MIRAGE: Modality-Aware 3D Vision-Language Model for Liver Lesion Diagnosis in MRI via Volumetric-Text Alignment

Haowen Zhu, Jingyang Zhang, and Chunfeng Yang*
School of Computer Science and Engineering, *Southeast University*, Nanjing, China
Email: {213233697, j.y.zhang}@seu.edu.cn, chunfeng.yang@seu.edu.cn (*Corresponding author)

***Abstract*—The accurate diagnosis of liver lesions from 3D multimodal Magnetic Resonance Imaging (MRI) is a critical yet challenging task in oncological decision-making. Existing deep learning approaches often struggle due to the scarcity of annotated volumetric medical data and the inherent heterogeneity of MRI sequences, where distinct physical parameters (e.g., T2-weighted vs. DWI) create significant distributional shifts. Furthermore, current medical Vision-Language Models (VLMs) adapted for 3D tasks often rely on static early fusion strategies, treating diverse MRI modalities as homogeneous input channels. This indiscriminate integration leads to feature interference and fails to replicate the dynamic, semantic-driven nature of radiological interpretation. To address these limitations, we propose MIRAGE (Multimodal MRI Representation Aggregation with Gated Experts). We introduce a library of modality-aware expert encoders, adopting a decoupled learning paradigm: experts are first contrastively pretrained to align specific physical imaging features with LLM-synthesized radiology semantics, ensuring robust representation learning prior to diagnosis. To mimic the selective attention of radiologists, we design a Modality Routing Gate (MRG) that dynamically prioritizes the most discriminative expert path for each patient, reducing redundancy and gradient conflicts. Additionally, a Cross-modal Interactive Feature Extraction and Refinement (CIFER) module is employed to fuse these selective features with textual knowledge. Extensive evaluations on the large-scale LLD-MMRI dataset demonstrate that MIRAGE achieves state-of-the-art performance with 91.57% accuracy, significantly outperforming existing medical VLMs while maintaining computational efficiency suitable for clinical deployment.**

***Index Terms*—Vision-Language Model, Multimodal MRI, Liver Lesion Diagnosis, Modality-Aware Learning, Medical Image Analysis**

I. Introduction

Accurate diagnosis of liver lesions—distinguishing hepatocellular carcinoma (HCC) and metastasis from benign cysts—is a cornerstone of oncological decision-making [1]. Magnetic Resonance Imaging (MRI) serves as the gold standard for this task, relying on a set of co-registered 3D sequences (e.g., T1, T2, DWI) acquired under distinct physical parameters. Each sequence captures unique physiological properties: DWI highlights cellular density indicative of malignancy, while dynamic contrast-enhanced phases reveal vascular patterns [2], [3].

Despite the richness of this data, automating diagnosis remains challenging. Traditional Deep Learning (DL) models are constrained by the scarcity of annotated 3D medical datasets [4]. The recent emergence of Vision-Language Models (VLMs) offers a promising solution by leveraging semantic knowledge from medical reports [5], [6]. However, adapting general VLMs to the nuances of liver MRI faces two fundamental hurdles.

First, the Dimensionality Gap: SOTA medical VLMs like MedCLIP [7] are predominantly 2D. Naive slice-based adaptation forfeits the volumetric spatial context essential for assessing lesion morphology [8].

Second, the Modality Heterogeneity: While multi-stream architectures exist, standard VLM adaptations often default to early fusion strategies, such as concatenating disparate MRI sequences into a unified input tensor [9]. This approach implicitly treats distinct physical modalities (e.g., fluid-sensitive T2 vs. diffusion-based DWI) as homogeneous channels. Such indiscriminate integration ignores the significant statistical distributional shifts between modalities, leading to feature interference where noise from less informative sequences dilutes the diagnostic signals of the dominant one.

