

DiA-gnostic VLVAE: Disentangled Alignment-Constrained Vision Language Variational AutoEncoder for Robust Radiology Reporting with Missing Modalities

Nagur Shareef Shaik¹, Teja Krishna Cherukuri¹, Adnan Masood², Dong Hye Ye¹,

¹Department of Computer Science, Georgia State University, Atlanta, GA, USA

²UST, Aliso Viejo, CA, USA

nshaik3@student.gsu.edu, tcherukuri1@student.gsu.edu, amasood@amp207.hbs.edu, dongye@gsu.edu

Abstract

The integration of medical images with clinical context is essential for generating accurate and clinically interpretable radiology reports. However, current automated methods often rely on resource-heavy Large Language Models (LLMs) or static knowledge graphs and struggle with two fundamental challenges in real-world clinical data: (1) **missing modalities**, such as incomplete clinical context, and (2) **feature entanglement**, where mixed modality-specific and shared information leads to suboptimal fusion and clinically unfaithful hallucinated findings. To address these challenges, we propose the **DiA-gnostic VLVAE**, which achieves robust radiology reporting through **Disentangled Alignment**. Our framework is designed to be resilient to missing modalities by disentangling shared and modality-specific features using a Mixture-of-Experts (MoE) based Vision-Language Variational Autoencoder (VLVAE). A constrained optimization objective enforces orthogonality and alignment between these latent representations to prevent suboptimal fusion. A compact LLaMA-X decoder then uses these disentangled representations to generate reports efficiently. On the IU X-Ray and MIMIC-CXR datasets, DiA has achieved competitive BLEU@4 scores of 0.266 and 0.134, respectively. Experimental results show that the proposed method significantly outperforms state-of-the-art models.

Introduction

Radiology report generation (RRG) is a critical task in medical imaging that aims to produce accurate and comprehensive reports from scans, which can help lessen the burden on radiologists. Despite progress in computer vision and natural language processing, RRG remains a significant challenge due to the need for precise clinical insight and coherent report synthesis. This is often complicated by imbalanced datasets where rare conditions are underrepresented, which can compromise diagnostic reliability (Yu et al. 2025).

Early models, such as R2Gen (Chen et al. 2020) and CvT2Dis (Nicolson et al. 2023), relied exclusively on image features, using transformers and contrastive learning to refine visual representations. However, this image-centric approach has difficulty capturing nuanced diseases and integrating clinical reasoning. Subsequent efforts focused on improving vision-language integration. For ex-

ample, XProNet utilized cross-modal prototypes for alignment (Wang et al. 2022), while METransformer used multiple learnable expert tokens to enhance textual consistency (Wang et al. 2023). Still, these models' reliance on image-centric patterns can lead to semantic discrepancies and clinical errors, especially when radiographic features of different diseases overlap, due to a lack of contextual grounding.

To address these limitations, recent models have begun to incorporate diagnostic context, such as disease pseudo-labels, knowledge graphs, or prior findings. Knowledge-driven approaches like MKSG (Yang et al. 2022) and M2KT (Yang et al. 2023) use medical knowledge graphs to improve factual accuracy. Context-aware models such as KiUT (Huang et al. 2023), DCL (Li et al. 2023b), EKA-Gen (Bu et al. 2024), and PromptMRG (Jin et al. 2024) have also integrated expert knowledge and prior reports through graphs and prompts. While these methods enhance the clinical relevance of the generated reports, they have several technical constraints. For instance, they often lack explicit disentanglement, making it difficult to separate modality-specific knowledge from shared information. Consequently, the absence of context can lead to incomplete reports due to inefficient multi-modal alignment. Additionally, prompt-based models often depend on templates constructed from pseudo-diagnoses, which limits their adaptability and can significantly increase computational overhead due to their use of Large Language Models (LLMs).

Retrieval-augmented methods like SEI (Liu et al. 2024) have advanced this area by extracting “factual entities” from a study, retrieving similar past cases, and using them to guide a cross-modal fusion decoder. However, this approach has its own issues. The entity-extraction and retrieval stages can be brittle, and the fusion network does not enforce explicit modality disentanglement or probabilistic feature gating. This leaves the model vulnerable to feature interference within what the authors term an “unstable fusion space”. Furthermore, when contextual information is missing, these models often fall back on deterministic rules instead of a principled probabilistic strategy, which can cause errors from earlier stages to propagate.

To tackle these challenges, we introduce the **DiA-gnostic VLVAE**, designed for **robust radiology reporting** by leveraging the principle of **Disentangled Alignment**. To handle

missing modalities and dynamic patient states, the framework uses real-time clinical data, including demographics, symptoms, and prior history, as dynamic context. Its core is a Vision-Language Variational Autoencoder (Mao et al. 2023) that disentangles modality-specific and shared latent representations, ensuring consistent vision-language alignment even when context is incomplete. This is supported by a Vision-Language Representation Learning module using Guided Context Attention (Cherukuri, Shaik, and Ye 2024) and a Modality Abstractor (Vaswani et al. 2017) for effective cross-modal feature fusion. Finally, a compact and efficient LLaMA-X decoder generates clinically precise reports, avoiding the template rigidity of prompt-based models (Jin et al. 2024) while outperforming more resource-intensive alternatives in adaptability and computational efficiency.

Related Work

Fusion of Heterogeneous Medical Data Fusing heterogeneous medical data, such as EHR, clinical notes, and various medical imaging types (Venugopalan et al. 2021; Mohsen et al. 2022), has shown significant potential for improving clinical tasks like prognosis prediction (Kline et al. 2022; Cheerla and Gevaert 2019), phenotyping (Hayat, Geras, and Shamout 2022), and medical image segmentation (Huang et al. 2020b). This integration of diverse data sources is a clear trend aimed at building more comprehensive and accurate clinical models (Huang et al. 2020a).

