Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

QSGT: Quantum-Inspired Sparse Gated Transformer for Faithful Multi-modal Radiology Report Summarization

Somenath Nag Choudhury
Indian Institute of Technology Patna
Anup Kumar Kolya
Maulana Abul Kalam Azad University of Technology, West Bengal
Amit Kumar Verma
Indian Institute of Technology Patna
Asif Ekbal
Indian Institute of Technology Patna
Samarjit Kar
samarjit.kar@maths.nitdgp.ac.in
National Institute of Technology Durgapur

Research Article

Keywords:

Posted Date: October 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9291473/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

QSGT: **Q**uantum-Inspired **S**parse **G**ated **T**ransformer for Faithful Multi-modal Radiology Report Summarization

Somenath Nag Choudhury[a], Anup Kumar Kolya[b], Amit Kumar Verma[c], Asif Ekbal[d] and Samarjit Kar[e,*]

[a]*Department of Computer Science & Engineering, IIT Patna, Patna, 801106, India*
[b]*Department of Computer Science & Engineering, RCCIIT, Kolkata, 700015, India*
[c]*Department of Civil & Environmental Engineering, IIT Patna, Patna, 801106, India*
[d]*Department of Mathematics, NIT Durgapur, Durgapur, 713209, India*

ARTICLE INFO

Keywords:
Multimodal Radiology Report Summarization
Faithful Summarization
Quantum Token Pruning
Doubly Occupied Configuration Interaction (DOCI)

ABSTRACT

Improving the alignment between visual evidence and textual descriptions in medical vision-language models (VLM) remains a critical challenge for achieving entity-level faithfulness in low-resource generation tasks. This paper proposes Quantum-Inspired Sparse Gated Transformer (QSGT), a hybrid approach of token pruning and mixture of experts (MoE), grounded on the conception of quantum many-body wavefunction. This hybrid approach was aimed at improving the selection of both modality-specific expert tokens and the formation of non-dominating experts to determine stable conjugates. The token pruning approach was improved using the quantum orbital style token assignments and the mixture of experts by the quantum doubly occupancy approximation. We also introduce two clinically motivated regularization terms, pair loss and seniority loss, further enforce semantic consistency and dominance of diagnostically relevant features. Our extensive experiments across seven thoracic disease categories demonstrate that QSGT significantly outperforms state-of-the-art medical VLMs on multiple evaluation metrics, including RadGraph (36.2), SPICE (42.68), and consistently high BERTScore (98), on the MIMIC-CXR test set. QSGT achieves substantial gains in clinical correctness and semantic accuracy, particularly for complex conditions such as pneumonia and pneumothorax. Ablation studies confirm the importance of each architectural component, while convergence analyses show stable and interpretable learning dynamics. These results indicate that QSGT provides a robust and clinically meaningful solution for multimodal medical report summarization, enabling more reliable, interpretable, and patient-centred clinical AI systems.

1. Introduction

The task of Multimodal Radiology Report Summarization (MRRS) is a sparse, closed-ended, and vision-conditioned text generation (Wang, Figueredo, Li, Zhang, Chen and Chen, 2025c), and there continues to exist unfaithfulness in attaining entity-level factuality in a low-resource environment (Xia, Chen, Tian, Gong, Hou, Xu, Wu, Fan, Zhou, Zhu et al., 2024). Although Transformers are theoretically capable of learning sparse token selection, explicit mechanisms remain essential in vision-language models to translate representational sparsity into computational efficiency, improve scalability to long multi-modal sequences, and support specialization across heterogeneous visual and linguistic tokens (Wang, Wei, Hsu and Lee, 2024). Even the Healthcare Foundation Models (HFMs) hinder widespread adoption, including issues with data quality and diversity, algorithmic reliability and adaptability, ethical concerns, and the computational infrastructure required for training and deployment, due to misleading tokens (He, Huang, Jiang, Nie, Wang, Wang and Chen, 2024). Thus, improving model robustness, enhancing clinical validation and interpretability, and fostering collaborative human-AI systems using some intelligent sparse mechanism, can be some promising future directions (Buess, Keicher, Navab, Maier and Tayebi Arasteh, 2025).

Sparsity in vision and language transformers is used specifically in two ways: i) **Sequence Level Sparsity or Token Pruning** (Touheed, Zubair, Sabir, Hassan, Butt, Riaz, Abdul and Ayub, 2024), which reduces the set of tokens, often yielding significant runtime and memory savings with minimal accuracy loss. If the saliency scores are: $s_1, s_2, \ldots, s_n$ for an input sequence of n tokens, token pruning selects a subset of the top-k most informative tokens, and the pruned token set can be formally defined as follows:

*Corresponding author
somenath_2121cs27@iitp.ac.in (S.N. Choudhury); anupkumar.kolya@rcciit.org.in (A.K. Kolya); akv@iitp.ac.in (A.K. Verma); asif@iitp.ac.in (A. Ekbal); skar.maths@nitdgp.ac.in (S. Kar)
ORCID(s):

$$S = \{\, i \,:\, s_i \text{ is among the top } k \,\}, \quad |S| = k, \tag{1}$$

$$\tilde{X} = \{\, x_i \,:\, i \in S \,\} \tag{2}$$

where $\tilde{X}$ contains only the k tokens with highest saliency scores. This reduces the effective sequence length for subsequent transformer processing, thereby lowering computational cost (e.g., self-attention FLOPs) with minimal impact on performance, and ii) **Parameter Level Sparsity or Mixture-of-Experts** (Cai, Jiang, Wang, Tang, Kim and Huang, 2025), that keeps the cost per token roughly constant even as total model parameters grow, because only a few experts are used. This allows huge capacity without proportionally increasing inference cost. Let $\mathcal{E} = \{1, \ldots, N\}$ be a set of N experts. For a given input x, the gating network computes routing weights $g_e(x)$ for each expert $e \in \mathcal{E}$. Only a small active subset $\mathcal{E}_{\text{active}} \subseteq \mathcal{E}$ is used for computation, typically selected based on the top-k gating scores. The MoE output is defined as

$$\text{MoE}(x) = \sum_{e \in \mathcal{E}_{\text{active}}} g_e(x) \cdot f_e(x) \tag{3}$$

where $f_e(\cdot)$ denotes the e-th expert's transformation and $g_e(x)$ is its associated gating weight. Typically, the cardinality of the active set satisfies (Eq.4), so that only a small number of experts are activated per input, allowing a model with a large total parameter count to maintain efficient per-token computation.

$$|\mathcal{E}_{\text{active}}| = k \ll N \tag{4}$$

Just as quantum many-body wavefunction theory, that models interacting particles beyond independent approximations, correlation-aware vision-language systems explicitly construct joint multi-modal representations to capture fine-grained cross-modal dependencies that are lost in mean-field encoding. Analogous to Hartree-Fock, most of the V-L system processes the vision and language independently (Eq.5), and the interaction is added late(Eq.6).

$$P(V, L) \simeq P(V)\, P(L) \tag{5}$$

$$\mathbf{Z}_{\text{joint}} = \sum_I c_I\, \mathbf{Z}_I^{(V,L)} \tag{6}$$

Thus, visualizing the text and vision sub-systems as two atoms with their own distribution (layer of importance) of tokens (electrons), possibly not fixed, like in a quantum orbital, and their cross-modal interaction (configuration interaction) to form a vision-language model (a system of quantum particles). Tokens that persist (valence shell electrons) across the pruning iteration and participate in multi-modal fusion may disappear and reappear multiple times in any orbital (concentric layers of importance) during this process(quantum refinement). The Configuration Interaction (CI) in quantum theory is a wavefunction-based electronic structure method in which the exact electronic wavefunction is approximated as a linear combination of Slater determinants (configurations).

$$|\Psi_{\text{CI}}\rangle = c_0|\Phi_0\rangle + \sum_i c_i|\Phi_i\rangle + \sum_{i<j} c_{ij}|\Phi_{ij}\rangle + \sum_{i<j<k} c_{ijk}|\Phi_{ijk}\rangle + \cdots \tag{7}$$

where, $|\Phi_0\rangle$ is the Hartree-Fock reference determinant, $|\Phi_i\rangle$, $|\Phi_{ij}\rangle$, and $|\Phi_{ijk}\rangle$ denote singly, doubly, and triply excited determinants, respectively, and c_0, c_i, c_{ij}, … are variational coefficients. Whereas the Doubly Occupied Configuration Interaction (DOCI) is a constrained form of configuration interaction in which the CI expansion is limited to seniority-zero determinants, retaining only those configurations where all electrons remain paired in doubly occupied orbitals, thereby efficiently capturing strong electron correlation while excluding unpaired excitations.

$$|\Psi_{\text{DOCI}}\rangle = \sum_{I \in S_0} c_I|\Phi_I\rangle \tag{8}$$

where S_0 denotes the set of all seniority-zero (doubly occupied) determinants, $|\Phi_I\rangle$ differs from $|\Phi_0\rangle$ only by pair excitations, and c_I are variationally optimized coefficients.