This limitation contrasts sharply with clinical workflow. The interpretation of liver MRI is inherently dynamic and semantic-driven. Radiologists do not weigh all sequences equally; instead, they prioritize specific modalities based on diagnostic cues found in reporting guidelines—for instance, focusing on the arterial phase to verify "hyper-vascularization" or DWI to confirm "restricted diffusion." Existing DL approaches, by forcing static fusion, struggle to replicate this selective attention mechanism. Furthermore, simultaneously optimizing for diverse physical modalities can create gradient conflicts, as different sequences may require contradictory feature extraction pathways to reach convergence. Consequently, there is a critical need for architectures that can mimic this radiological workflow by dynamically routing attention to the most discriminative modality.

To address these limitations, we propose **MIRAGE** (**M**ultimodal MRI **R**epresentation **A**ggregation with **G**ated **E**xperts). Unlike methods that enforce a shared encoder or static fusion, MIRAGE adopts a "Modality-Aware" philosophy. We introduce a library of specialized expert encoders, utilizing a pretraining stage to explicitly align each MRI sequence with LLM-generated radiology semantic descriptions. This ensures that each expert learns robust, modality-specific representations prior to the diagnostic task. Crucially, to han-

dle the multimodal input efficiently, we design a Modality Routing Gate (MRG). Mimicking the radiologist's selective process, the MRG prioritizes the most discriminative expert path during inference—such as focusing on arterial experts when HCC features are present—thereby reducing redundancy and enhancing interpretability.

Our contributions are:

- We propose a modality-aware 3D VLM framework that explicitly models the heterogeneity of eight MRI sequences via specialized expert pretraining.
- We introduce the MRG and CIFER modules to dynamically select and deeply fuse diagnostic features, effectively bridging the gap between clinical reasoning and model inference.
- Evaluations on the LLD-MMRI dataset show that MIRAGE achieves state-of-the-art performance, outperforming strong baselines like MCPL.

II. Related Work

A. Deep Learning for Liver Lesion Analysis

Deep learning has revolutionized liver lesion diagnosis. Early approaches relied on 2D CNNs applied to individual slices, which were later superseded by 3D networks like 3D-ResNet and V-Net that capture volumetric context [10]. Recent Transformer-based models, such as TransLiver [11], have further improved performance by modeling long-range dependencies. However, these fully supervised methods depend heavily on large-scale voxel-wise annotations, which are scarce and expensive to obtain. Moreover, they operate as "closed-set" classifiers, failing to leverage the rich, unstructured semantic knowledge embedded in clinical radiology reports [12].

B. Vision-Language Models in Medicine

Vision-Language Models (VLMs) have emerged to address data scarcity by learning from image-text pairs. General domain models like CLIP [5] have been adapted to medicine (e.g., MedCLIP [7], BiomedCLIP [13]), showing remarkable zero-shot capabilities [14], [15]. A comprehensive review by Lin et al. [6] highlights the rapid evolution of this field. Despite these advances, a significant gap remains in 3D Volumetric Imaging. Most existing medical VLMs are designed for 2D X-rays or pathology slides. Recent attempts to extend VLMs to 3D, such as M3D [9] or CT-CLIP [16], primarily focus on CT data [17]. They typically employ a unified visual encoder for all inputs, which is suboptimal for multi-sequence MRI where each sequence (e.g., T1, DWI) has distinct physical properties.

C. Learning from Heterogeneous Modalities

Handling the heterogeneity of multimodal medical data is a persistent challenge [18]. Recent works have explored advanced graph-based learning [19] to model complex inter-modality interactions, or distilled prompt learning [20] to handle incomplete modalities during inference. Our work aligns with the Mixture-of-Experts (MoE) philosophy [21], which posits that specialized sub-networks learn better representations than a single shared network. While MoE has been applied to natural language processing and general computer vision, MIRAGE is the first to tailor this concept for 3D Multimodal MRI. By assigning dedicated experts to different MRI sequences and coordinating them via a routing gate, we explicitly resolve the feature interference problem inherent in single-stream architectures.