Handling Missing Modality In practice, some clinical data modalities are inevitably missing (Huang et al. 2020a). A common solution is late fusion, where predictions from independently modeled modalities are aggregated at the decision level (Yoo et al. 2019; Steyaert et al. 2023). However, this approach can be suboptimal as it fails to capture the interactions between modalities (Huang et al. 2020a). More recent research has explored generative methods to impute or reconstruct missing data at the feature or instance level (Ma et al. 2021; Zhang et al. 2022; Sharma and Hamarneh 2019). These techniques may use a Bayesian meta-learning framework (Ma et al. 2021) or impute features in the latent space with auxiliary information (Zhang et al. 2022). Despite these advances, results from generated data may not be robust (Li et al. 2023a; Yao et al. 2024), and handling missing data in highly heterogeneous settings like image-and-text fusion remains an open challenge (Yao et al. 2024).

Disentangled Representation Learning A promising approach for handling both missing data and modal inconsistency is to disentangle shared and modality-specific information (Yao et al. 2024; Liu et al. 2025; Robinet et al. 2024). The goal is to learn representations that separate common, patient-related information from unique, modality-specific details (Robinet et al. 2024). This is often achieved by imposing explicit constraints on the latent space. Common techniques include enforcing orthogonality between shared and specific representations to minimize redundancy (Braman et al. 2021; Yao et al. 2024) or minimizing their mutual information, often via an adversarial objective (Sanchez, Serrurier, and Ortner 2020; Liu et al. 2025; Robinet et al. 2024). Concurrently, the alignment of shared representations

is enforced using methods like Jensen-Shannon divergence (JSD) (Yao et al. 2024) or contrastive objectives (Robinet et al. 2024). While most prior work focused on more homogeneous modalities like different MRI scans (Chen et al. 2019; Shen and Gao 2019), DiA introduces a probabilistic tri-factor decomposition that leverages a Vision-Language VAE with a shared-gate Mixture-of-Experts and a unified Disentangled-Alignment constraint, enabling robust radiology reporting from highly heterogeneous inputs with missing modalities.

Methodology

The **DiA-gnostic VLVAE** is a principled probabilistic approach for **robust radiology reporting** designed to be resilient to **missing modalities** such as incomplete clinical context. The framework is built on the principle of **Disentangled Alignment**, which it achieves by learning a tri-factor latent space that explicitly separates modality-specific (vision, language) features from shared cross-modal semantics. To handle missing data, the shared latent is inferred via a Mixture-of-Experts (MoE) posterior, a theoretically grounded method that allows the model to marginalize an absent expert while preserving inferential integrity. This factorization is guided by a dual-consistency constraint: an **orthogonality** term disentangles the latent factors, while a **contrastive alignment** term ensures the shared space is predictive of each modality, leading to robust and faithful generation. This disentangled structure is learned by our novel Vision-Language Mixture-of-Experts Variational Auto-Encoder (VL-MoE-VAE) module and is used to drive report generation through an efficient LLaMA-X decoder.

Problem Formulation

Let our dataset be $\mathcal{D} = \{(V_i, L_i, R_i)\}_{i=1}^N$, where for each subject i , $V_i \in \mathbb{R}^{H \times W \times C}$ represents a medical image (e.g., Chest X-Ray), $L_i = \{l_{i,k}\}_{k=1}^{K_i}$ captures clinical indications (e.g., patient demographics, symptoms, prior history) with K elements, and $R_i = \{r_{i,t}\}_{t=1}^{T_i}$ is the corresponding radiology report. Our primary objective is to learn a conditional generative model $p(R | V, L)$ that maximizes the likelihood of producing the correct report R given the image V and the accompanying clinical context L . A critical principle for achieving robust reporting is *modality resilience*: the framework must remain effective even when one modality is absent, particularly the clinical context L . Consequently, the framework must also support principled inference for the marginal scenario $p(R | V)$.

Feature Extraction and Fusion

Before probabilistic modeling, we transform the raw, high-dimensional inputs into a unified, semantically rich feature space. This stage serves as a powerful feature extraction baseline, complementing DiA.

Vision & Language Feature Extractor We leverage a pre-trained convolutional neural network, EfficientNetB₀ (Tan and Le 2019), to extract high-level features from input image V . To capture clinically relevant global patterns that are often missed by local receptive fields, we