1.1. Our Contribution

We propose QSGT (Quantum-Inspired Sparse Gated Transformer), a hybrid approach of token pruning and mixture of experts, to perform the modality-specific token refinement and gated cross-modal interaction, respectively, based on quantum principles(Fig. 1). We have used a progressive and adaptive token pruning, individually for both text and vision, followed by an intelligent cross-modal configuration interaction strategy. We can summarize our contributions as follows:

- **Orbital Style Unimodal Refinement:** This is an iterative method of progressively finding the sparse set of candidate unimodal tokens, participating in cross-modal interaction. Unlike the classical approach, the orbit assignment mechanism is iteration-dependent and emergent, not static, similar to quantum orbitals. Also, the subset ensures a differentiable selection based on Gumbel Top-K.
- **DOCI Inspired Multi-modal Fusion:** Originated from quantum many-body wavefunction theory, this DOCI (Doubly Occupied Configuration Interaction) is an inspirational multi-modal fusion approach to improve cross-modal relevance. This quantum-inspired pairing constraints enforce only the coherent and co-activated fusion states. To encourage a paired co-activation of selected vision and text tokens, the seniority-zero(zero unpaired electrons) regularization, a cross-reinforcing modality feature (penalizes activations of unimodal latent dimensions) is used. It is then followed by the representative fused vector, a weighted superposition of multiple pair configurations (DOCI-style sparse gated configuration activations).

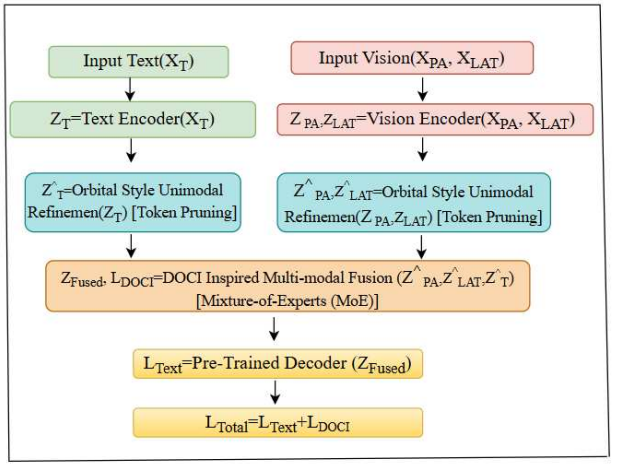


Figure 1: QSGT Workflow Diagram

2. Related Works

In this section, we have analyzed three advanced fusion mechanisms from recent vision-language research and demonstrated how they are adapted and unified in our QSGT architecture.

2.1. Token Pruning Approaches

Recent token pruning methods efficiently reduce computation in transformer architectures based on learned or attention-based scores. DynamicViT introduces a dynamic token sparsification framework that hierarchically prunes up to 66% of input tokens while maintaining less than 0.5% accuracy loss on image classification tasks (Rao, Zhao, Liu, Lu, Zhou and Hsieh, 2021). Learned Token Pruning (LTP) adapts thresholds during training to prune tokens without expensive top-k selection, leading to significant FLOPs reduction and improved throughput (Kim, Shen, Thorsley, Gholami, Hassoun and Keutzer, 2021). HeatViT further incorporates hardware-aware design to adaptively prune tokens for FPGA acceleration with substantial speedups (Dong, Sun, Lu, Xie, Liu, Kong, Meng, Li, Lin, Fang and Wang, 2022). Extensions of token pruning show practical utility in dense prediction settings like object detection

and instance segmentation, and provide better performance (Liu, Gehrig, Messikommer, Cannici and Scaramuzza, 2023). The background-aware token pruning improves throughput in detection pipelines (Sah, Kumar, Rohmetra and Saboori, 2024), and dynamic token pruning tailored for semantic segmentation achieves significant cost reduction with maintained accuracy (Tang, Zhang, Liu, Liu and Liu, 2023).

2.2. Mixture of Expert Approaches

In recent research, Mixture of Experts (MoE) architectures are extensively used for medical vision-language understanding, where multimodal fusion of clinical images and text poses unique challenges. For example, MedMoE introduces a modality-specialised MoE framework that dynamically routes visual features through expert branches conditioned on report type, improving alignment and retrieval across diverse medical imaging modalities such as X-ray and CT within a unified vision-language pipeline (Chopra, Mao, Sanchez-Rodriguez, Feola, Li and Kira, 2025; Jiang, Zheng, Zhang, Jin, Yuan and Liu, 2024). To handle heterogeneous clinical imaging tasks, MoME extends the MoE paradigm by integrating multi-scale visual and textual embeddings for robust medical image segmentation across large CT datasets (Izhar, Japar, Idris and Dang, 2025b; Rezvani, Yan, Wu, Han, Khosravi and Xie, 2025). In radiology report generation, DiA-gnostic VLVAE leverages a disentangled MoE-based vision-language VAE to address missing modality issues and enhance report accuracy by separating the shared and modality-specific features (Zhu, Hu, Mo, Li and Wu, 2024). Earlier work on Med-MoE shows that domain-specific expert routing can achieve competitive performance on medical visual question answering and classification tasks while reducing activated parameters (Huang, Shen, Liu, Shang, Li, Huang and Yang, 2024). Collectively, these studies highlight that MoE strategies contribute to more adaptive, efficient, and clinically interpretable multimodal models in medical AI.

2.3. Quantum Inspired VLM

The quantum approaches were introduced to implement novel vision-language architectures to capture the existing modality-specific uncertainty of candidate tokens and domain-independent representation approximations, and thus improve the multi-modal reasoning (Singh, Kumar, Ahuja and Barua, 2025; Xu, Chen, Chai, Xiao, Yan, Ding, Wang, Jin, Jiao, Yue et al., 2026; Izhar, Idris and Japar, 2025a). Assuming the different heterogeneous clinical vision modalities as the quantum states, the QNMF (Qu, Li and Tiwari, 2023) proposed a quantum-inspired fusion mechanism, based on quantum unitary transformations and measurement operations using a neural network to improve the cross-modal correlation. Experiments were conducted in a small-scale environment, and reported to improve the robustness through this compact representation. The quantum superposition principle can also be considered as a potential candidate for the joint multi-modal representation, and thus can be used to enhance cross-modal representation learning. QViLa (Mukesh, Jayaprakash and Kumar, 2024) successfully applied this at the token-level with an attention mechanism, and reported to achieve better cross-modal alignments for multi-modal understanding benchmarks. Because of the existence of multiple report-texts against the same vision label, modeling of cross-modal causal dynamics can also help in multi-modal reasoning. CDMRNet (Wang, Chen, Yu, Zhao and Duan, 2025b), introduces such a causal graph-based multi-modal interaction, optimized via co-evolving quantum states. This approach is reported to better capture the complex inter-class dependencies, with improved interpretability and grounding. Together, these works demonstrate effective use of quantum principles for robust multi-modal reasoning and faithfulness.

Other notable approaches, like prompt-guided (Huang, Guo, Bai, Zhao and Tao, 2025; Luo, Jiang, Wang, Cai, Gao and Yang, 2025), vision region or paired context aware (He, Wong and Yoo, 2025; Gao, Ni, Liu, Ni, Xin, Hu and Zhang, 2025; Li, Wang, Sun, He and Feng, 2025; Kommanduri, Matta and Imandi, 2025), pure and hybrid quantum approaches (Relan and Rambola, 2025; Sankaradass and Manish, 2026; Wang, Mao, Li, Li and Li, 2025a), multi-feature (Su, Song, Lin, Wu, Weng, Mao, Liu, Yin and Yang, 2025; Cao, Ding, Zhang and Hei, 2025; Lu, Zhu, Zhang and Yao, 2025; Zhou, Liu, Yang, Li and Xiao, 2026) has also contributed to improving the vision-language alignment and thus improving the multiple medical downstream tasks, but with limited or no focus on faithfulness.

3. Method

The input to our QSGT is two vision components, PA and LAT, and one text component, the radiology conception text. This conception text is a representative text and is being selected based on the existing entity priority in the findings section of the associated clinical report. In QSGT, the unimodal encoders return the tokens and the token attention. These are then used as the input of our first module, named Orbital Refinement, which refines the tokens and

produces a subset for each of them. The refined tokens are then passed to the next module, named DOCI-Inspired multi-modal fusion for cross-modal interaction. This then produces the expert mixture of embedding with their cumulative cross-modal losses (Fig. 2). The following sections provide a detailed discussion of the two modules.