III. Methodology

We formulate liver lesion diagnosis as a multimodal reasoning task. Let a patient case be represented by a set of K co-registered MRI sequences $\mathcal{V} = \{V^1, ..., V^K\}$ (e.g., T2, DWI, Arterial), where each $V^k \in \mathbb{R}^{D \times H \times W}$. Our goal is to predict the lesion category y by leveraging both visual data and clinical semantic knowledge. The MIRAGE framework (Fig. 1) consists of two phases: Modality-Aware Pretraining and Task-Specific Fine-tuning.

A. LLM-Assisted Radiology Caption Generation

To bridge the semantic gap in data-scarce medical domains, we utilize GPT-4o to generate synthetic radiology reports. For each volume in the training set labeled with triplet $\{c_i, m_i, o_i\}$ (disease, modality, organ), we construct structured prompts to generate detailed descriptions. Clinical Validity Strategy: To address potential hallucinations or biases common in general LLMs (a concern raised in recent studies), our generation pipeline is strictly conditioned on ground-truth labels. Furthermore, we instructed the model to generate $M = 5$ diverse candidate reports per volume, focusing on distinct radiological attributes (e.g., texture, boundaries). A subset of these captions underwent rigorous review by senior radiologists to ensure clinical accuracy, resulting in a robust report set $\mathcal{T}$ for pretraining.

B. Modality-Aware Vision-Language Pretraining

Standard 3D encoders often struggle with the heterogeneity of MRI sequences (e.g., fluid-sensitive T2 vs. cellular-sensitive DWI). We tackle this by creating K modality-aware expert encoders $\{E_k\}_{k=1}^K$, where each expert is a 3D ResNet-based backbone specialized for a specific physical modality.

During pretraining, for an input pair (V_i^k, T_i) where V_i^k is a volume of modality k and T_i is its report, we extract the visual embedding $v_i \in \mathbb{R}^d$ and text embedding $t_i \in \mathbb{R}^d$ via non-linear projection heads. We employ a symmetric Cross-Modal Contrastive (CMC) loss to align these representations. The volume-to-text loss is formulated as:

$$\mathcal{L}_{\text{v2t}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(sim(v_i, t_i)/\tau)}{\sum_{j=1}^{N} \exp(sim(v_i, t_j)/\tau)} \tag{1}$$

where $sim(\cdot, \cdot)$ denotes cosine similarity and τ is a learnable temperature parameter. The total pretraining objective averages the bidirectional losses: $\mathcal{L}_{\text{CMC}} = (\mathcal{L}_{\text{v2t}} + \mathcal{L}_{\text{t2v}})/2$. This step ensures that each expert E_k learns to extract features specific to its modality's physical characteristics, establishing a shared latent space for downstream tasks.

