivial margins. The VAE-only model yielded measurable gains, with Hits@1 improving from 0.573 to 0.595 and MRR from 0.721 to 0.730. The GAN-only model performed better, achieving 0.603 on Hits@1 and 0.742 on MRR. Synthesizing missing visual representations introduces discriminative information that text alone cannot provide, resulting in more stable rankings when KB-side visual semantics are absent.[49] The GAN variant marginally outperformed VAE on MRR and Hits@5, suggesting that adversarially regularized features offer slightly better separability for closely related medical

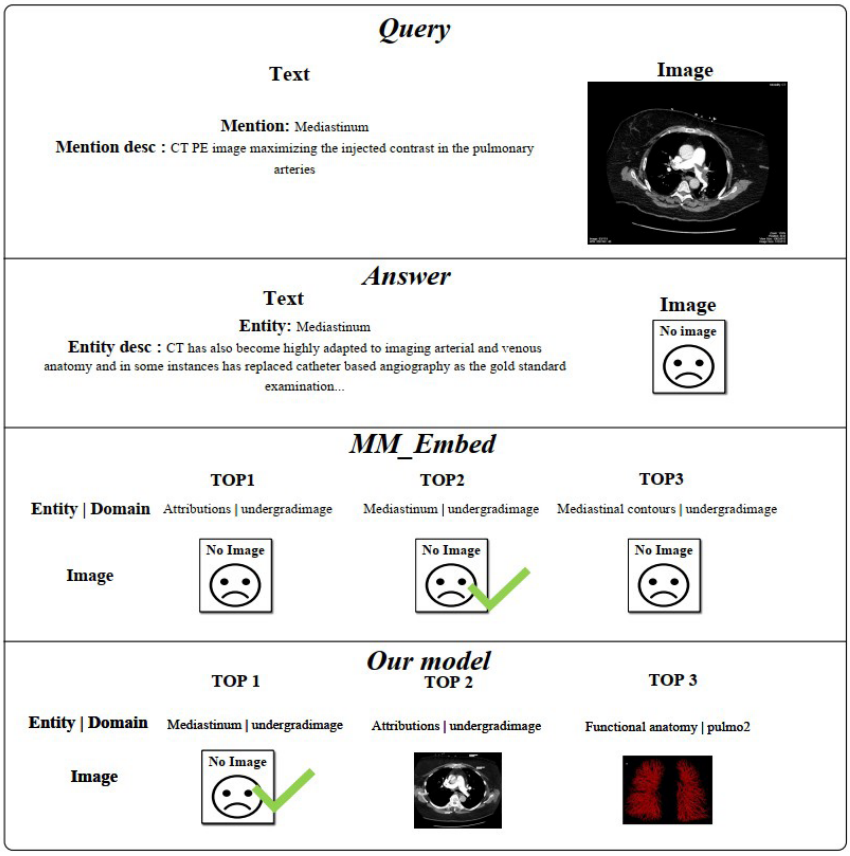


Figure 3. Representative medical cases illustrating modality missingness and modality imbalance in real-world medical multimodal retrieval. We compared against MM_Embed,[5] a multimodal embedding retrieval baseline that maps mention/entity text and images (when available) into a shared embedding space and ranks candidates by similarity.
Abbreviation: CT PE: Computed tomography pulmonary angiogram.