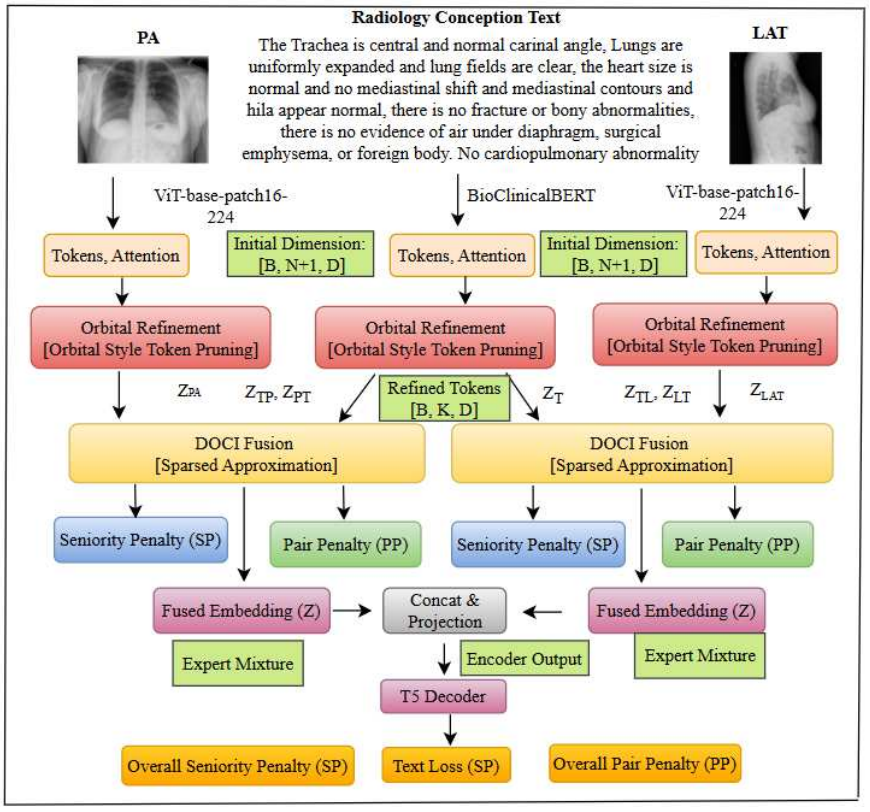


Figure 2: QSGT: Overall Architecture

3.1. Orbital Refinement: Attention-Guided Adaptive Token Pruning

This Orbital Refinement (Fig. 3) is an iterative sparsification algorithm that treats [CLS] as a nucleus, a semantic gravitational centre and tokens as electrons. The orbits collapse iteratively based on intrinsic energy (independent appearance) and nucleus interaction (contextual presence). Tokens exchange information with [CLS] via attention, compete via energy scoring, and are pruned through differentiable Top-K selection until the system reaches a stable configuration. Energy is reduced in our implementation by progressively shrinking the allowed token set and reoptimising saliency under attention constraints, rather than by explicitly changing a scalar energy value. This conceptually reflects the optimization goal of selecting a subset of tokens that is both low-energy (highly relevant) and low-uncertainty (highly confident/concentrated), and can be expressed as follows:

$$\min_{S\subseteq\{1,\dots,n\}} \quad \sum_{i\in S} E_t(i) \;-\; \tau\, H(S) \qquad\qquad \text{s.t.} \quad |S| \;\le\; k_t, \tag{9}$$

Here, $S \subseteq \{1, \dots, n\}$ means that S is a subset of the set of indices 1 through n, i.e., every element of S is an element of $\{1, \dots, n\}$, consistent with standard subset notation, $|S| \le k_t$ restricts the maximum number of selected elements to be at most k_t, where $|\cdot|$ denotes the cardinality (size) of a finite set, $\sum_{i\in S} E_t(i)$ denotes the sum of energy scores $E_t(i)$ over all elements i in the subset S, aggregating the scores of selected indices, and $H(S)$ is a set-based entropy or score term defined on the subset S, and it is weighted by the scalar τ in the objective to control the influence of uncertainty or spread in S.

The Orbital Refinement Algorithm described in Algo. 1 takes two inputs, token embeddings and the attention map, and produces a refined subset of tokens for the participating modalities, of the same energetic compatibility and not in

Algorithm 1 Orbital Refinement Algorithm

Require: Token embeddings $\mathbf{X} \in \mathbb{R}^{B\times N\times D}$
Require: Attention maps $\mathbf{A} \in \mathbb{R}^{B\times H\times N\times N}$
Ensure: Refined tokens $\mathbf{X}'$, mask history $\mathcal{M}$

Separate CLS and non-CLS tokens:
$(\mathbf{x}_{\text{CLS}}, \mathbf{X}^{(0)}) \leftarrow \mathbf{X}$
Compute CLS-relative attention saliency:
$\mathbf{S}^{\text{attn}} \leftarrow \text{mean}_h(\mathbf{A})_{0,1:} + \text{mean}_h(\mathbf{A})_{1:,0}$
Initialize iteration counter and history:
$t \leftarrow 0, \quad \hat{\mathbf{S}}^{(t-1)} \leftarrow \varnothing, \quad \mathcal{M} \leftarrow \varnothing$
while $t < T_{\max}$ **do**
  Compute energy-based saliency:
  $\mathbf{E}^{(t)} \leftarrow f_\theta(\mathbf{X}^{(t)})$
  Resize attention saliency to current token count:
  $\tilde{\mathbf{S}}^{\text{attn},(t)} \leftarrow \text{Interp}(\mathbf{S}^{\text{attn}}, |\mathbf{X}^{(t)}|)$
  Fuse energy and attention saliency:
  $\mathbf{S}^{(t)} \leftarrow \alpha\, \mathbf{E}^{(t)} + (1-\alpha)\, \tilde{\mathbf{S}}^{\text{attn},(t)}$
  Normalize saliency distribution:
  $\hat{\mathbf{S}}^{(t)} \leftarrow \mathbf{S}^{(t)} / (\text{std}(\mathbf{S}^{(t)}) + \varepsilon)$
  Determine adaptive token budget:
  $k_t \leftarrow \max\left(N_{\min}, \left\lfloor |\mathbf{X}^{(t)}| \gamma^t \right\rfloor\right)$
  Apply differentiable Gumbel-Top-k selection:
  $\mathbf{m}^{(t)} \sim \text{GumbelTopK}(\hat{\mathbf{S}}^{(t)}, k_t, \tau)$
  Prune token set:
  $\mathbf{X}^{(t+1)} \leftarrow \mathbf{X}^{(t)} \odot \mathbf{m}^{(t)}$
  Store selection mask:
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{\mathbf{m}^{(t)}\}$
  Check adaptive stopping criterion:
  if $\text{Stop}(\hat{\mathbf{S}}^{(t-1)}, \hat{\mathbf{S}}^{(t)})$ **then**
    break
  end if
  $t \leftarrow t + 1$
end while
Reattach CLS token:
$\mathbf{X}' \leftarrow [\mathbf{x}_{\text{CLS}}, \mathbf{X}^{(t)}]$
return $\mathbf{X}'$, $\mathcal{M}$

cardinality, to be used further by the multi-modal fusion module 3.2. Let,

$$\mathbf{X} \in \mathbb{R}^{B\times N\times D}$$

denote a batch of token embeddings produced by a unimodal Transformer encoder, where the first token corresponds to a special classification token (CLS), and

$$\mathbf{A} \in \mathbb{R}^{B\times H\times N\times N}$$

be the corresponding last-layer self-attention maps, the goal is to produce a refined token set

$$\mathbf{X}' \in \mathbb{R}^{B\times(1+N')\times D}, \quad N' \ll N,$$

such that the CLS token is preserved and only semantically stable non-CLS tokens remain.

3.1.1. Nucleus–Electron Decomposition

We explicitly separate the CLS token from the remaining tokens:

$$\mathbf{x}_{\text{cls}} = \mathbf{X}_{:,0,:}, \quad \mathbf{X}_{\text{tok}} = \mathbf{X}_{:,1:,:}. \tag{10}$$

The CLS token is treated as a fixed *nucleus*, while non-CLS tokens act as *electrons* subject to refinement. All pruning operations are applied exclusively to $\mathbf{X}_{\text{tok}}$.

3.1.2. Attention-Guided Saliency

For each non-CLS token, to capture their contextual importance, we computed a bidirectional CLS-relative (nucleus) token saliency relationally rather than independently, as follows:

$$S_i^{\text{attn}} = \frac{1}{H} \sum_{h=1}^{H} \left(A_{0,i}^{(h)} + A_{i,0}^{(h)} \right) \tag{11}$$

where $A_{0,i}^{(h)}$ denotes attention from CLS to token i, and $A_{i,0}^{(h)}$ denotes attention from token i to CLS.

3.1.3. Token Saliency

We have computed the token saliency based on two factors, its correspondence and its individuality. Correspondence score, already calculated in section 3.1.2, and thus we define the individuality of a token using a learnable energy function,

$$E(\mathbf{x}_i) = f_\theta(\mathbf{x}_i), \quad f_\theta : \mathbb{R}^D \to \mathbb{R} \tag{12}$$

which is a lightweight multilayer perceptron applied to non-CLS tokens, to generate a scaler score for each non-CLS token. Then, at iteration t, we have combined the scores for token i to compute its saliency as,

$$S_i^{(t)} = \alpha E(\mathbf{x}_i^{(t)}) + (1 - \alpha)\, \tilde{S}_i^{\text{attn}} \tag{13}$$

where $\tilde{S}_i^{\text{attn}}$ denotes the attention saliency interpolated to the current token count, and $\alpha \in [0, 1]$ balances intrinsic and contextual importance. The linear interpolation used can be defined as:

$$y[j] = x[i] \cdot (1 - \delta) + x[i + 1] \cdot \delta \tag{14}$$

where, y[i] is the interpolated value, and x[i] and x[i+1] are the adjacent values of original sequence. The scores are then normalized to stabilize the optimization as:

$$\hat{S}_i^{(t)} = \frac{S_i^{(t)}}{\text{std}(S^{(t)}) + \epsilon} \tag{15}$$

3.1.4. Differentiable Top-k Selection

To retain the most salient tokens while preserving gradient flow, we employ a Gumbel-Top-k estimator. For each token, we compute its temperature-controlled stochastic perturbation as:

$$\tilde{S}_i^{(t)} = \frac{\hat{S}_i^{(t)} + G_i}{\tau}, \quad G_i \sim \text{Gumbel}(0, 1) \tag{16}$$

followed by a probability distribution, a continuous relaxation of the categorical variable as:

$$\mathbf{p}^{(t)} = \text{softmax}(\tilde{\mathbf{S}}^{(t)}) \tag{17}$$

A hard top-k mask is extracted in the forward pass, while gradients flow through the soft probabilities via a straight-through estimator(STE). The token budget at iteration t is

$$k_t = \max\left(N_{\min}, \left\lfloor N_0 \cdot \gamma^t \right\rfloor\right) \tag{18}$$

where $N_{\min}$ enforces a minimum representational capacity and $\gamma \in (0, 1)$ controls contraction speed.