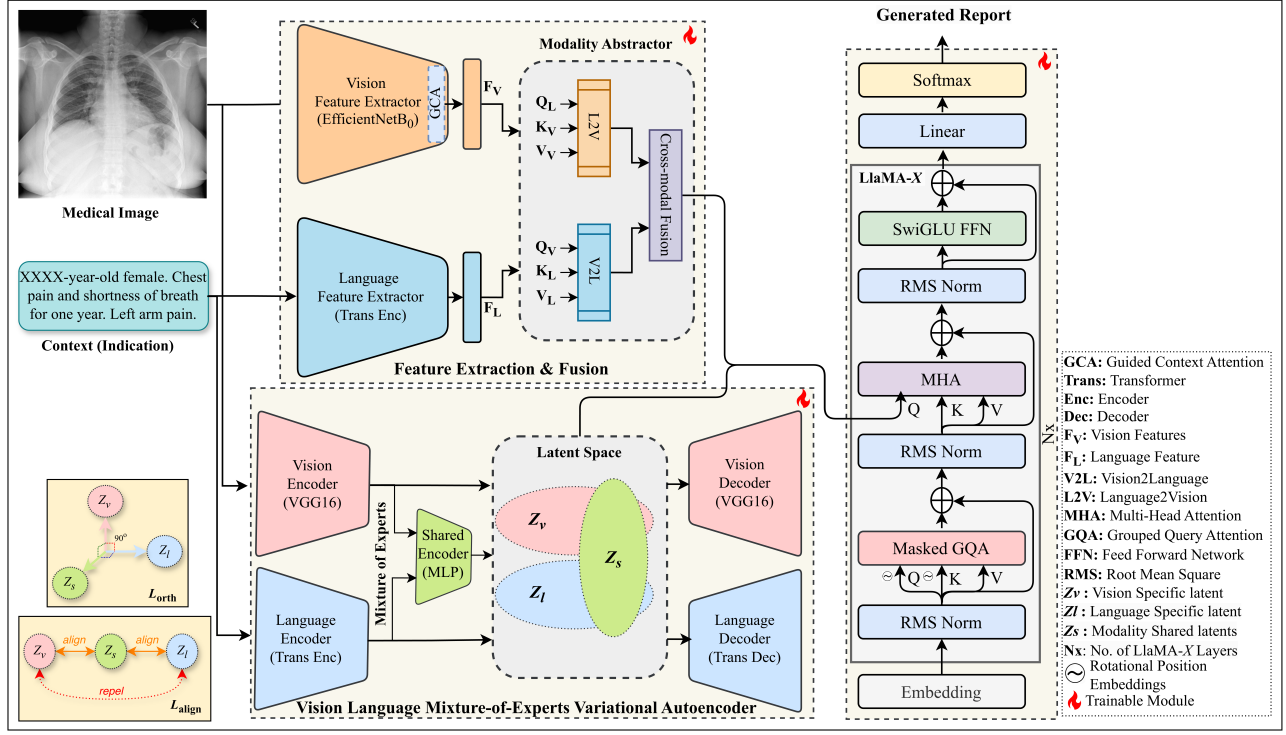


Figure 1: **Architecture of DiA**: Extracts vision features using *EfficientNetB₀* with *Guided Context Attention* and language features via a *Transformer Encoder*, fused by a *Modality Abtractor*; learns modality-specific latents (Z_v, Z_l) using VAEs (*VGG16* and *Transformer*) and shared latent (Z_s) through a *Mixture-of-Experts Shared Encoder*, disentangled via \mathcal{L}_{orth} , aligned with \mathcal{L}_{align} ; generate reports using *LlaMA-X* Decoder.

augment the backbone with a Guided Context Attention (GCA) (Cherukuri, Shaik, and Ye 2024) mechanism. This module produces a spatially-aware feature map that is projected into the final vision feature, $F_V \in \mathbb{R}^{S_V \times E}$, where S_V captures spatial dimensions and E is the number of feature channels. The clinical context L is tokenized and processed by a standard Transformer encoder (Vaswani et al. 2017) to capture complex semantic relationships, producing a sequence of contextualized embeddings $F_L \in \mathbb{R}^{S_L \times E}$, where S_L is the maximum sequence length.

Modality Abtractor To align and integrate these heterogeneous features, we use a Modality Abtractor based on bidirectional cross-attention (Vaswani et al. 2017). First, the vision features F_V and language features F_L are projected into query (Q), key (K), and value (V) representations using learnable weight matrices. The module then allows features from each modality to query the other, dynamically highlighting visually-grounded clinical terms and text-relevant image regions. This process computes both vision-to-language F_{V2L} and language-to-vision F_{L2V} representations via multi-head attention:

$$F_{V2L} = F_V + \text{Softmax} \left(\frac{Q_V \cdot K_L^T}{\sqrt{d_k}} \right) \cdot V_L \quad (1)$$

$$F_{L2V} = F_L + \text{Softmax} \left(\frac{Q_L \cdot K_V^T}{\sqrt{d_k}} \right) \cdot V_V \quad (2)$$

where d_k is the key vector’s dimension. The resulting features are concatenated to form a unified multi-modal representation F_{VL} , integrating complementary features for downstream VLVAE module.

Vision-Language Mixture-of-Experts VAE

We formulate DiA’s probabilistic framework using a Multi-modal Variational Autoencoder (MVAE) (see Fig. 1) that learns a **Tri-factor Latent Decomposition**. This decomposition is designed to disentangle the sources of variation in vision-language data into three distinct latent variables: a vision-specific latent Z_v , a language-specific latent Z_l , and a shared, cross-modal latent Z_s . As the true posterior over the latents, $p_\theta(Z_v, Z_l, Z_s | V, L)$, is intractable, we introduce a variational approximation with a specific factorization: $q_\phi(Z_v, Z_l, Z_s | V, L) \sim q_{\phi_v}(Z_v | V) \cdot q_{\phi_l}(Z_l | L) \cdot q_{\phi_s}(Z_s | V, L)$. Here, q_{ϕ_v} and q_{ϕ_l} are encoders for the modality-specific latents, while q_{ϕ_s} is a joint encoder for the shared latent, which uses a Mixture-of-Experts (MoE) strategy to ensure robustness against missing modalities.

Modality-Specific Latent Inference The model’s structure is guided by its generative process, which assumes that each observed modality is generated independently from its corresponding specific latent variable. For the vision modality, a latent variable Z_v is sampled from a prior distribution $p(Z_v)$, and the image is generated by a decoder $p_{\theta_v}(V | Z_v)$,

parameterized by θ_v . Similarly, the language latent Z_l is sampled from its prior $p(Z_l)$ to generate the clinical context via $p_{\theta_l}(L | Z_l)$, with parameters θ_l . This design introduces a critical inductive bias: all information necessary to reconstruct a modality must be encoded in its specific latent variable, which enforces representational independence and facilitates disentangled learning.

To learn the parameters, we need to infer the values of the latent variables from the data. This requires computing the true posterior distributions, $p_{\theta_v}(Z_v | V)$ and $p_{\theta_l}(Z_l | L)$, which are intractable to compute directly. To overcome this, we employ variational inference, introducing encoder networks to approximate these true but intractable posteriors. The vision encoder, $q_{\phi_v}(Z_v | V)$, uses a pre-trained VGG16 network (Simonyan and Zisserman 2014) followed by a fully connected layer to produce the Gaussian parameters (μ_v, σ_v^2) for the approximate posterior over Z_v . The language encoder, $q_{\phi_l}(Z_l | L)$, is a Transformer-based encoder (Liu and Liu 2019) that outputs (μ_l, σ_l^2) for the approximate posterior over the language-specific latent Z_l .

Shared Latent Inference via Mixture-of-Experts To model the shared latent variable Z_s , DiA employs a Mixture-of-Experts (MoE) strategy (Shi, Paige et al. 2019) via a dedicated shared encoder. This approach contrasts with Product-of-Experts (PoE) approaches (Wu and Goodman 2018), which can produce overconfident posterior estimates and degrade significantly when a modality is missing. The MoE formulation provides a more robust alternative for learning from partially observed data.