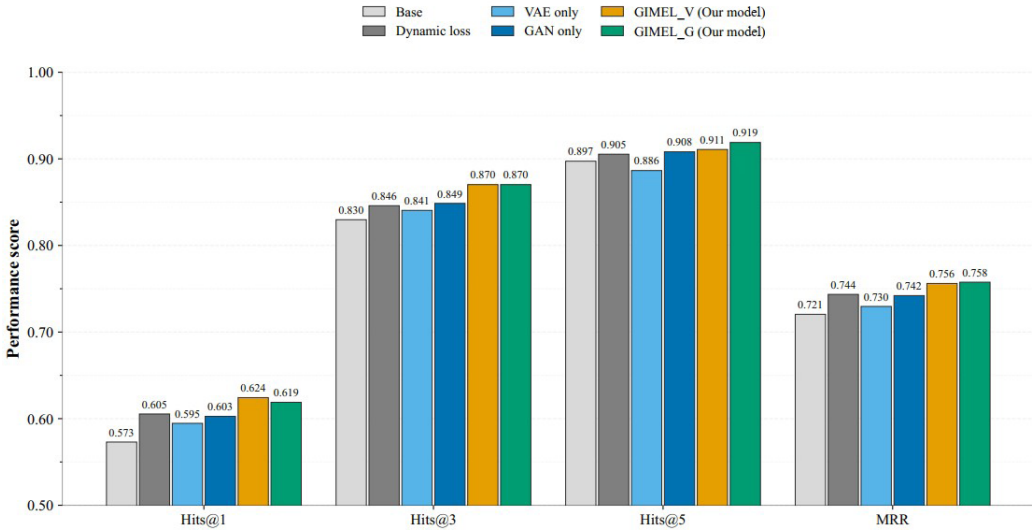


Figure 4. Ablation study results analyzing the contribution of different components in our framework
Abbreviations: GAN: Generative adversarial network; GIMEL: Generative incomplete-modality multimodal entity linking; MRR: Mean reciprocal rank; VAE: Variational autoencoders.

concepts, though the gap remains small. This is expected: the goal here was not unrestricted image generation but retrieval-oriented feature completion, where recovering sufficient discriminative semantics matters more than sample realism. Once the features were generated, different generator families yielded only modest differences in final ranking metrics. Overall, these results confirm that the conditional generative component effectively restores missing modality semantics under realistic KB-side incompleteness. Since incompleteness in the current benchmark predominantly occurs on the KB side, recovering even partial visual information for KB entities reshapes the candidate space and improves cross-modal comparability, explaining why generative completion consistently and structurally improves ranking quality.

It is worth emphasizing that this section focused on the effectiveness of generative completion itself (research question 2). As shown in Figure 4, even without any additional balancing strategy (VAE-only/GAN-only), completing missing visual representations yielded stable and evident improvements. This corroborates that missing modality semantics is a key bottleneck in our setting: when KB-side images are largely unavailable (Table 4), recovering the missing modality directly enhances the discriminability and consistency of candidate ranking. Meanwhile, the complete models (GIMEL-V/G) further achieved the best overall performance, suggesting that generative completion and a better training objective/optimization scheme are complementary. We defer a detailed investigation of the optimization aspect to research question 3, where we analyzed why such complementarity led to improved training stability and robustness.

4.7. Dynamic loss balancing for mitigating optimization imbalance

Research question 3: Can dynamic loss balancing mitigate modality-induced optimization bias and improve training stability as well as final performance?

To validate the effect of dynamic loss balancing on optimization imbalance, we analyzed the training process from two complementary perspectives: (i) training dynamics, by examining how key loss terms evolve across epochs; and (ii) final effectiveness, by comparing the performance with and without dynamic balancing under the same experimental setting.

Figure 5 illustrates the evolution of two key losses (the overall objective and the image–text related loss) before and after applying loss balancing. When no balancing was applied, the two losses behaved quite differently during training—their minima landed at different epochs, and both curves showed fairly pronounced fluctuations. A likely explanation is that gradients from different modality-specific objectives alternate in dominance, steering optimization toward one modality at a time and undermining the coordinated learning that multimodal training is intended to promote.[53] Once dynamic loss balancing was introduced, the picture changed