3.1.5. Adaptive Stopping Criterion

This refinement proceeds until the saliency distribution stabilizes. Let $\hat{S}^{(t)}$ and $\hat{S}^{(t-1)}$ denote normalized scores from consecutive iterations, then refinement stops if any of the following three conditions hold:

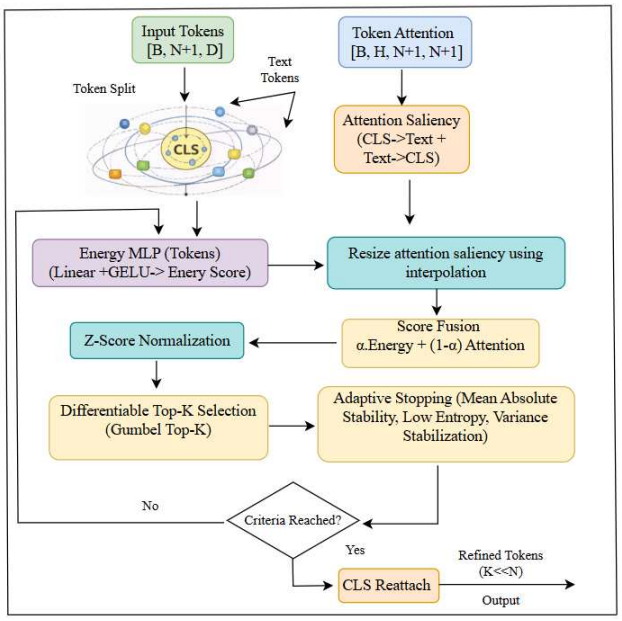


Figure 3: QSGT: Orbital Refinement Module

Distribution Stability: This measures the average absolute change in score values from the previous step. If the average change falls below a small threshold, the system isn't changing much and further refinement may be unnecessary, and can be represented as follows:

$$\mathbb{E}\left[\left|\hat{S}^{(t)} - \hat{S}^{(t-1)}\right|\right] < \varepsilon \tag{19}$$

Entropy Collapse: Entropy measures uncertainty, where a high entropy indicates a distribution is spread out (high uncertainty about which tokens matter), and Low entropy ensures the distribution is peaked (few tokens strongly distinguished). Since low entropy indicates high confidence in the current ranking of tokens, refinement can stop early.

$$H(\mathbf{p}^{(t)}) = -\sum_i p_i^{(t)} \log p_i^{(t)} < H_{\min} \tag{20}$$

Variance Stabilization: Variance reflects the spread of the normalized saliency scores. A small change means the overall distribution shape isn't evolving much, another indicator of stabilization. If it's below a threshold, then the refinement isn't making structural changes.

$$\left|\mathrm{Var}(\hat{S}^{(t)}) - \mathrm{Var}(\hat{S}^{(t-1)})\right| < \varepsilon \tag{21}$$

Thus, this adaptive termination ensures refinement halts once semantic probability density becomes invariant to further energy reduction.

3.2. DOCI Inspired Multi-modal Fusion

This module performs token-level multimodal fusion between the refined text and vision tokens, N_t and N_v, respectively, received from the orbital refinement module. It applies a sparse configuration-interaction formulation instead of standard cross-attention, as represented in Fig. 4. Here, the objective is to explicitly model fine-grained cross-modal token interactions while suppressing spurious pairings through energy-based configuration selection, aligned using pair and seniority constraints. Treating each token (text and vision) as an orbital, each pair forms a double-occupancy configuration. The ReLU-based pair penalty, which enforces orbital compatibility by penalizing mismatched activations between text and vision orbitals, and thereby favours cross-modal pairs that share aligned latent occupation patterns. Whereas the seniority-zero penalty suppresses one-sided cross-modal interactions by penalizing configurations where the joint occupancy gate exceeds the minimum individual modality occupancy, thereby

enforcing only the effective seniority-zero (paired) multi-modal configurations. Finally, the sparsity regularization, which promotes a compact configuration-interaction expansion by penalizing diffuse CI coefficients, ensures that only a small subset of energetically and structurally consistent cross-modal token pairs contribute to multi-modal fusion.

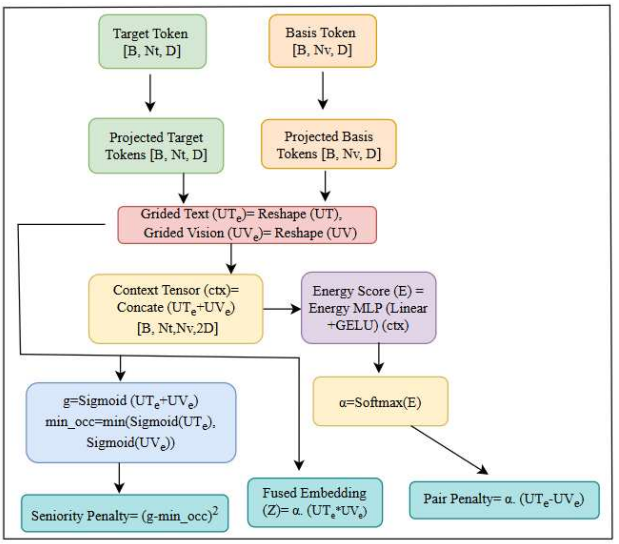


Figure 4: QSGT: DOCI Token Fusion Module

Given text and vision token embeddings $X_T \in \mathbb{R}^{B\times N_t\times d}$ and $X_V \in \mathbb{R}^{B\times N_v\times d}$, we first project them into a shared orbital space via $U_T = W_T X_T$ and $U_V = W_V X_V$. Although text and vision embeddings share the same dimensionality, they reside on distinct modality-specific manifolds with misaligned coordinates; thus, these learned linear projections map both modalities into a shared orbital basis, rendering element-wise pairing and DOCI constraints semantically meaningful. The ϕ_i and the C_i in the configuration interaction (CI) expression are the i-th configuration basis state and the i-th energy-based CI coefficient, respectively. Similarly, each token pair (t, v), we construct a DOCI pair configuration ($P_{t,v}$=ϕ_i) using element-wise double occupancy (configuration contribution), a configuration interaction (CI) energy ($E_{t,v}$), and the CI coefficients ($\alpha_{t,v}$=C_i) obtained by temperature-scaled normalized energy as follows:

$$\Psi = \sum_i c_i \, \Phi_i \tag{22}$$

$$P_{t,v} = \mathrm{LN}\Big(U_T^{(t)} \odot U_V^{(v)}\Big) \tag{23}$$

$$E_{t,v} = MLP\Big([U_T^{(t)}, U_V^{(v)}]\Big) \tag{24}$$

$$\alpha_{t,v} = \mathrm{softmax}\left(\frac{E_{t,v}}{\tau}\right) \tag{25}$$

Similar to Eq. 22, the final fused multi-modal representation can then be expressed as:

$$z = \sum_{t=1}^{N_t}\sum_{v=1}^{N_v} \alpha_{t,v} P_{t,v} \tag{26}$$

To enforce DOCI-consistent interactions, we introduce two quantum-inspired regularizers. The *pair penalty* (L_{pair}) constrains orbital compatibility by penalizing activation mismatch, while the *seniority penalty* (L_{sen}) enforces joint occupancy by discouraging single-modality excitation. The additional sparsity regularizer $\mathcal{L}_{\text{sparse}} = \lambda_{\text{sparse}}\|\alpha\|_1$, promotes only the compact CI expansions. The regularizations are represented as follows:

$$\mathcal{L}_{\text{pair}} = \sum_{t,v} \alpha_{t,v} \left\| \mathrm{ReLU}(U_T^{(t)}) - \mathrm{ReLU}(U_V^{(v)}) \right\|_1 \tag{27}$$

Algorithm 2 DOCI-Inspired Multimodal Fusion Algorithm

Require: Text tokens $X_T \in \mathbb{R}^{B\times N_t\times d}$; Vision tokens $X_V \in \mathbb{R}^{B\times N_v\times d}$;
Ensure: Fused multimodal representation $z \in \mathbb{R}^{B\times d}$; DOCI regularization loss $\mathcal{L}_{\text{DOCI}}$

Project tokens to orbital space
$U_T \leftarrow W_T X_T$ ▷ $U_T \in \mathbb{R}^{B\times N_t\times d}$
$U_V \leftarrow W_V X_V$ ▷ $U_V \in \mathbb{R}^{B\times N_v\times d}$

Enumerate DOCI pair configurations
for $t = 1$ to N_t **do**
  for $v = 1$ to N_v **do**
    $P_{t,v} \leftarrow \text{LN}(U_T^{(t)} \odot U_V^{(v)})$
    $E_{t,v} \leftarrow MLP([U_T^{(t)}, U_V^{(v)}])$
  end for
end for

Compute CI coefficients
$\alpha_{t,v} \leftarrow \text{softmax}\left(\frac{E_{t,v}}{\tau}\right)$
if k is specified **then**
  Retain top-k entries of α and set others to zero
end if

Fuse multimodal representation
$z \leftarrow \sum_{t=1}^{N_t}\sum_{v=1}^{N_v} \alpha_{t,v} P_{t,v}$
$z \leftarrow \text{LN}(W_o z)$

Compute pair penalty (orbital compatibility)
$\mathcal{L}_{\text{pair}} \leftarrow \sum_{t,v} \alpha_{t,v} \left\|\text{ReLU}(U_T^{(t)}) - \text{ReLU}(U_V^{(v)})\right\|_1$