The shared encoder approximates the posterior over Z_s as a weighted combination of unimodal *expert* posteriors. For each modality $M \in \{V, L\}$, the encoder outputs parameters (μ_s, σ_s^2) and corresponding mixture weights π_M . The overall MoE posterior is then defined as:

$$q_{\phi_s}(Z_s | V, L) = \sum_{M \in \{V, L\}} \pi_M \cdot q_{\phi_s}(Z_s | M), \quad (3)$$

where the mixture coefficients π_M are non-negative and sum to one. This allows the model to adaptively the contribution of each modality to the shared representation.

Learning Objective The overall learning objective for the proposed VL-MoE-VAE is to maximize the Evidence Lower Bound (ELBO) (Mao et al. 2023) on the marginal log-likelihood. The ELBO balances accurate reconstruction with structured regularization over the latent space to enforce the desired disentangled alignment across Z_v , Z_l , and Z_s . The full objective is defined as:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} = & \mathbb{E}_{q_{\phi_s}(Z_s|V,L)} \left[\mathbb{E}_{q_{\phi_v}(Z_v|V)} [\log p_{\theta_v}(V|Z_v)] \right. \\ & \left. + \mathbb{E}_{q_{\phi_l}(Z_l|L)} [\log p_{\theta_l}(L|Z_l)] \right] \\ & - \left[D_{\text{KL}}(q_{\phi_v}(Z_v|V) \| p(Z_v)) + D_{\text{KL}}(q_{\phi_l}(Z_l|L) \| p(Z_l)) \right] \\ & - \text{JSD}(q_{\phi_s}(Z_s|V, L), p(Z_s)) \end{aligned} \quad (4)$$

This objective function evaluates the model’s ability to reconstruct the input modalities (V, L) from their respective

specific latents Z_v and Z_l , conditioned on a shared latent variable Z_s . It also encourages the modality-specific posteriors $q_{\phi_v}(Z_v | V)$ and $q_{\phi_l}(Z_l | L)$ to remain close to standard Gaussian priors $\mathcal{N}(0, I)$ via a Kullback-Leibler (KL) divergence penalty.

A key aspect of our Mixture-of-Experts (MoE) formulation is the use of Jensen-Shannon Divergence (JSD) (Menéndez et al. 1997) to regularize the shared latent Z_s . Unlike the standard KL divergence, which can lead to *component collapse* where only one expert contributes to the posterior (Minka et al. 2005), the symmetric and bounded nature of JSD is more suitable for mixture distributions. It encourages the entire mixture to align with the prior, promoting stability and ensuring all experts contribute meaningfully to the shared latent representation, a choice consistent with recent findings in multimodal generative modeling (Sutter, Daunhawer, and Vogt 2020; Yao et al. 2024).

Disentangled Alignment Constraint

The ELBO objective alone does not guarantee that the latent factors are either semantically meaningful or disentangled. To explicitly enforce the desired properties of disentanglement between shared and modality-specific factors, and strong alignment within the shared space, we introduce a novel Disentangled Alignment Constraint, which combines two regularization terms detailed below.

Orthogonality for Disentanglement To promote statistical independence between modality-specific and shared latent representations, we introduce an orthogonality constraint on the latent space, a technique demonstrated to be effective in structured representation learning (Bousmalis et al. 2016). Specifically, we enforce uncorrelatedness between the latent variables Z_v , Z_l , and Z_s by first applying a whitening transformation to each, resulting in zero-mean, unit-covariance representations denoted as $(\tilde{Z}_v, \tilde{Z}_l, \tilde{Z}_s)$. This is implemented via a batch normalization layer applied to each latent subspace. The orthogonality loss is then formulated as the sum of squared Frobenius norms of the pairwise cross-covariance matrices:

$$\mathcal{L}_{\text{orth}} = \|\tilde{Z}_s^\top \tilde{Z}_v\|_F^2 + \|\tilde{Z}_s^\top \tilde{Z}_l\|_F^2 + \|\tilde{Z}_v^\top \tilde{Z}_l\|_F^2 \quad (5)$$

Minimizing $\mathcal{L}_{\text{orth}}$ penalizes any statistical correlation between the latent subspaces, thereby encouraging disentanglement. This uncorrelation, when combined with whitening, approximates statistical independence under the assumption of non-Gaussianity, a core principle underlying Independent Component Analysis (ICA) (Hyvärinen, Hurri, and Hoyer 2001).

Contrastive Alignment of the Shared Space While orthogonality promotes statistical independence, it does not inherently guarantee the semantic relevance of the shared representation Z_s . To address this, we introduce a contrastive alignment objective based on the InfoNCE loss (Rusak et al. 2024), which aligns Z_s with the modality-specific latents Z_v and Z_l . This objective encourages Z_s to exhibit higher similarity with its corresponding modality-specific latent while

treating the other as a negative sample. Formally, the alignment loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{align}} = & -\mathbb{E}_{q(Z_v, Z_s)} \left[\log \frac{\exp(\text{sim}(Z_s, Z_v)/\tau)}{\sum_{Z' \in \{Z_v, Z_l\}} \exp(\text{sim}(Z_s, Z')/\tau)} \right] \\ & - \mathbb{E}_{q(Z_l, Z_s)} \left[\log \frac{\exp(\text{sim}(Z_s, Z_l)/\tau)}{\sum_{Z' \in \{Z_v, Z_l\}} \exp(\text{sim}(Z_s, Z')/\tau)} \right] \quad (6) \end{aligned}$$

where $\text{sim}(\cdot)$ denotes cosine similarity, and τ is a temperature parameter. This formulation ensures that Z_s remains semantically coherent with both modalities. From an information-theoretic perspective, minimizing $\mathcal{L}_{\text{align}}$ effectively maximizes the mutual information between the shared and specific latents ($I(Z_s; Z_v)$ and $I(Z_s; Z_l)$), ensuring that the shared latent Z_s captures semantic information common to both modalities (Poole et al. 2019).

When combined, the orthogonality and alignment objectives enable the model to learn latent spaces that are both statistically disentangled and semantically rich. This dual constraint is crucial for improving the model’s generalization, robustness, and interpretability in multi-modal settings.