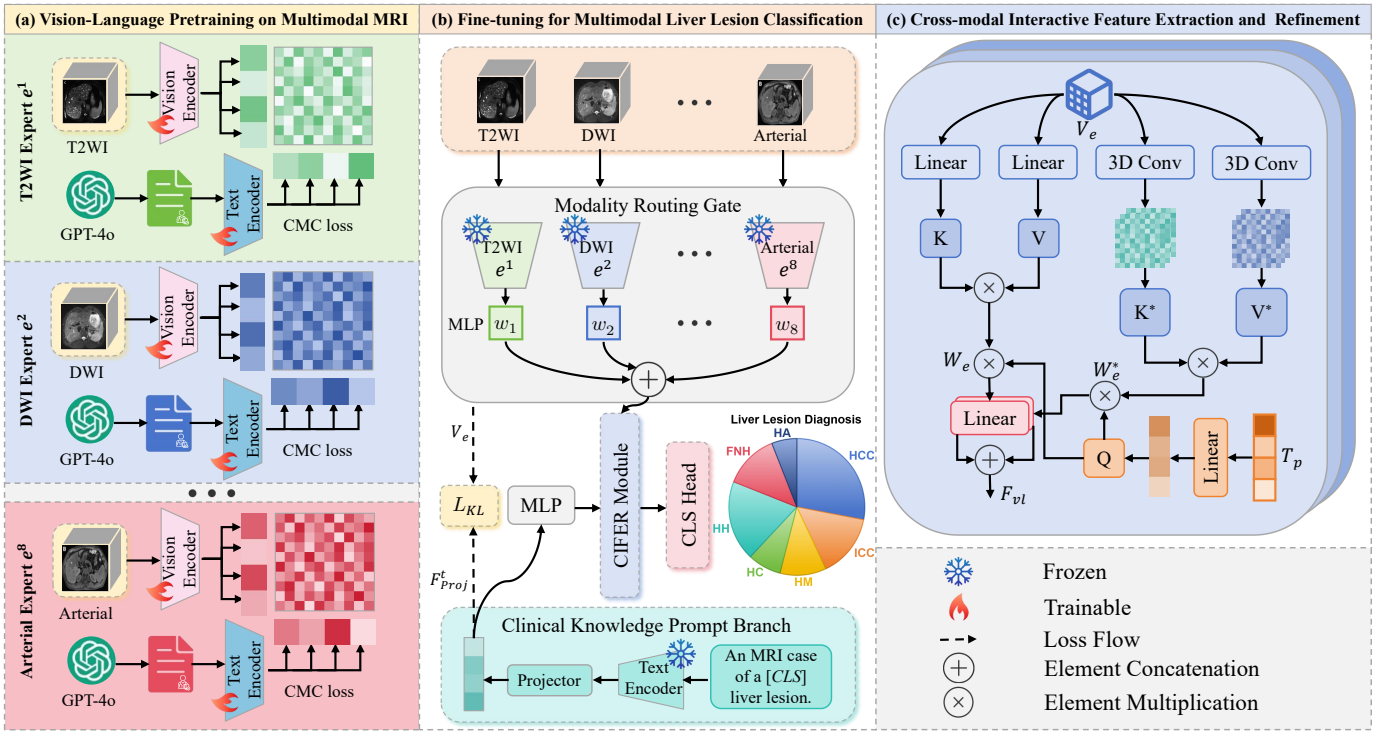

Fig. 1. Overview of the MIRAGE framework. (a) Modality-Aware Pretraining: Specialized experts learn sequence-specific representations aligned with LLM-generated reports. (b) Fine-tuning: The MRG dynamically selects the most discriminative expert to minimize inference redundancy, while the CIFER module performs deep vision-text fusion.

C. Fine-tuning with Modality Routing Gate (MRG)

During fine-tuning, the input is the patient's full multimodal set $\mathcal{V}$. Processing all K volumes with deep experts simultaneously is computationally prohibitive ($8\times$ FLOPs) and prone to feature interference. We introduce the Modality Routing Gate (MRG) to optimize inference efficiency and focus.

The MRG consists of a shallow 3D CNN feature extractor followed by a two-layer Multi-Layer Perceptron (MLP). It computes a routing probability vector $\mathbf{w} \in \mathbb{R}^K$. To simulate the radiologist's workflow of identifying the "key diagnostic sequence" (e.g., wash-out in Delayed phase) and to avoid gradient conflict between modalities, we employ a Top-1 Routing strategy:

$$V_e = E_{k^*}(V^{k^*}), \quad k^* = \arg\max_k \mathbf{w}_k \tag{2}$$

This mechanism activates only the single most relevant expert backbone for each patient sample. Unlike soft-attention mechanisms that process all inputs, our hard-routing strategy significantly lowers the computational cost and enforces parameter specialization. Simultaneously, we construct a Clinical Knowledge Prompt (CKP) T_p (e.g., "An MRI case of [CLS]"), encoded by the frozen BioBERT to guide the visual features with prior semantic context.