Compute seniority penalty (double occupancy)
$\mathcal{L}_{\text{sen}} \leftarrow \sum_{t,v} \alpha_{t,v} \left\|\sigma(U_T^{(t)} + U_V^{(v)}) - \min\big(\sigma(U_T^{(t)}), \sigma(U_V^{(v)})\big)\right\|_2^2$

Enforce sparsity of CI expansion
$\mathcal{L}_{\text{sparse}} \leftarrow \|\alpha\|_1$

Aggregate DOCI regularization
$\mathcal{L}_{\text{DOCI}} \leftarrow \lambda_{\text{pair}}\mathcal{L}_{\text{pair}} + \lambda_{\text{sen}}\mathcal{L}_{\text{sen}} + \lambda_{\text{sparse}}\mathcal{L}_{\text{sparse}}$

return $z, \mathcal{L}_{\text{DOCI}}$

$$\mathcal{L}_{\text{sen}} = \sum_{t,v} \alpha_{t,v} \left\|\sigma(U_T^{(t)} + U_V^{(v)}) - \min\big(\sigma(U_T^{(t)}), \sigma(U_V^{(v)})\big)\right\|_2^2 \tag{28}$$

Thus, the total DOCI regularization term (L_{DOCI}), yielding an interpretable and physically motivated multi-modal fusion framework, can be represented as:

$$\mathcal{L}_{\text{DOCI}} = \lambda_{\text{pair}}\mathcal{L}_{\text{pair}} + \lambda_{\text{sen}}\mathcal{L}_{\text{sen}} + \mathcal{L}_{\text{sparse}} \tag{29}$$

4. Experiment

To experiment, we have used two publicly available datasets, Open-I and MIMIC-CXR, for training and testing, respectively. We have chosen the Kaggle platform to run and test our model. The description of the modified datasets and the platform specifications used is explained in detail in the following sections.

4.1. Dataset

The initial Open-I dataset comprised 3,955 text reports and 7,290 image files, with each report linked to one or more corresponding images. For our study, we focused specifically on reports associated with exactly two images, Posteroanterior (PA) and Lateral (LAT) views. After filtering, 3,127 such reports were identified, but only 2,704 contained all the necessary sections. Of these, 1,953 were classified as clear chest cases, 311 involved chest-related conditions (target entities), and 440 were categorized as others. From the 2,264 reports (1,953 clear + 311 disease-related) that included both findings and impressions, we generated a CSV file with 6,470 entries. This was achieved by replicating disease findings and their corresponding impression statements, using specific entities as keywords. The maximum lengths observed were 160 words for findings (source sequence) and 66 words for impressions (target sequence), with vocabulary sizes of 1,580 and 880 words for the source and target, respectively. Similar to the Open-I dataset, the MIMIC-CXR contains 227,835 patient reports associated with a total of 377,110 radiographs (PA+LAT), indicating more than one radiograph (in some cases, it is five). Also, the number of impression statements varies from one to nine. Since we have trained our model on reports with two radiographs, one PA and one LAT, for each target medical entity, we have considered the samples with the same input restrictions for MIMIC-CXR to test. Here also, we have curated the samples according to the target medical entities and tested the system to report the individual entity-wise performance. Each case was tested five times with 100 random samples, and the best result obtained is reported.

4.2. Implementation Details

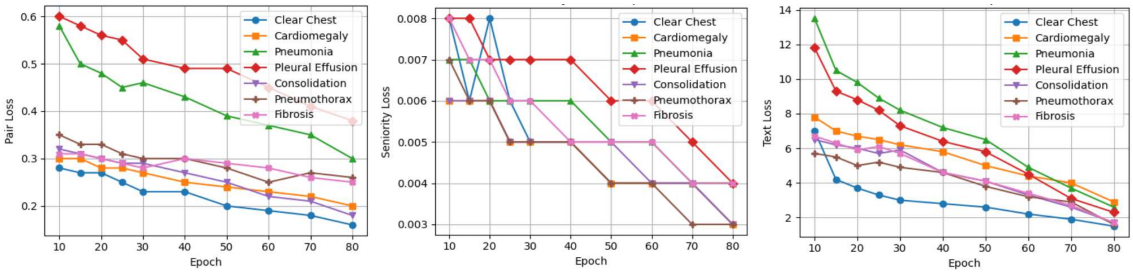
The proposed QSGT is a lightweight and unified model with approximately 220M trainable parameters, trained using task-invariant joint-space representation learning. For unimodal encoding, we have used publicly available pre-trained checkpoints, BioClinicalBERT() and ViT-base(), for text and vision, respectively, each with 12 transformers and 768 hidden state dimensions. Assuming the joint representation as the encoder output, we have used T5-small as the decoder for summary generation. The model was trained for 80 epochs for a batch size of 8. The training was guided by AdamW optimizer with learning rate 2×10^{-4}, weight decay 10^{-2}, and linear warm-up over the first 10% steps. Our orbital refinement module employs the following five hyperparameters: i) temperature $\tau = 0.3$, ii) saliency energy mixing coefficient $\alpha = 0.5$, iii) least four tokens to retain $N_{min} = 4$ per modality, iv) contraction rate per refinement step as 70%, and v) max_steps for refinement as 3. During inference, we have used a beam width of $k = 5$ to generate a text between a length of 5 and 64. All random seeds were fixed for better reproducibility. All experiments were conducted using an NVIDIA GPU with 32 GB of memory using PyTorch. QSGT requires 11.8 GB of memory during training and 6.2 GB during inference. Each epoch takes 90-100 min on average, with each sample taking 3-4 sec, and a maximum of 12 hours to complete the training for all classes.

5. Results

We have measured the performance of QSGT both quantitatively (loss and metric scores) and qualitatively (radiologists' remarks, attention localization). To measure the robustness of the model, we have also performed a sensitivity analysis (ablation study) and model confidence(statistical analysis), along with a performance comparison with the SOTA VLM.

5.1. Loss Function Behavior

The overall loss of QSGT is the cumulative contribution of pair-penalty loss(PP), seniority-penalty loss(SP), as returned by the DOCI module and the text loss from the decoder. All the losses exhibit a steady convergence, indicating improvements with epoch progression (Fig5). The text loss, which is an indicator of exact textual overlap between the reference and generated text, was found to be the most consistent for all disease classes across all the losses. In such a decoder-only model, this confirms the effective vision-conditioned language modeling using quantum principles, which helps improve the class-level textual fluency. Text loss for Pneumonia from 13.5 to 2.6, and for Pleural Effusion from 11.8 to 2.3 are some reflections. The progressive decrease of the pair-penalty loss (PP) demonstrates its effectiveness



(a) Pair Loss (b) Seniority Loss (c) Text Loss

Figure 5: QSGT: Epoch vs. Loss Plot

on the design goal to penalize the cross-modal token misalignment. The PP found Pneumonia and Pleural Effusion with difficult initial alignments, but helped them converge below 0.40. This indicates the contribution of PP in cross-modality conflict suppression. The seniority-penalty (SP) was designed to discourage the unimodal dominations and encourage multimodal joint clinical decisions. The SP loss for all the diseases reaches a stable range of 0.003-0.004, confirming the model's adaptability to cross-modal fused features.

5.2. Metric-wise Performance

Another quantitative measure of such models is to capture the nature of generation, summarization and faithfulness. The generation scores were observed using METEOR, SacreBLEU, and SPICE (Fig 6); the summarization scores were observed using ROUGE-2, ROUGE-L, and BERT (Fig 7); and the factuality score (Fig 8) was observed using RadGraphF1. The SPICE score exhibits the highest gain among all the generation scores, indicating a near-exact lexical match, specifically for clear chest. Clear chest results in all observations being the best, and Fibrosis is poor due to the obvious data imbalance. Otherwise, for pneumonia (16 to 66), cardiomegaly(0 to 68), and pneumothorax(0 to 55) are positive indications of accurate lexical and semantic consistencies in terms of SPICE. Except for fibrosis (32), the final METEOR scores for all other disease classes range between 42 (consolidation) and 50 (cardiomegaly), with the highest being clear chest (82). This showcases the improvement in structural alignments. A similar trend was observed for SacreBLEU, where the score ranges mostly between 20(fibrosis) and 30(cardiomegaly), indicating the absence of lexical similarity. This result is supportive of the abstractive nature of the summary, where the summaries must be semantically consistent, if not structurally.

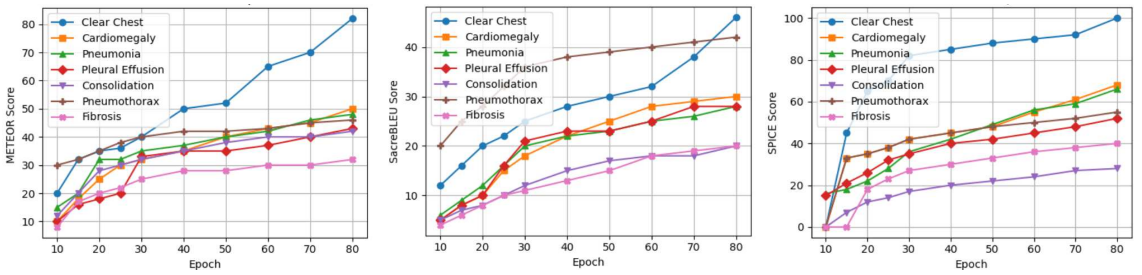
The summarization scores, specifically ROUGE-2 and ROUGE-L, not only check for exact matching but also token placements. Thus, a paraphrased summary will face a lower ROUGE score, but higher generation scores. Except for clear chest with the best summarization scores (85, 92), scores for the other disease classes were found to be comparative. The BERT score exceeds 98 across all classes, indicating a good embedding level similarity across the classes. The RadGraph score was observed to be initially zero for most of the diseases, but exhibits a steady and progressive escalation with epochs, indicating clinically sound and faithful generations. Although the minimum score observed was 37 (consolidation), pneumothorax (77), fibrosis (50), and pleural effusion (57) are some noticeable improvements, with Pneumonia and consolidation being the toughest to attain.