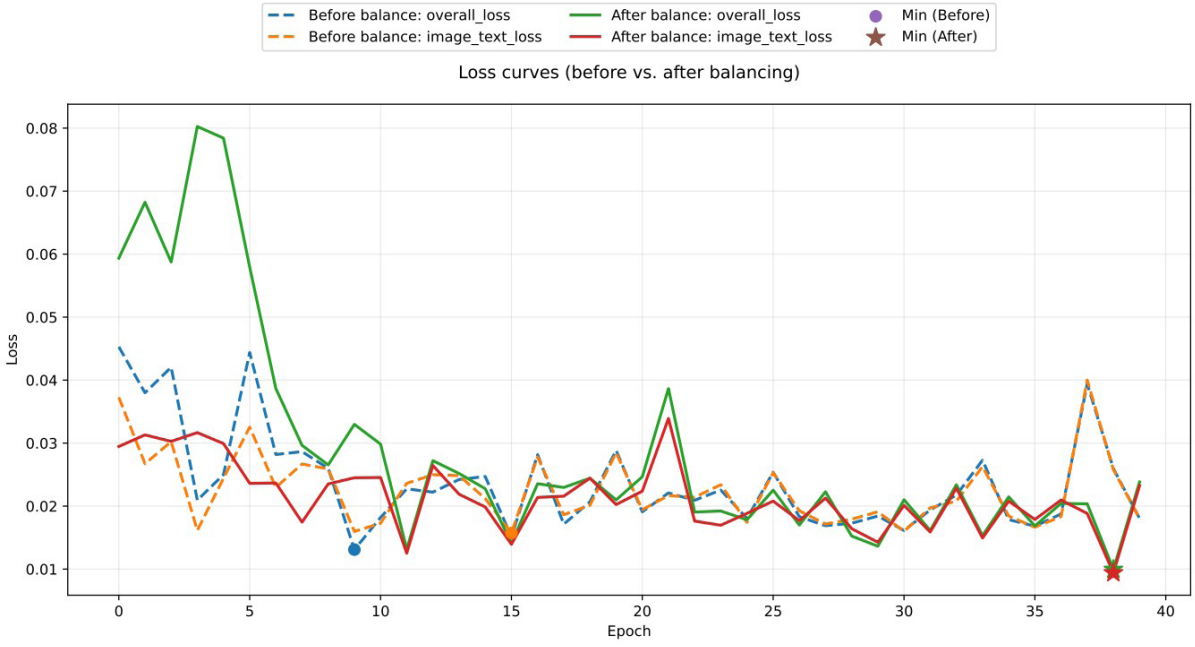


Figure 5. Training loss curves before and after loss balancing. We compared overall_loss and image_text_loss across epochs; markers indicate the minimum loss of each curve, highlighting improved convergence after balancing.

noticeably. With balancing, the two loss curves grew more aligned—their minima were closer together in time, and the fluctuations largely disappeared. This was not a minor cosmetic difference; the tighter coupling between objectives helped the model learn both modalities more coherently, rather than lurching between them.

Figure 5 adds another angle to this: using dynamic loss alone, without any of the other components, produced consistent gains over the base model across ranking metrics. Loss balancing is therefore not merely a matter of stabilizing the training process; it has a tangible impact on the model's ability to retrieve and link entities.

Taken together, these results indicate that optimization imbalance is not a theoretical aspect of the missing-modality problem—it is reflected in empirical performance. The improvements from dynamic balancing cannot be attributed solely to enhanced representations; rather, maintaining alignment among training objectives appears to play an independent and critical role. This is essentially the core argument of the study: handling missing modalities well requires both recovering missing semantic content and preventing the optimization from drifting, particularly when modality availability is as uneven as it tends to be in real medical data.

5. Conclusion

In this study, we investigated medical multimodal entity linking under realistic modality-missing conditions, where incomplete visual evidence and modality imbalance often reduce retrieval reliability. To address this issue, we proposed GIMEL, a unified framework that improves robustness from two complementary aspects: a plug-and-play conditional generative module for recovering missing visual semantics, and a dynamic loss balancing strategy for mitigating optimization bias caused by asymmetric modality availability. Experiments on a benchmark covering 11 medical domains showed that the proposed framework consistently improved ranking performance and training stability over strong baselines. These results further suggest that the challenge of medical multimodal entity linking is shaped not only by semantic ambiguity, but also by the structural incompleteness of real-world multimodal data.

Despite these encouraging results, this work still has several limitations. The current benchmark was reorganized from an existing medical retrieval dataset and does not yet cover all clinical scenarios or forms of multimodal evidence. In addition, the framework mainly focused on missing visual modality and emphasized robustness at the retrieval stage, without exploring broader end-to-end optimization. Future work could build larger, more diverse benchmarks, explore more general completion mechanisms across multiple degraded modalities, and integrate robust retrieval more tightly with downstream medical reasoning and retrieval-augmented generation systems.

Acknowledgments

None.

Funding

This work is partly supported by the Youth Talent Support Program of Anhui Association for Science and Technology (No. RCTJ202420), the Hefei Key Technology R & D Champion-Based Selection Project (No. 2024SGJ010), and the Open Project of Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China), under grant number BigKEOpen2025-08.