D. Cross-modal Interactive Feature Refinement

To deeply fuse the selected expert feature V_e with the prompt T_p, we utilize the CIFER module via Multi-Head Cross-Attention (MHCA). T_p serves as the Query (Q), driving the retrieval of relevant anatomical details from V_e, which serves as Key (K) and Value (V). To capture visual features at multiple abstraction levels, we adopt a dual-path design. We compute attention weights for both the raw expert features V_e and a deeper abstraction V_e^* processed by additional $3\times3\times3$ convolutional layers:

$$\begin{aligned} W_e &= \text{Softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V, \\ W_e^* &= \text{Softmax}\left(\frac{Q(K^*)^T}{\sqrt{d_h}}\right)V^* \end{aligned} \tag{3}$$

where d_h is the dimension of attention heads. The final multimodal representation is obtained by concatenating the outputs: $F_{vl} = [W_e \oplus W_e^*]$. This hierarchical interaction ensures the model captures both fine-grained texture (from V_e) and global semantic patterns (from V_e^*).

E. Curriculum-Based Hybrid Optimization

The model is optimized using a hybrid objective combining classification accuracy and semantic alignment. To address the distribution mismatch, we define the alignment loss using Kullback-Leibler (KL) divergence on the *softmax-normalized probability distributions* of the features, rather than raw vectors. Let $\sigma(\cdot)$ be the Softmax function. The total loss is:

$$\mathcal{L}_{total} = \lambda_c \mathcal{L}_{\text{BCE}}(Pred, y) + \lambda_s \mathcal{L}_{\text{KL}}(\sigma(V_e/\tau)||\sigma(T_p/\tau)) \tag{4}$$

Crucially, we employ a Curriculum Learning schedule for the coefficients λ_c and λ_s to stabilize training. We use a ramp-down/ramp-up strategy:

$$\lambda_c = 1.0 \cdot e^{-5(\frac{t}{T})}, \quad \lambda_s = 0.1 \cdot e^{-5(1-\frac{t}{T})} \tag{5}$$

where t is the current epoch and T is the max epochs. This schedule prioritizes discriminative classification (high λ_c) in early stages and gradually emphasizes semantic consistency (increasing λ_s), preventing premature overfitting to noisy visual cues while ensuring task-specific convergence.

IV. Experimental Setup and Results

A. Dataset Description

We evaluate MIRAGE on the public LLD-MMRI benchmark [22], which represents one of the largest available cohorts for multi-sequence liver MRI. While containing 498 patients, the dataset is volumetrically rich, comprising 3,984 co-registered 3D MRI volumes. The data spans seven distinct lesion categories: four benign (Hemangioma, Cysts, FNH, Abscess) and three malignant (HCC, ICC, Metastasis). Crucially, each patient case includes eight aligned imaging phases with distinct contrast mechanisms: Non-contrast, Arterial, Venous, Delayed, T2-weighted (T2WI), Diffusion-weighted (DWI), T1 in-phase, and T1 out-of-phase. For evaluation, we strictly adhere to the official split: 316 patients for training, 78 for validation, and 104 for testing.

B. Implementation Details

Network Architecture: We utilize the 3D image encoder from the CLIP-Driven Foundation Model [17] as the backbone for our expert encoders due to its robust initialization. For the text branch, we employ BioBERT [23] to ensure domain-specific semantic alignment.

Training Protocol: Experiments were conducted on a cluster of 4 NVIDIA A40 GPUs using PyTorch.

- *Pretraining:* The model is trained for 300 epochs (Batch size=32) using AdamW optimizer (lr $= 10^{-4}$, weight decay $= 0.01$). The temperature τ is initialized to 0.07.
- *Fine-tuning:* We fine-tune for 500 epochs with a reduced learning rate of 10^{-5}. Input volumes are normalized via Min-Max scaling and resized to $128 \times 128 \times 128$.