5.3. Qualitative Analysis

This section represents the qualitative analysis of QSGT, based on the target task of summarization, and its spatial grounding.

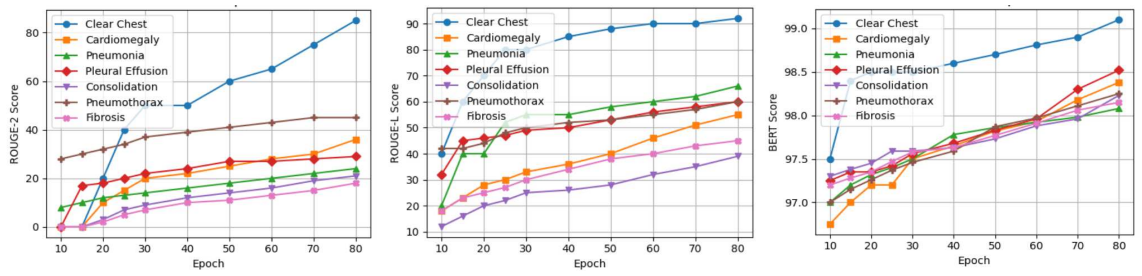
5.3.1. Quality of Summarization

Table 1 presents the qualitative evaluation of the proposed QSGT across seven representative thoracic conditions based on the RadGraph score, highlighting the clinical congruence of the generated summary with the reference reports, as verified by expert remarks. Random samples were drawn from each class, and individual scores were computed. For a clear chest, the entity and its negation were successfully captured, but failed to predict the severity term ("acute"). Although the RadGraph score of 80 can also be obtained from "acute cardiopulmonary disease", QSGT correctly predicts the negation ("no"), to justify the SPICE score. For both Pneumonia and Pneumothorax, QSGT accurately retain the core pathological entities, with position or description, enabling the model to score 66.67 for both cases.



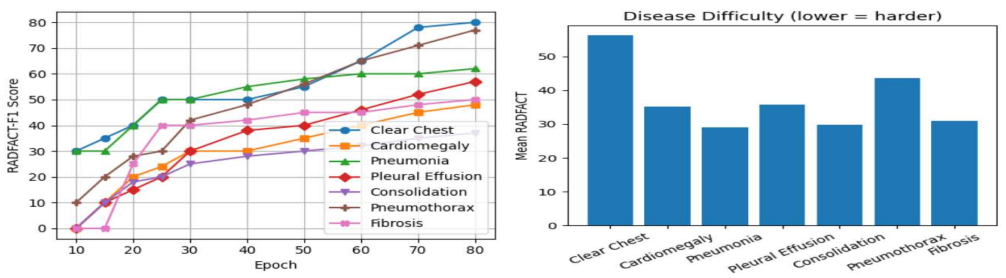
(a) METEOR Score (b) SacreBLEU Score (c) SPICE Score

Figure 6: QSGT: Epoch-wise average generation scores



(a) ROUGE-2 Score (b) ROUGE-L Score (c) BERT Score

Figure 7: QSGT: Epoch-wise average summarization scores



(a) RAD Score (b) Disease Difficulty

Figure 8: QSGT: Epoch-wise factuality scores and disease difficulty

A moderate performance of QSGT was observed in RadGraph scores for Cardiomegaly(50), Consolidation(50) and fibrosis (40), due to missing additional descriptions like "opacity in the left base", "focal pulmonary", and "cystic". Overall observation of the quality of generation reveals the model's ability to prioritize the dominant medical entity and its most suitable neighbouring tokens, which well justifies the objective of the principles used. Increasing the N_{min}, minimum and maximum generation length, and decoder may help improve the entire sequence while preserving the same principles.

5.3.2. Quality of Spatial Localization

We have represented the few-shot attention visualization of QSGT in Figure 9, for PA radiographs only. Since QSGT is searching for key localized patches in vision, supporting the underlying radiographic conception, it focuses mainly on the lung fields, their presence and absence, shadowing, collapse or shrinking, etc. Each row corresponds to a specific class, with a best, moderate and poor assessment level. For a clear chest, since no abnormalities exist, with lung fields equally spread, QSGT best generates a balanced attention. Poor attention, unable to highlight the uniformity in the trachea, or mediastinal shift. In cardiomegaly, where in a radiograph we have an enlarged heart, although the QSGT

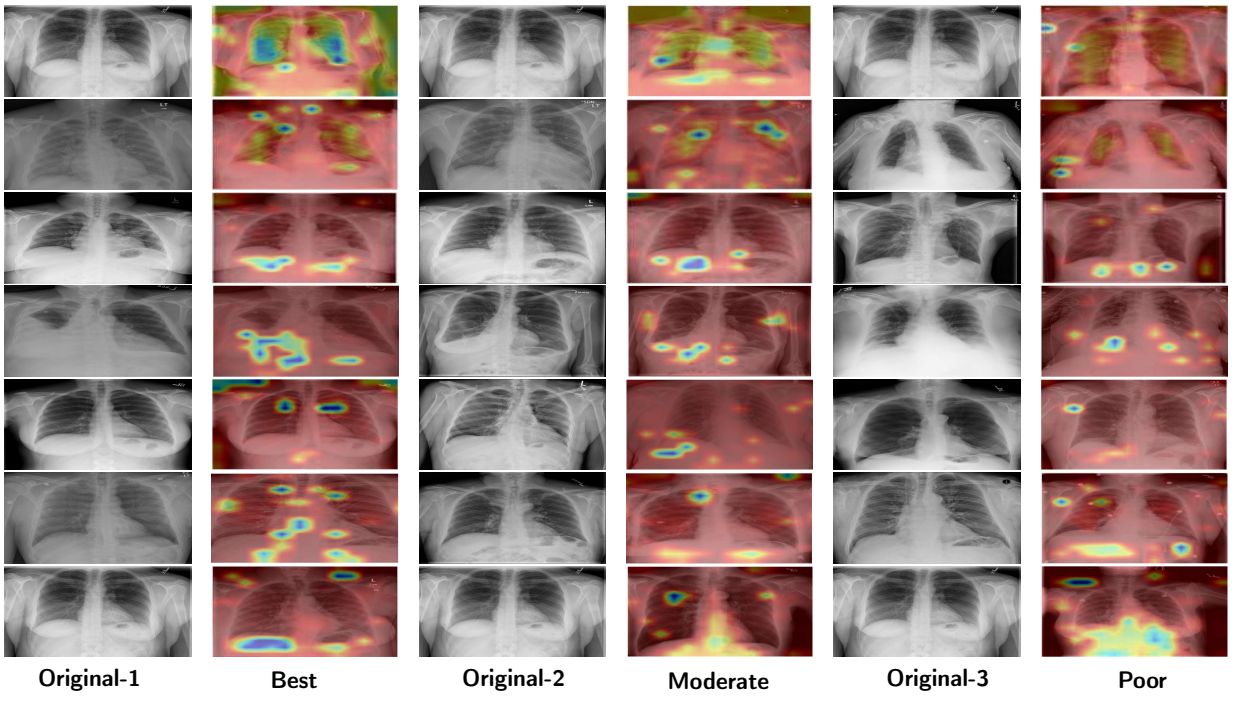


Figure 9: QSGT: Few-shot visualization of randomly chosen samples from IU_CXR. We present both the original image(left) and attention maps generated: Best, Moderate and Poor (right). Each row corresponds to a specific case: clear chest, cardiomegaly, pneumonia, pleural effusion, consolidation, pneumothorax and fibrosis (From top to bottom).

was not able to identify the width, it captured the areas near the left heart border, indicating cardiac enlargement. The pneumonia visualization suggests inflammation of the lung parenchyma appears in forms of airspace opacity. QSGT here correctly identifies the lower lung areas in the best case, while in the worst case, it falsely identifies the area. In a pleural effusion radiograph, a homogeneous dense opacity is observed and correctly identified by QSGT in a good case, while in a poor case, it captures a scattered pattern instead. Consolidation is a case of patchy airspace shadowing. In a good performance, QSGT highlights the parenchymal regions on both lungs. While for moderate or poor, it invokes dispersed attention. For pneumothorax, it correctly identifies the lung margins that For a clear chest, since no abnormalities exist, with lung fields equally spread, QSGT best generates a balanced attention. Poor attention, unable to highlight the uniformity in the trachea, or mediastinal shift. In cardiomegaly, where in a radiograph we have an enlarged heart, although the QSGT was not able to identify the width, it captured the areas near the left heart border, indicating cardiac enlargement. The pneumonia visualization suggests inflammation of the lung parenchyma appears in forms of airspace opacity. QSGT here correctly identifies the lower lung areas in the best case, while in the worst case, it falsely identifies the area. In a pleural effusion radiograph, a homogeneous dense opacity is observed and correctly identified by QSGT in a good case, while in a poor case, it captures a scattered pattern instead. Consolidation is a case of patchy airspace shadowing. In a good performance, QSGT highlights the parenchymal regions on both lungs. While for moderate or poor, it invokes dispersed attention. For pneumothorax, it correctly identifies the lung margins to support the concept of the lung edge seen, but no lung marking beyond the lung edge. In case of fibrosis, QSGT to ground the phrases like "Reticular or Nodular Shadowing", can emphasize on the interstitial regions, unlike the moderate and poor cases, where it incorrectly localize the stomach area.