LLaMA-X Decoder

The final report is generated by the LLaMA-X Decoder, which is trained to model the dependencies between the report text and the fused multi-modal representations from the preceding modules. The entire DiA framework is optimized end-to-end with a composite loss function.

The LLaMA-X Decoder is a compact adaptation of LLaMA (Touvron et al. 2023). It uses a GPT-derived Cross-Attention (Brown 2020) to condition the report generation on the fused multi-modal representations from both the Modality Abstractor (F_{VL}) and VL-MoE-VAE (Z_v, Z_l, Z_s). The architecture incorporates several optimizations for efficiency and performance: (1) Rotary Positional Encodings (RoPE) which embed relative positional information via rotation matrices in the query and key vectors to efficiently handle long sequence lengths; (2) Grouped Query Attention which partitions queries into groups and leverages Key-Value (KV) caching to minimize redundant computations during inference; (3) SwiGLU Feed-Forward Network (FFN) that is defined as $\text{SwiGLU}(x) = (xW_1) \odot \sigma(xW_2)W_3$, with SiLU activation $\sigma(\cdot)$ to enhance feature transformation and mitigate the vanishing gradient problem; (4) RMS Pre-Normalization that is defined as $x' = x/\sqrt{\text{mean}(x^2) + \epsilon}$ to stabilize the inputs to the attention and feed-forward layers.

The decoder is trained by optimizing a standard cross-entropy loss, $\mathcal{L}_{\text{CE}} = -\sum_{i=1}^N \sum_{j=1}^T r_{ij} \log(\hat{r}_{ij})$ to align predicted reports \hat{r} with ground-truth r over T tokens. The overall objective for the DiA framework integrates this generation loss with previously defined objectives for the VL-MoE-VAE and the Disentangled Alignment Constraint. The total loss is a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{ELBO}} + \lambda_1 \mathcal{L}_{\text{orth}} + \lambda_2 \mathcal{L}_{\text{align}}, \quad (7)$$

where λ_1 and λ_2 are hyperparameters that balance the contributions of the orthogonality and alignment losses, respectively. This composite objective ensures that the model learns to generate accurate reports while maintaining a robust, disentangled latent structure.

Inference with Missing Context

A key advantage of the DiA framework is its inherent robustness to missing modalities, a common scenario in clinical workflows where the image V is present but the clinical context L may be absent. This resilience is a direct consequence of using a Mixture-of-Experts (MoE) posterior to infer the shared latent Z_s . At inference time, if a modality L is unavailable, a designated “null” token is passed to corresponding expert. As the MoE router was exposed to the same token during training, it learns to down-weight the unavailable modality automatically, i.e. $\pi_L \approx 0$ and $\pi_V \approx 1$ in Eq. (3). This allows the posterior to gracefully reduce to being conditioned only on the available data $q_{\phi_s}(Z_s | V)$ without requiring any imputation or architectural changes.

This process is theoretically sound. By substituting the reduced posterior into the training objective in eq. (4) and discarding terms involving the missing modality L , the objective becomes a marginal ELBO.

$$\begin{aligned} \mathcal{L}_{\text{ELBO}}^{(V)} = & \mathbb{E}_{q_{\phi_v}(Z_v | V)} [\log p_{\theta_v}(V | Z_v)] \\ & - D_{\text{KL}}(q_{\phi_v}(Z_v | V) \| p(Z_v)) \\ & - \text{JSD}(q_{\phi_s}(Z_s | V), p(Z_s)) \quad (8) \end{aligned}$$

This new objective $\mathcal{L}_{\text{ELBO}}^{(V)}$ remains a valid lower bound on the marginal log-likelihood of the observed data ($\mathcal{L}_{\text{ELBO}}^{(V)} \leq \log p_{\theta}(V)$), ensuring the learning procedure is principled for any subset of modalities.

The model’s effective performance in this scenario stems from the contrastive alignment term $\mathcal{L}_{\text{align}}$ applied during training. By maximizing the mutual information between the shared latent and each specific modality $I(Z_s; Z_v)$ and $I(Z_s; Z_l)$, the shared latent Z_s learns to encode salient cross-modal semantics. Consequently, even when inferred from a single modality, Z_s still provides the LLaMA-X decoder with sufficient information to generate clinically faithful reports, leading to a graceful degradation in performance rather than a catastrophic failure.

Experiments

Experimental Settings

Datasets and Preprocessing We evaluate DiA on two standard radiology report generation benchmarks: IU X-Ray (Demner-Fushman et al. 2016) and MIMIC-CXR (Johnson et al. 2019), both comprising paired chest X-ray images, free-text reports, and structured clinical metadata, enabling assessment under both complete and missing modality conditions.

IU X-Ray, consists of 7,470 frontal-view X-ray images and 3,955 reports. We adopt a 70%/10%/20% train/validation/test split and use a 1,000 word vocabulary. Approximately 2% of the test samples in this dataset are missing

Type	Model	IU X-Ray				MIMIC-CXR			
		B@1	B@4	R-L	F ₁	B@1	B@4	R-L	F ₁
Img	R2Gen (Chen et al. 2020)	0.470	0.165	0.371	-	0.353	0.103	0.277	-
	CvT2Dis (Nicolson et al. 2023)	0.473	0.175	0.376	-	0.392	0.127	0.285	0.384
KG	METransformer (Wang et al. 2023)	0.483	0.172	0.380	-	0.386	0.124	0.291	0.311
	Clinical BERT (Yan and Pei 2022)	0.495	0.170	0.376	-	0.383	0.106	0.275	0.415
	M2KT (Yang et al. 2023)	0.497	0.174	0.399	-	0.386	0.111	0.274	0.352
	MKSG (Yang et al. 2022)	0.496	0.178	0.381	-	0.363	0.115	0.284	0.371
CA	XProNet (Wang et al. 2022)	0.525	0.199	0.411	-	0.344	0.105	0.279	-
	PromptMRG (Jin et al. 2024)	0.401	0.098	0.281	0.211	0.398	0.112	0.268	0.476
	KiUT (Huang et al. 2023)	0.525	0.185	0.409	-	0.393	0.113	0.285	0.321
	EKAGen (Bu et al. 2024)	0.526	0.203	0.404	-	0.411	0.119	0.217	0.499
Ours	SEI (Liu et al. 2024)	-	-	-	-	0.382	0.135	0.299	0.460
	DiA	0.616	0.266	0.516	0.298	0.415	0.134	0.369	0.497

Table 1: Performance comparison of our proposed DiA with state-of-the-art models on the IU X-Ray and MIMIC-CXR datasets, reporting NLG and CE metrics; Methods grouped as Image (Img), Knowledge-Guided (KG), & Context-Aware (CA).