Baseline Settings: To ensure a fair comparison and address potential concerns regarding zero-shot performance, all baseline methods were fully fine-tuned on the LLD-MMRI training set under identical protocols and data augmentations.

C. Comparison with State-of-the-Art

We benchmark MIRAGE against five diverse SOTA methods: the Baseline (Standard 3D CLIP) [17], causal inference method TDSASH [24], and medical VLMs including MedCLIP [7], TraumaDet [25], and MCPL [26].

Quantitative Analysis: As reported in Table I, MIRAGE achieves state-of-the-art performance with 91.57% Accuracy and 88.14% AUC. While achieving a competitive F1 score comparable to the strong baseline (MCPL), our method significantly outperforms it in overall accuracy by a margin of 3.7%. This margin validates our core hypothesis: treating heterogeneous MRI sequences (e.g., T1 vs. DWI) uniformly leads to feature interference. Our Modality-Aware Experts disentangle these conflicts, allowing the model to leverage the unique diagnostic value of each sequence.

TABLE I
Performance Comparison on the LLD-MMRI Dataset. All methods are fully fine-tuned. Best results are in **bold**.

Method	ACC (%)	F1 (%)	AUC (%)
Baseline [17]	82.86 ± 0.17	74.86 ± 1.18	75.15 ± 2.32
TDSASH [24]	83.22 ± 1.46	78.26 ± 2.23	80.36 ± 1.13
MedCLIP [7]	85.53 ± 1.54	84.33 ± 1.49	84.21 ± 3.45
TraumaDet [25]	84.24 ± 2.21	85.33 ± 2.37	84.53 ± 2.32
MCPL [26]	87.87 ± 2.41	$\mathbf{87.56 \pm 2.41}$	86.49 ± 2.54
MIRAGE (Ours)	$\mathbf{91.57 \pm 1.56}$	87.33 ± 3.57	$\mathbf{88.14 \pm 2.25}$

Qualitative Visualization: Fig. 2 illustrates the ROC curves, demonstrating that MIRAGE maintains high sensitivity even at low false-positive rates. Complementing this, the Radar chart in Fig. 3 highlights our balanced performance across diverse pathologies, accurately classifying challenging malignant subtypes (e.g., HCC, ICC) where baselines often fail due to subtle textural differences.

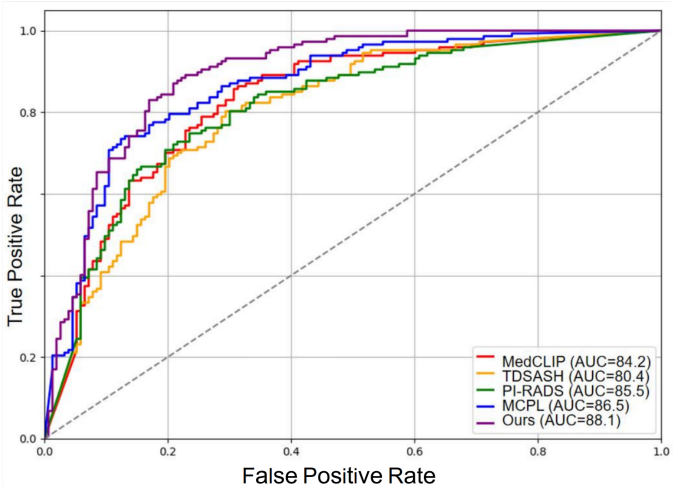

Fig. 2. ROC curves comparison on the LLD-MMRI dataset. MIRAGE (purple curve) encloses the largest area (AUC=88.1%), demonstrating superior sensitivity and specificity balance compared to baselines.