Conflict of interest

Chenyang Bu, He Ren, and Weiping Ding are Guest Editors of this special issue, with Chenyang Bu also serving as an Editorial Board Member for the journal. Separately, all the authors declared that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Author contributions

Conceptualization: He Ren, Chenyang Bu
Formal analysis: Shiji Zang
Funding acquisition: He Ren, Chenyang Bu
Investigation: Shiji Zang, Yunpeng Hong
Methodology: Yunpeng Hong
Project administration: Chenyang Bu
Software: Shiji Zang
Supervision: Chenyang Bu, Weiping Ding
Writing–original draft: Shiji Zang
Writing–review & editing: Chenyang Bu, Yunpeng Hong, Weiping Ding

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The source code is available at: https://github.com/zsjreal/GIMEL_medicine. The raw data used in this study are derived from the publicly available M3Retrieve benchmark, specifically the M3Retrieve_VisualContextRetrieval subset. We reorganized the retrieval-style corpus into a

medical multimodal entity linking (MEL) benchmark by treating each query as a mention and each retrievable passage as a KB entity, and by constructing fixed train/dev/test splits as described in the manuscript. The processed benchmark files and the code used for data preprocessing and experiments are available from the corresponding author upon reasonable request via email.

References

1. Li M, Kilicoglu H, Xu H, Zhang R. BiomedRAG: a retrieval augmented large language model for biomedicine. *J Biomed Inform*. 2025;162:104769.

   doi: 10.1016/j.jbi.2024.104769.

2. Xiong G, Jin Q, Wang X, Zhang M, Lu Z, Zhang A. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In: *Biocomputing 2025*. World Scientific. 2024:199-214.

   doi: 10.1142/9789819807024_0015

3. Akhtar ZB. Artificial intelligence within medical diagnostics: A multi-disease perspective. *Artif Intell Health*. 2025;2(3):44-62.

   doi: 10.36922/aih.5173.

4. Neverova N, Wolf C, Taylor GW, Nebout F. ModDrop: adaptive multi-modal gesture recognition. *IEEE Trans Pattern Anal Mach Intell*. 2016;38(8):1692-1706.

   doi: 10.1109/TPAMI.2015.2461544.

5. Chen Z, Guo L, Fang Y, *et al*. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In: *Lecture Notes in Computer Science*. Cham, Switzerland: Springer Nature. 2023:121-139.

   doi: 10.1007/978-3-031-47240-4_7

6. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, *et al*. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). 2021. Available from: https://proceedings.mlr.press/v139/radford21a/radford21a.pdf.

7. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2022:3876-3887.

   doi: 10.18653/v1/2022.emnlp-main.256

8. Zhang S, Xu Y, Usuyama N, *et al*. BiomedCLIP: a multimodal biomedical foundation model pretrained on biomedical images and text. *arXiv*. 2023.

   doi: 10.48550/arXiv.2303.00915

9. Lin W, Zhao Z, Zhang X, Wu C, Zhang Y, Wang Y, Xie W. PMC-CLIP: contrastive language-image pre-training using biomedical documents. *arXiv*. 2023.

   doi: 10.48550/arXiv.2303.07240

10. Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE. 2018:7482-7491.

    doi: 10.1109/CVPR.2018.00781

11. Chen Z, Badrinarayanan V, Lee CY, Rabinovich A. GradNorm: gradient normalization for adaptive loss balancing in deep multitask networks. In: Proceedings of the International Conference on Machine Learning (ICML); 2018. Available from: https://proceedings.mlr.press/v80/chen18a.html.

12. Yu T, Kumar S, Gupta A, Levine S, Hausman K, Finn C. Gradient surgery for multi-task learning. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2020. Available from: https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

13. Li P, Wang Y. A multimodal entity linking approach incorporating topic concepts. *IEEE Trans Knowl Data Eng*. 2021;27(2):491-494.

    doi: 10.1109/CISAI54367.2021.00100

14. Bu C, Chang G, Chen Z, *et al*. Query-driven multimodal GraphRAG: dynamic local knowledge graph construction for online reasoning. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics; 2025:21360-21380.

    doi: 10.18653/v1/2025.findings-acl.1100

15. Gao X, Lu K. RefSAM3D: adapting the segment anything model with cross-modal references for three-dimensional medical image segmentation. *Artif Intell Health*. 2025;2(4):114-128.

    doi: 10.36922/AIH025080010

16. Zhao X, Zhao Q, Tanaka T. EpilepsyLLM: fine-tuning large language models for Japanese epilepsy knowledge representation. *Artif Intell Health*. 2025;3(1):104-115.

    doi: 10.36922/AIH025180042

17. Gan J, Luo J, Wang H, Wang S, He W, Huang Q. Multimodal entity linking: a new dataset and a baseline. In: Proceedings of the ACM International Conference on Multimedia. ACM. 2021:993-1001.

    doi: 10.1145/3474085.3475400.