5.4. Ablation Study

Reports of the ablation results of QSGT are tabulated under three complementary evaluation metrics (SPICE in Table 2, BERTScore in Table 3, and RadGraph Score in Table 4), across seven disease categories. We have used the abbreviations: CC (Clear Chest), CAR (Cardiomegaly), PNI (Pneumonia), PLE(Pleural effusion), CON(Consolidation), PNEU (Pneumothorax), and FIB (Fibrosis) for all analyses. The full QSGT model consistently attains the highest scores

Table 1
Qualitative Analysis of Summarization

Disease	Reference	Generated	RAD Score	Remark
Clear Chest	no acute cardiopulmonary disease.	no cardiopulmonary disease	80	Good
Cardiomegaly	1. unchanged cardiomegaly 2. opacity in the left base atelectasis	pulmonary cardiomegaly with atelectasis	50	Moderate
Pneumonia	left lower lobe airspace disease consistent with pneumonia	disease airspace pneumonia	66.67	Good
Pleural Effusion	bilateral pleural effusions, right larger than left.	right larger pleural effusions	60	Good
Consolidation	no focal pulmonary consolidation	no consolidation	50	Moderate
Pneumothorax	no significant change in right pneumothorax	no change of right pneumothorax	66.67	Good
Fibrosis	cystic fibrosis with interstitial opacity	fibrosis with interstitial opacity	40	Moderate

Table 2
Ablation study: Effect on the maximum of the average SPICE Scores computed across all samples

Variant	CC	CAR	PNI	PLE	CON	PNEU	FIB
W/O Orbital	90.18	55.21	56.32	45.22	24.26	50.19	36.34
W/O Pair Penalty	82.18	42.06	36.11	35.24	17.81	42.29	27.56
W/O Seniority Penalty	85.62	45.09	42.80	40.57	20.76	45.37	30.28
W/O DOCI	70.33	38.47	28.75	32.88	14.90	38.46	23.66
W/O Orbital + DOCI	45.14	33.23	18.53	21.27	7.73	33.48	0.0
QSGT	**100**	**68.25**	**66.19**	**52.23**	**28.46**	**55.41**	**40.83**

Table 3
Ablation study: Effect on the maximum of the average BERT Scores computed across all samples

Variant	CC	CAR	PNI	PLE	CON	PNEU	FIB
W/O Orbital	98.81	97.93	97.92	97.96	97.88	97.97	97.91
W/O Pair Penalty	98.50	97.50	97.50	97.56	97.59	97.46	97.58
W/O Seniority Penalty	98.60	97.65	97.78	97.68	97.63	97.59	97.63
W/O DOCI	98.50	97.20	97.40	97.42	97.59	97.37	97.58
W/O Orbital + DOCI	98.40	97.00	97.20	97.35	97.38	97.15	97.28
QSGT	**99.10**	**98.38**	**98.08**	**98.52**	**98.23**	**98.25**	**98.15**

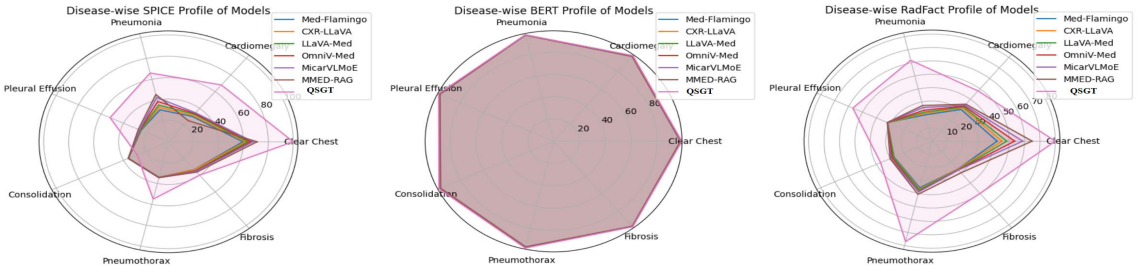
for all diseases and metrics, indicating that performance gains are systematic rather than disease-specific or metric-dependent. Removing DOCI produces the largest average degradation, with SPICE decreasing by up to 38 points (Pneumonia), RadGraph Score by up to 47 points (Pneumothorax), and BERTScore dropping by more than 1.0 across multiple diseases. This large drift makes the DOCI module the dominant contributor to the model performance, and the importance of the togetherness of both designed loss functions. Although their individual contribution to the model's robustness is less significant than that of the orbital refinement module, they help maintain a steady performance and metric consistency. A drastic fall in all three metric scores across all diseases was witnessed during the absence of both the orbital and DOCI module, which signifies the tight correlation and hierarchically supplementary nature.

Table 4
Ablation study: Effect on the maximum of the average RadGraph scores computed across all samples

Variant	CC	CAR	PNI	PLE	CON	PNEU	FIB
W/O Orbital	65.15	40.18	60.26	46.71	32.46	65.18	48.33
W/O Pair Penalty	50.22	30.66	50.72	30.58	25.73	42.65	42.51
W/O Seniority Penalty	50.10	30.38	55.49	38.27	28.36	48.76	45.88
W/O DOCI	50.29	24.44	50.28	20.45	20.31	30.63	40.51
W/O Orbital + DOCI	35.66	10.15	30.45	10.08	10.25	20.18	25.13
QSGT	**80.18**	**48.22**	**62.24**	**57.34**	**37.41**	**77.31**	**50.08**

5.5. Comparative Analysis

QSGT is compared with six recent multimodal medical VLMs based on SPICE, BERT and RadGraph scores, across seven thoracic disease categories (Fig 10, 11). MMED-RAG and Med-Flamingo were found to be the strongest and weakest baselines for all three scores. For a clear chest, the SPICE scores are highest for all the models, and for pleural effusion, they are lowest. Comparison yields a maximum improvement in SPICE score of 42.68 (25.32 to 68) for cardiomegaly and a minimum of 3.65 (33.65 to 40) for fibrosis, from the best baseline. Except for the clear chest, the SPICE score for all the baselines varies between 25 and 43, while for QSGT, it ranges from 40 to 68, except for consolidation. This confirms the semantic awareness of QSGT over the baselines. Although a very small margin of improvement was witnessed in the BERT score by QSGT, unlike the baselines, it never falls below 98 for any of the classes. This signifies the model's ability to generate better contextual fused embeddings than the baselines. While focusing on factuality, QSGT achieved significant gains for pneumonia (34.66) and pneumothorax (36.2), compared with the best SOTA model. Even for complex classes, it achieves satisfactory results (fibrosis=50). This showcases the model's consistency in capturing vital clinical information across the disease.



(a) SPICE Score (b) BERT Score (c) RAD Score

Figure 10: Disease-wise score profiles of the models

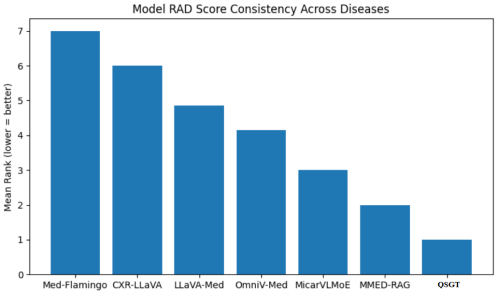


Figure 11: RAD Score Consistency Comparative

5.6. Statistical Analysis

The Wilcoxon signed-rank test indicates a statistically significant improvement in RadGraph-F1 scores across all diseases ($p < 0.01$). Effect sizes were quantified using Cliff's delta, revealing a large effect, thereby confirming that performance gains are substantial rather than incremental. Learning curves demonstrate monotonic improvement with score stabilization after epochs 75-80, suggesting training convergence and robustness. To quantify uncertainty in the final performance, we report a 95% confidence interval (CI) over disease-wise RadGraph-F1 scores at the final epoch. Using Student's t-distribution ($n = 7$), the mean final-epoch score of 58.71 yields a 95% CI of [44.9, 72.5]. This interval confirms statistically reliable performance gains across diverse thoracic pathologies while accounting for inter-disease variability.

6. Conclusion

This paper introduces a novel unified vision-language model. The conception of quantum many-body can be used to approximate a vision-language system and improve the downstream generation task by introducing quantum style token refinement and togetherness losses, by visualizing the vision-language systems as two atomic subsystems. The process of token pruning has replaced by the quantum orbital style electron assignment process for token refinement, and the expert configurations by the quantum joint approximation, to ensure no modality or spin is alone. The presence and contribution of each of these modules in improving the model's performance in terms of generation, summarization and faithfulness metrics were found to be evident and essential. All the results obtained in a low-resource environment proved the research objective in a sparse mode. The class-specific training has helped the model to better understand both the radiographic vision appearances and their associated lexical contingency. Achieving a very high quantitative metric with semantically sound phrasal consistency did not take place, and also not expected from this compression-based approach. Thus, a semantic consistency-improving module needed to be implanted. The use of the same training environment with the same hyperparameters for all classes, the use of pre-trained encoders instead of finetuned, not considering the concept of co-evolving states, and verification of one approximation function may have restricted the model from its best performance. We are further interested in enhancing the model for different anatomies, downstream tasks, and large vision language modeling.