Context	Baseline	VL-MoE-VAE	DA	IU X-Ray				MIMIC-CXR			
				B@1	B@4	R-L	F ₁	B@1	B@4	R-L	F ₁
✓	✓	✗	✗	0.602	0.262	0.435	0.358	0.386	0.114	0.260	0.446
	✓	✓	✗	0.655	0.319	0.548	0.381	0.423	0.140	0.343	0.551
	✓	✓	✓	0.691	0.357	0.624	0.396	0.447	0.158	0.399	0.621
✗	✓	✗	✗	0.276	0.079	0.185	0.166	0.295	0.049	0.176	0.219
	✓	✓	✗	0.365	0.174	0.374	0.204	0.356	0.093	0.315	0.394
	✓	✓	✓	0.387	0.198	0.421	0.213	0.371	0.104	0.350	0.438

Table 2: Ablation Study: Incremental effects of VL-MoE-VAE (\mathcal{L}_{ELBO}) and Disentangled Alignment (DA) ($\mathcal{L}_{orth} + \mathcal{L}_{align}$) across with-context (✓) and missing-context (✗) scenarios

clinical context, providing a controlled setting to test for modality resilience. **MIMIC-CXR** is a much larger dataset with 473,057 images and 206,563 reports across 64,588 patients. We use the official split from (Chen et al. 2020), comprising 270,790 training, 2,130 validation, and 3,858 test samples. Reports are tokenized, lower-cased, and filtered to remove non-alphabetic tokens; words appearing < 4 are discarded, resulting in a vocabulary of 4,000 tokens. This dataset presents a more significant challenge for model robustness, as approximately 45% of its test samples have missing clinical indications.

Implementation and Training Details DiA was implemented in PyTorch and trained for 25 epochs on an NVIDIA A40 GPU using the AdamW optimizer (Loshchilov and Hutter 2017) with a learning rate of $1e-4$ and a weight decay of $1e-5$. We used a batch size of 4 and set the maximum report length of 50 words. The model’s compact architecture is defined by an embedding dimension E of 1024, a latent dimension for (Z_v, Z_l, Z_s) of 256, 6 Transformer encoder-decoder layers, 8 attention heads, and 2 key-value (KV) heads. A dropout rate of 0.1 was used to mitigate overfitting, while the loss term coefficients were set to $\lambda_1, \lambda_2 = 0.3$. These values were determined empirically from a search

range of 0.1 to 0.5. To ensure consistent results, the Transformer’s weight initialization was controlled by setting a random seed. We assess model performance using natural language generation (NLG) metrics including BLEU (Papineni et al. 2002) and ROUGE-L (Lin 2004), and a clinical efficacy (CE) metric such as F₁ score. Following (Nicolson et al. 2023), the F₁ score is calculated by converting the generated reports into 14 disease classification labels using the CheXbert labeler (Smit et al. 2020).

Evaluation

Comparison with State-of-the-Art Methods As shown in Table 1, DiA demonstrates superior performance compared to state-of-the-art (SOTA) methods on both IU X-Ray and MIMIC-CXR datasets. The evaluation spans Image-specific (Img), Knowledge-Guided (KG), and Context-Aware (CA) approaches, with DiA excelling in both natural language generation (NLG) and clinical efficacy (CE) metrics. On IU X-Ray, DiA achieves a BLEU@4 score of 0.266, surpassing the best KG model (XProNet) by 0.067, while an F₁ score of 0.298, outperforming the best CA model (PromptMRG) by 0.087. On the more challenging MIMIC-CXR dataset, DiA’s performance is highly competitive;

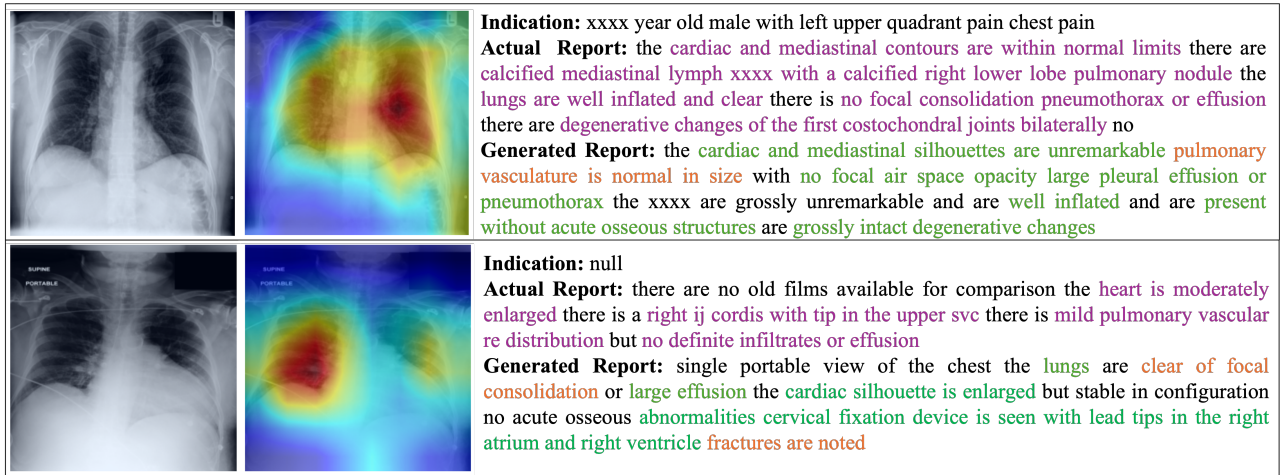


Figure 2: Comparison of actual and generated reports with chest X-rays and attention maps. Purple highlights key findings in the actual report, green indicates matched findings in the report, and amber marks mismatches / additional generated findings.

while SEI shows a marginal lead in BLEU@4 (0.135 vs. 0.134), DiA’s higher ROUGE-L score indicates enhanced report coherence. Its F_1 score of 0.497 nearly matches the top performer, EKAGen (0.499). These results highlight DiA’s adept integration of vision-language contexts, surpassing advanced CA methods that struggle with longer reports.