18. Sanz-Cruzado J, Lever J. Accelerating cross-encoders in biomedical entity linking. In: Proceedings of the 24th Workshop on Biomedical Language Processing. Association for Computational Linguistics. 2025:136-147.

    doi: 10.18653/v1/2025.bionlp-1.13

19. Mumtaz U, Ahmed A, Mumtaz S. LLMs-Healthcare: Current

applications and challenges of large language models in various medical specialties. *Artif Intell Health.* 2024;1(2):16-28.

doi: 10.36922/aih.2558.

20. Song S, Zhao S, Wang C, *et al.* A dual-way enhanced framework from text matching point of view for multimodal entity linking. *AAAI*. 2024;38(17):19008-19016.

doi: 10.1609/aaai.v38i17.29867

21. Xing S, Zhao F, Wu Z, Li C, Zhang J, Dai X. DRIN: dynamic relation interactive network for multimodal entity linking. In: Proceedings of the ACM International Conference on Multimedia. ACM. 2023:3599-3608.

doi: 10.1145/3581783.3612575

22. Yao BM, Chen Y, Wang Q, *et al.* AMELI: enhancing multimodal entity linking with fine-grained attributes. *arXiv*. 2023.

doi: 10.48550/arXiv.2305.14725

23. Shi S, Xu Z, Hu B, Zhang M. Generative multimodal entity linking. *arXiv*. 2023.

doi: 10.48550/arXiv.2306.12725

24. Liu Q, He Y, Xu T, *et al.* UniMEL: a unified framework for multimodal entity linking with large language models. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM). ACM. 2024:1909-1919.

doi: 10.1145/3627673.3679793.

25. Vishwanath AB, Srinivasalu VK, Subramaniam N. Role of large language models in improving provider–patient experience and interaction efficiency: A scoping review. *Artif Intell Health.* 2025;2(2):1-10.

doi: 10.36922/aih.4808.

26. Vashishth S, Newman-Griffis D, Joshi R, Dutt R, Rosé CP. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *J Biomed Inform.* 2021;121:103880.

doi: 10.1016/j.jbi.2021.103880

27. Zeng J, Zhou J, Liu T. Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. *IEEE Trans Multimed.* 2023;25(9):6301-6314.

doi: 10.1109/TMM.2022.3207572.

28. Huan R, Zhong G, Chen P, Liang R. UniMF: a unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. *IEEE Trans Multimed.* 2024;26:5753-5768.

doi: 10.1109/TMM.2023.3338769.

29. Luo P, Xu T, Wu S, Zhu C, Xu L, Chen E. Multi-grained multimodal interaction network for entity linking. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD). ACM. 2023:1583-1594.

doi: 10.1145/3580305.3599439

30. Wang H, Chen Y, Ma C, Avery J, Hull L, Carneiro G. Multi-modal learning with missing modality via shared-specific feature modelling.In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2023:15878-15887.

doi: 10.1109/cvpr52729.2023.01524

31. Akhtar MS, Chauhan DS, Ghosal D, Poria S, Ekbal A, Bhattacharyya P. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv*. 2019.

doi: 10.48550/arXiv.1905.05812

32. Wang Q, Zhan L, Thompson P, Zhou J. Multimodal learning with incomplete modalities by knowledge distillation. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM. 2020:1828-1838.

doi: 10.1145/3394486.3403234

33. Xue Z, Marculescu R. Dynamic multimodal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2023:2575-2584.

doi: 10.1109/CVPRW59228.2023.00256

34. Xaviar S, Yang X, Ardakanian O. Robust multimodal fusion for human activity recognition. *arXiv*. 2023.

doi: 10.48550/arXiv.2303.04636

35. Chen D, Zhang R. Building multimodal knowledge bases with multimodal computational sequences and generative adversarial networks. *IEEE Trans Multimed.* 2024;26:2027-2040.

doi: 10.1109/TMM.2023.3291503

36. Zhang Y, Peng C, Wang Q, Song D, Li K, Zhou SK. Unified multi-modal image synthesis for missing modality imputation. *IEEE Trans Med Imaging.* 2024;43(1):225-236.

doi: 10.1109/TMI.2024.3424785.