7. Acknowledgment

The authors gratefully acknowledge the contribution of Dr. Ashok Kumar Maulik (MD AIIMS, Radiodiagnosis) for his expert assistance in preparing the knowledge base and verifying the qualitative study and attention maps of the model.

References

Buess, L., Keicher, M., Navab, N., Maier, A., Tayebi Arasteh, S., 2025. From large language models to multimodal ai: A scoping review on the potential of generative ai in medicine. Biomedical Engineering Letters 15, 845–863.

Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., Huang, J., 2025. A survey on mixture of experts in large language models. IEEE Transactions on Knowledge and Data Engineering .

Cao, Y., Ding, H., Zhang, Y., Hei, Y., 2025. Radiology report generation based on adaptive enhanced fusion of multi features. Computers in Biology and Medicine 193, 110494.

Chopra, S., Mao, L., Sanchez-Rodriguez, G., Feola, A.J., Li, J., Kira, Z., 2025. Medmoe: Modality-specialized mixture of experts for medical vision-language understanding. arXiv abs/2506.08356. URL: https://arxiv.org/abs/2506.08356. preprint.

Dong, P., Sun, M., Lu, A., Xie, Y., Liu, K., Kong, Z., Meng, X., Li, Z., Lin, X., Fang, Z., Wang, Y., 2022. Heatvit: Hardware-efficient adaptive token pruning for vision transformers. arXiv abs/2211.08110. URL: https://arxiv.org/abs/2211.08110. adaptive token selector inserted before transformer blocks for hardware-efficient pruning and significant compute reduction.

Gao, Y., Ni, Z., Liu, W., Ni, L., Xin, L., Hu, L., Zhang, L., 2025. Abnormal-region-aware multi-modal feature fusion for medical report generation. Knowledge-Based Systems , 113538.

He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., Chen, H., 2024. Foundation model for advancing healthcare: Challenges, opportunities and future directions. IEEE Reviews in Biomedical Engineering .

He, Z., Wong, A.N.N., Yoo, J.S., 2025. Radiology report generation using automatic keyword adaptation, frequency-based multi-label classification and text-to-text large language models. Computers in Biology and Medicine 196, 110625.

Huang, L., Guo, S., Bai, T., Zhao, R., Tao, K., 2025. Prompt-guided orthogonal multimodal fusion for cancer survival prediction. Information Sciences , 122242.

Huang, X., Shen, L., Liu, J., Shang, F., Li, H., Huang, H., Yang, Y., 2024. Towards a multimodal large language model with pixel-level insight for biomedicine. arXiv abs/2412.09278. URL: https://arxiv.org/abs/2412.09278. preprint.

Izhar, A., Idris, N., Japar, N., 2025a. Medical radiology report generation: A systematic review of current deep learning methods, trends, and future directions. Artificial intelligence in medicine , 103220.

Izhar, A., Japar, N., Idris, N., Dang, T., 2025b. Micarvlmoe: A modern gated cross-aligned vision-language mixture of experts model for medical image captioning and report generation. arXiv abs/2504.20343. URL: https://arxiv.org/abs/2504.20343. preprint.

Jiang, S., Zheng, T., Zhang, Y., Jin, Y., Yuan, L., Liu, Z., 2024. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics. pp. 3843–3860. URL: https://aclanthology.org/2024.findings-emnlp.221/.

Kim, S., Shen, S., Thorsley, D., Gholami, A., Hassoun, J., Keutzer, K., 2021. Learned token pruning for transformers. arXiv abs/2107.00910. URL: https://arxiv.org/abs/2107.00910. threshold-based learned token pruning that adaptively removes less important tokens to improve efficiency.

Kommanduri, R., Matta, K.K., Imandi, R., 2025. An anatomical-aware three-stream ensemble network for brain tumor classification. Information Sciences , 123043.

Li, H., Wang, H., Sun, X., He, H., Feng, J., 2025. Context-enhanced framework for medical image report generation using multimodal contexts. Knowledge-Based Systems 310, 112913.

Liu, Y., Gehrig, M., Messikommer, N., Cannici, M., Scaramuzza, D., 2023. Revisiting token pruning for object detection and instance segmentation. arXiv abs/2306.07050. URL: https://arxiv.org/abs/2306.07050. investigates token pruning strategies in dense prediction tasks and proposes dynamic pruning designs.

Lu, S.Y., Zhu, Z., Zhang, Y.D., Yao, Y.D., 2025. Tuberculosis and pneumonia diagnosis in chest x-rays by large adaptive filter and aligning normalized network with report-guided multi-level alignment. Engineering Applications of Artificial Intelligence 158, 111575.

Luo, Z., Jiang, Z., Wang, M., Cai, X., Gao, D., Yang, L., 2025. Chatgpt based contrastive learning for radiology report summarization. Expert Systems with Applications 267, 125827.

Mukesh, K., Jayaprakash, S., Kumar, R.P., 2024. Qvila: Quantum infused vision-language model for enhanced multimodal understanding. SN Computer Science 5, 1023.

Qu, Z., Li, Y., Tiwari, P., 2023. Qnmf: A quantum neural network based multimodal fusion system for intelligent diagnosis. Information Fusion 100, 101913.

Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J., 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification, in: NeurIPS 2021. URL: https://arxiv.org/abs/2106.02034. dynamic token pruning framework for vision transformers, reducing up to 66% of tokens with minimal accuracy drop.

Relan, S., Rambola, R.K., 2025. Big bird-disentangled representation and adaptive contrast transformer for abstractive text summarization with contrast attention. Engineering Applications of Artificial Intelligence 162, 112536.

Rezvani, A., Yan, X., Wu, A.T., Han, K., Khosravi, P., Xie, X., 2025. Mome: Mixture of visual language medical experts for medical imaging segmentation. arXiv abs/2510.26996. URL: https://arxiv.org/abs/2510.26996. preprint.

Sah, S., Kumar, R., Rohmetra, H., Saboori, E., 2024. Token pruning using a lightweight background aware vision transformer. arXiv abs/2410.09324. URL: https://arxiv.org/abs/2410.09324. uses a background-aware pre-processing block to identify and prune background tokens, improving throughput in object detection.

Sankaradass, V., Manish, V.M., 2026. Quantum-inspired predictive healthcare system for personalised oncology using multimodal data fusion and secure blockchain protocols. Engineering Applications of Artificial Intelligence 167, 113730.

Singh, S.S., Kumar, S., Ahuja, R., Barua, J., 2025. Fusion of quantum computing and explainable ai: A comprehensive survey on transformative healthcare solutions. Information Fusion , 103217.

Su, Z., Song, H., Lin, Y., Wu, Y., Weng, X., Mao, Z., Liu, B., Yin, H., Yang, J., 2025. Medkit: Multi-level feature distillation with knowledge injection for radiology report generation. Expert Systems with Applications , 129003.

Tang, Q., Zhang, B., Liu, J., Liu, F., Liu, Y., 2023. Dynamic token pruning in plain vision transformers for semantic segmentation, in: arXiv. URL: https://arxiv.org/abs/2308.01045. dynamic token pruning tailored for semantic segmentation tasks, maintaining computational efficiency.

Touheed, M., Zubair, U., Sabir, D., Hassan, A., Butt, M.F.U., Riaz, F., Abdul, W., Ayub, R., 2024. Applications of pruning methods in natural language processing. IEEE Access 12, 89418–89438.

Wang, A., Mao, D., Li, X., Li, T., Li, L., 2025a. Hqnet: A hybrid quantum network for multi-class mri brain classification via quantum computing. Expert Systems with Applications 261, 125537.

Wang, S., Chen, K., Yu, M., Zhao, P., Duan, H., 2025b. Cdmrnet: multimodal meta-adaptive reasoning network with dynamic causal modeling and co-evolution of quantum states. Scientific Reports 15, 26370.

Wang, X., Figueredo, G., Li, R., Zhang, W.E., Chen, W., Chen, X., 2025c. A survey of deep-learning-based radiology report generation using multimodal inputs. Medical Image Analysis 103, 103627. URL: https://www.sciencedirect.com/science/article/pii/S1361841525001744, doi:https://doi.org/10.1016/j.media.2025.103627.

Wang, Z., Wei, S., Hsu, D., Lee, J.D., 2024. Transformers provably learn sparse token selection while fully-connected nets cannot. arXiv preprint arXiv:2406.06893 .

Xia, P., Chen, Z., Tian, J., Gong, Y., Hou, R., Xu, Y., Wu, Z., Fan, Z., Zhou, Y., Zhu, K., et al., 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. Advances in Neural Information Processing Systems 37, 140334–140365.

Xu, D., Chen, Y., Chai, Z., Xiao, Y., Yan, Y., Ding, W., Wang, H., Jin, Z., Jiao, W., Yue, W., et al., 2026. Knowledge fusion in deep learning-based medical vision-language models: A review. Information Fusion 125, 103455.

Zhou, F., Liu, X., Yang, Z., Li, Z., Xiao, H., 2026. Enhancing medical mllms with dual vision encoders and moe-based modality projector. Knowledge-Based Systems , 115275.

Zhu, X., Hu, Y., Mo, F., Li, M., Wu, J., 2024. Uni-med: A unified medical generalist foundation model for multi-task learning via connector-moe. arXiv abs/2409.17508. URL: https://arxiv.org/abs/2409.17508. preprint.