D. Computational Efficiency Analysis

Beyond diagnostic accuracy, we analyze the computational feasibility of MIRAGE. While utilizing $K = 8$ modality-specific experts increases static parameter storage, our Top-1 Modality Routing Gate (MRG) ensures inference efficiency. Unlike ensemble approaches that aggregate outputs from all sub-networks (incurring $8\times$ computational cost), our routing mechanism dynamically activates only the single most discriminative expert path for each patient case. Consequently, the inference FLOPs remain comparable to a standard single-stream 3D backbone ($\approx$ 45 GFLOPs). Crucially, since only one expert is active at a time, the peak memory footprint during inference is significantly minimized compared to parallel processing architectures. This design achieves an optimal

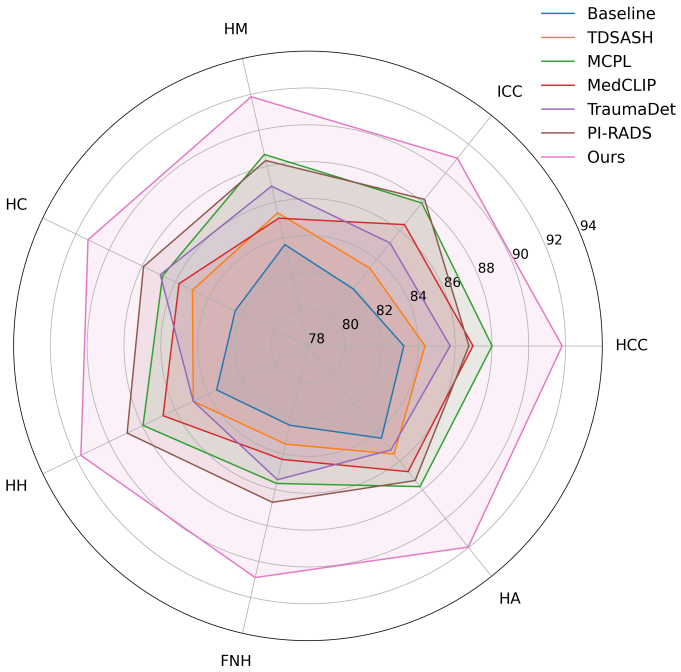

Fig. 3. Class-wise accuracy analysis. The Radar chart highlights that MIRAGE maintains balanced high performance across all 7 categories, effectively identifying challenging malignant subtypes (e.g., HCC, ICC).

trade-off, offering the high representational capacity of a large-scale model during training while maintaining the low latency and memory efficiency required for deployment on standard clinical workstations without high-end GPU clusters.

V. Ablation Study and Analysis

To validate the architectural decisions within MIRAGE, we conducted ablation studies on the LLD-MMRI dataset.

A. Component-wise Contribution

We first dissect the incremental gains of each module: Vision-Language Pretraining (VLP), Clinical Knowledge Prompts (CKP), and the CIFER fusion module. Table II details the results.

Impact of VLP: Modality-aware pretraining yields 84.22% accuracy, surpassing the scratch baseline (82.86%). Confirming LLM-generated reports provide richer semantic signals than class labels, establishing a superior starting point for encoders.

Impact of CKP: Clinical prompts improve accuracy by 3.03%, validating text embeddings as semantic anchors that guide visual features toward diagnostic regions.

Impact of CIFER: The largest boost (+4.32%) comes from CIFER, proving simple concatenation is insufficient. Our hierarchical cross-attention effectively bridges visual textures and textual semantics, resolving complex lesion ambiguities.

B. Modality-Awareness vs. Shared Parameters

A central debate in 3D medical imaging is whether to use a shared backbone (parameter efficient) or specialized backbones (representation rich). We compared our Modality-Aware Experts ($N = 8$) against a Single-Expert ($N = 1$) baseline where all MRI sequences share weights (similar to standard 3D classification approaches). As shown in Table III (Top), the multi-expert design outperforms the shared approach by 4.14% (91.57% vs. 87.43%). **Analysis:** The shared encoder suffers from "negative transfer," as it struggles to optimize simultaneously for contradictory features (e.g., T2 hyper-intensity vs. T1 hypo-intensity). Our expert design disentangles these modalities. Furthermore, combined with our Top-1 routing, we achieve this performance gain without inflating inference cost, offering a superior trade-off compared to soft-weighting or ensemble averaging strategies.