Ablation Study: Impact of Core Components Table 2 presents an ablation study quantifying the impact of DiA’s core components, the VL-MoE-VAE and the Disentangled Alignment (DA) constraint—under both complete (✓) and missing (✗) context scenarios. When clinical context is available, adding the VL-MoE-VAE to the baseline significantly boosts performance, improving the F_1 score on MIMIC-CXR by 0.105 and the BLEU@4 on IU X-ray by 0.057, which demonstrates the benefit of modeling a shared latent structure. Incorporating the DA constraint ($\mathcal{L}_{\text{orth}} + \mathcal{L}_{\text{align}}$) further enhances performance, with full DiA model achieving the highest scores across all metrics (e.g., MIMIC-CXR: F_1 0.621, ROUGE-L 0.399).

Under missing context, DiA shows remarkable resilience. While the baseline’s F_1 score on MIMIC-CXR drops by 0.227, with image only input, while DiA drops by only 0.183, outperforming the baseline by a margin of +0.219 in this challenging setting. The resilience is also evident on IU X-Ray, where DiA’s BLEU@4 remains more than $2\times$ higher than baseline’s (0.198 vs. 0.079). Comparing the start-to-end gains on MIMIC-CXR, the full DiA model improves over the baseline by 0.175 on the F_1 score with context and by 0.219 without context, demonstrating even greater relative benefit in the challenging incomplete-input scenario.

These findings confirm that DiA’s latent structure effectively infers missing semantics, establishing DiA as a modality-resilient, high-performance report generator.

Analysis of Architectural Choices and Efficiency Table 3 summarizes a comparison of encoder and decoder variants on MIMIC-CXR to validate DiA’s architectural design. For **encoder** variants, we compared DiA’s custom fea-

Variant	Params	FLOPs	B@4	F_1
RAD+CXR-BERT	568.7	81.1	0.121	0.441
Transformer	591.2	80.6	0.126	0.479
GPT-2	746.9	86.4	0.116	0.419
DiA LLaMA-X	589.7	51.1	0.134	0.497

Table 3: Comparison of encoder-decoder variants on MIMIC-CXR. RAD-DINO + CXR-BERT replaces DiA’s custom feature extractor and latent encoder; decoder across all variants is LLaMA-X unless otherwise noted.

ture extraction pipeline against a pre-trained RAD-DINO + CXR-BERT setup. (Perez-Garcia et al. 2025; Boecking et al. 2022) Despite using powerful pre-trained models, the RAD-DINO + CXR-BERT configuration achieved lower performance (BLEU@4 = 0.121, F_1 = 0.441) and incurred higher computational cost (81.1 GFLOPs). DiA’s end-to-end learned encoder proved more effective and efficient (BLEU@4 = 0.134, F_1 = 0.497 at 51.1 GFLOPs). For **decoder** variants, the LLaMA-X architecture outperformed standard Transformer (BLEU@4 = 0.126, F_1 = 0.479 at 80.6 GFLOPs) and GPT-2 decoders (BLEU@4 = 0.116, F_1 = 0.419 at 86.4 GFLOPs) in both accuracy and efficiency. These results demonstrate that DiA’s lightweight yet expressive components offer superior performance-to-cost trade-off. DiA’s efficiency is demonstrated by its training and inference times on an NVIDIA A40 GPU. Training on IU X-Ray takes 2.8 hours, with a 0.15-second inference time. For the larger MIMIC-CXR dataset, training takes 79.8 hours with a 0.18-second inference time. With 589.7M parameters and a computational cost of 51.14 GFLOPs, DiA maintains consistent computational efficiency.

Qualitative Visual Inspection As shown in Figure 2, visual inspection of the model’s attention maps reinforces its strengths. The heatmaps highlight that DiA focuses on key

clinical regions in the chest X-rays, both with and without the presence of clinical context in the input. The strong alignment between the generated reports and the ground-truth reports underscores the effective synergy of all of DiA's components.

Conclusion

This research introduces DiA, a cutting-edge framework that advances radiology report generation by effectively integrating medical scans with real-time clinical indications. The core of DiA is its ability to disentangle and align modality-specific and shared latent representations, enabling the generation of coherent reports even with incomplete context. As a result, DiA outperforms state-of-the-art methods on the IU X-Ray and MIMIC-CXR datasets. This proven robustness in handling missing data underscores DiA's potential to enhance diagnostic accuracy and support radiologists in real-world clinical scenarios. Overall, DiA significantly advances automating radiology reporting, promising to improve efficiency and reliability of medical imaging workflows.