37. Zhao J, Li R, Jin Q. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. 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). Association for Computational Linguistics. 2021:2608-2618.

doi: 10.18653/v1/2021.acl-long.203

38. Zhou J, Tang J, Zuo Y, Wan P, Zhang D, Shao W. Robust Multimodal Survival Prediction with Conditional Latent Differentiation Variational AutoEncoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025:10384-10393

39. Emami H, Dong M, Glide-Hurst C. CL-GAN: contrastive learning-based generative adversarial network for modality transfer with limited paired data. In: *Lecture Notes in Computer Science*. Cham, Switzerland: Springer Nature. 2023:527-542.

    doi: 10.1007/978-3-031-25066-8_30

40. Zhang S, Xu Y, Usuyama N, *et al.* A multimodal biomedical foundation model trained from fifteen million image-text pairs. *NEJM AI*. 2025;2(1):AIoa2400640.

    doi: 10.1056/AIoa2400640.

41. Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv*. 2013.

    doi: 10.48550/arXiv.1312.6114

42. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2014;27. Available from: https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

43. Wang J, Black M, Rankin D, *et al.* A bagging ensemble machine learning method for imbalanced data to predict anxiety disorders and analysis of risk factors in older people: observational study. *Artif Intell Health*. 2025;1-22.

    doi: 10.36922/AIH025070009

44. Acharya A, Ghosh A, Verma P, Pasupa K, Saha S, Singh P. M3Retrieve: benchmarking multimodal retrieval for medicine. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2025:15274-15287.

    doi: 10.18653/v1/2025.emnlp-main.

45. Lee C, Roy R, Xu M, *et al.* NV-Embed: improved techniques for training LLMs as generalist embedding models. In: Proceedings of the International Conference on Learning Representations (ICLR). 2025. Available from: https://openreview.net/pdf?id=lgsyLSsDRe.

46. Lin J, Ma X, Lin SC, Yang JH, Pradeep R, Nogueira R. Pyserini: a Python toolkit for reproducible information retrieval research with sparse and dense representations. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM. 2021:2356-2362.

    doi: 10.1145/3404835.3463238.

47. Byrne B, Chen J, Coca A, Lin W, Mei J. Fine-grained Late-interaction Multi-modal Retrieval for Retrieval Augmented Visual Question Answering. In: *Advances in Neural Information Processing Systems 36*. San Diego, CA: Neural Information Processing Systems Foundation, Inc. (NeurIPS). 2023:22820-22840.

    doi: 10.52202/075280-0990

48. Lao M, Li Z, Guo Y, Zhang X, Cai S, Ding Z, *et al.* Boosting discriminability for robust multimodal entity linking with visual modality missing. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery. 2025:989-999.

    doi: 10.1145/3726302.3729906

49. Xiao G, Zeng W, Zhang S, Lao M, Zhao X. Multi-Modal Entities Matter: Benchmarking Multi-Modal Entity Alignment. In: Proceedings of the 31st International Conference on Computational Linguistics. Association for Computational Linguistics. 2025:8714-8724.

50. Wang Y, Sun H, Wang J, *et al.* Towards semantic consistency: Dirichlet energy driven robust multi-modal entity alignment. In: Proceedings of the 2024 IEEE 40th International Conference on Data Engineering (ICDE). IEEE. 2024:3559-3572.

    doi: 10.1109/ICDE60146.2024.00274

51. Sui X, Zhang Y, Zhao Y, Song K, Zhou B, Yuan X. MELOV: Multimodal Entity Linking with Optimized Visual Features in Latent Space. In: Proceedings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics. 2024:816-826.

    doi: 10.18653/v1/2024.findings-acl.46

52. Kwon J, Kim M, Lee E, Choi J, Kim Y. See-Saw Modality Balance: See Gradient, and Sew Impaired Vision-Language Balance to Mitigate Dominant Modality Bias. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics. 2025:4364-4378.

    doi: 10.18653/v1/2025.naacl-long.222