References

- Boecking, B.; Usuyama, N.; Bannur, S.; et al. 2022. Making the most of text semantics to improve biomedical vision-language processing. In *European conference on computer vision*, 1–21. Springer.
- Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 343–351.
- Braman, N.; Gordon, J. W. H.; Goossens, E. T.; Willis, C.; et al. 2021. Deep Orthogonal Fusion: Multimodal Prognostic Biomarker Discovery Integrating Radiology, Pathology, Genomic, and Clinical Data. In *Medical Image Computing and Computer Assisted Intervention*, 667–677.
- Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Bu, S.; Li, T.; Yang, Y.; and Dai, Z. 2024. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14194–14204.
- Cheerla, A.; and Gevaert, O. 2019. Deep learning with multimodal representation for pan-cancer prognosis prediction. *Bioinformatics*, 35(14): i446–i454.
- Chen, C.; Dou, Q.; Jin, Y.; et al. 2019. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *Medical Image Computing and Computer Assisted Intervention*, 447–456. Springer.
- Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.
- Cherukuri, T. K.; Shaik, N. S.; and Ye, D. H. 2024. Guided Context Gating: Learning to Leverage Salient Lesions in Retinal Fundus Images. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE.
- Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; et al. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.
- Hayat, N.; Geras, K. J.; and Shamout, F. E. 2022. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182, 479–503. PMLR.
- Huang, S.-C.; Pareek, A.; Seyyedi, S.; et al. 2020a. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 3(1): 136.
- Huang, S.-C.; Pareek, A.; Zamanian, R.; Banerjee, I.; and Lungren, M. P. 2020b. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific Reports*, 10(1): 22147.
- Huang, Z.; et al. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19809–19818.
- Hyvärinen, A.; Hurri, J.; and Hoyer, P. O. 2001. Independent component analysis. In *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, 151–175. Springer.
- Jin, H.; Che, H.; Lin, Y.; and Chen, H. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2607–2615.
- Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Kline, A.; Wang, H.; Li, Y.; Dennis, S.; Hutch, M.; Xu, Z.; Wang, F.; Cheng, F.; and Luo, Y. 2022. Multimodal machine learning in precision health: A scoping review. *NPJ Digital Medicine*, 5(1): 171.
- Li, L.; Ding, W.; Huang, L.; Zhuang, X.; and Grau, V. 2023a. Multi-modality cardiac image computing: A survey. *Medical Image Analysis*, 85: 102869.
- Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023b. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3334–3343.
- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
- Liu, C.; Huang, Z.; Chen, Z.; Tang, F.; Tian, Y.; Xu, Z.; Luo, Z.; Zheng, Y.; and Meng, Y. 2025. Incomplete Modality Disentangled Representation for Ophthalmic Disease Grading and Diagnosis. *arXiv preprint arXiv:2502.11724*.
- Liu, D.; and Liu, G. 2019. A transformer-based variational autoencoder for sentence generation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.

- Liu, K.; Ma, Z.; Kang, X.; et al. 2024. Structural Entities Extraction and Patient Indications Incorporation for Chest X-Ray Report Generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 433–443. Springer.
- Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2302–2310.
- Mao, Y.; Zhang, J.; Xiang, M.; Zhong, Y.; and Dai, Y. 2023. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 954–965.
- Menéndez, M. L.; Pardo, J. A.; Pardo, L.; and Pardo, M. d. C. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2): 307–318.
- Minka, T.; et al. 2005. Divergence measures and message passing. *Technical Report MSR-TR-2005-173*.
- Mohsen, F.; Ali, H.; El Hajj, N.; and Shah, Z. 2022. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports*, 12(1): 17981.
- Nicolson, A.; et al. 2023. Improving chest X-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144: 102633.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.
- Perez-Garcia, F.; Sharma, H.; Bond-Taylor, S.; et al. 2025. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1): 119–130.
- Poole, B.; van den Oord, A.; Hjelm, R. D.; Maaløe, L.; Dhariwal, P.; Kingma, D. P.; and Alemi, A. A. 2019. Variational Inference with Mutual Information Constraints. *arXiv preprint arXiv:1907.00030*.
- Robinet, L.; Berjaoui, A.; Kheil, Z.; and Cohen-Jonathan Moyal, E. 2024. DRIM: Learning Disentangled Representations from Incomplete Multimodal Healthcare Data. *arXiv preprint arXiv:2409.17055*.
- Rusak, E.; Reizinger, P.; Juhos, A.; Bringmann, O.; Zimmermann, R. S.; and Brendel, W. 2024. InfoNCE: Identifying the Gap Between Theory and Practice. *arXiv preprint arXiv:2407.00143*.
- Sanchez, E. H.; Serrurier, M.; and Ortner, M. 2020. Learning Disentangled Representations via Mutual Information Estimation. In *Computer Vision – ECCV 2020*, 205–221.
- Sharma, A.; and Hamarneh, G. 2019. Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. In *IEEE Transactions on Medical Imaging*, volume 39, 1170–1183.
- Shen, Y.; and Gao, M. 2019. Brain tumor segmentation on MRI with missing modalities. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019*, 417–428. Springer.
- Shi, Y.; Paige, B.; et al. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems*, 32.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; et al. 2020. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 1500–1519. Association for Computational Linguistics.
- Steyaert, S.; Qiu, Y. L.; Zheng, Y.; Mukherjee, P.; Vogel, H.; and Gevaert, O. 2023. Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. In *Communications Medicine*, volume 3, 1–15.
- Sutter, T.; Daunhawer, I.; and Vogt, J. 2020. Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in neural information processing systems*, 33: 6100–6110.
- Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Venugopalan, J.; Tong, L.; Hassanzadeh, H. R.; and Wang, M. D. 2021. Multimodal deep learning models for early detection of Alzheimer’s disease stage. *Scientific reports*, 11(1): 3254.
- Wang, J.; et al. 2022. Cross-modal prototype driven network for radiology report generation. In *Computer Vision – ECCV 2022*, 563–579. Springer.
- Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023. Me-transformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11558–11567.
- Wu, M.; and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31.
- Yan, B.; and Pei, M. 2022. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2982–2990.
- Yang, S.; Wu, X.; Ge, S.; Zheng, Z.; Zhou, S. K.; and Xiao, L. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86: 102798.
- Yang, S.; Wu, X.; Ge, S.; Zhou, S. K.; and Xiao, L. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80: 102510.

Yao, W.; Yin, K.; Cheung, W. K.; et al. 2024. DrFuse: Learning Disentangled Representation for Clinical Multi-Modal Fusion with Missing Modality and Modal Inconsistency. In *The Thirty-Eighth AAAI Conference on Artificial Intelligence*.

Yoo, Y.; Tang, L. Y.; Li, D. K.; Metz, L.; et al. 2019. Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. In *Computer Methods in Biomechanics and Biomedical Engineering*, volume 7, 250–259.

Yu, T.; Lu, W.; Yang, Y.; Han, W.; Huang, Q.; Yu, J.; and Zhang, K. 2025. Adapter-Enhanced Hierarchical Cross-Modal Pre-training for Lightweight Medical Report Generation. *IEEE Journal of Biomedical and Health Informatics*.

Zhang, C.; Chu, X.; Ma, L.; et al. 2022. M3Care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2418–2428.