TABLE II
Step-wise Ablation Study of Key Components

Baseline	VLP	CKP	CIFER	Avg. ACC (%)
✓				82.86 ± 0.17
✓	✓			84.22 ± 2.32
✓	✓	✓		87.25 ± 1.63
✓	✓	✓	✓	$\mathbf{91.57 \pm 1.56}$

C. Encoder Architecture Selection

We investigated the choice of text and image backbones (Table III, Bottom). Replacing the general-domain CLIP text encoder with BioBERT yields a gain of approximately 3%. This confirms that medical-specific vocabulary understanding is crucial. Similarly, using the CLIP-Driven Foundation Model as the vision initializer significantly outperforms standard SwinViT, benefiting from its broad pretraining on diverse imaging data.

TABLE III
Analysis of Expert Strategy and Backbone Architectures

Expert Strategy	**ACC (%)**
Single-Expert (Shared Weights)	87.43 ± 1.32
Modality-Aware Experts (Ours)	$\mathbf{91.57 \pm 1.56}$
Backbone (Image + Text)	**ACC (%)**
SwinViT [27] + CLIP [5]	83.23 ± 1.13
SwinViT [27] + BioBERT [23]	84.42 ± 1.53
Foundation [17] + CLIP [5]	88.56 ± 1.26
Foundation [17] + BioBERT [23]	$\mathbf{91.57 \pm 1.56}$

D. Qualitative Visualization

To provide qualitative evidence of the CIFER module's efficacy, we visualize the t-SNE embeddings of the test set in Fig. 4. **Without CIFER (Fig. 4a):** The feature space is cluttered; malignant classes like HCC (blue) and ICC (orange) overlap significantly with benign classes, explaining the lower sensitivity of baseline models. **With CIFER (Fig. 4b):** The manifold structure is transformed. Lesion categories form distinct, compact clusters with wide inter-class margins. This proves that the cross-modal attention successfully filters out noise and refines the decision boundaries.

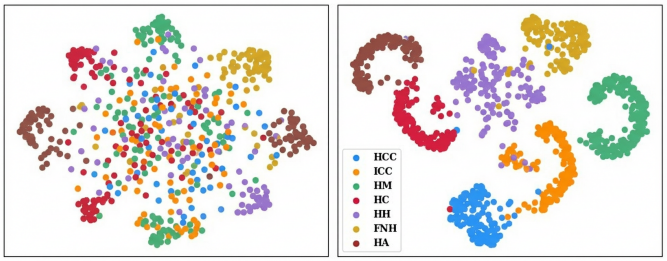

Fig. 4. t-SNE visualization of feature distributions. (a) Without CIFER, classes overlap significantly. (b) With CIFER, the model learns a highly discriminative manifold with clear class separation. Note that both subplots share the common legend indicating the seven lesion categories.

VI. Conclusion

In this paper, we proposed MIRAGE, a modality-aware 3D VLM framework designed for liver lesion diagnosis in multimodal MRI. By leveraging a library of specialized expert encoders and a Modality Routing Gate, our model addresses the inherent heterogeneity of MRI sequences while mimicking the selective diagnostic process of radiologists. Experimental results on the LLD-MMRI dataset demonstrate that MIRAGE significantly outperforms existing methods in classification accuracy and feature representation. Future work will focus on extending this gated expert architecture to longitudinal lesion tracking and broader multi-organ diagnostic tasks.

References

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17–48, 2023.

[2] Z. Fan, M. Jin, L. Zhang *et al.*, "From clinical variables to multiomics analysis: a margin morphology-based gross classification system for hepatocellular carcinoma stratification," *Gut*, vol. 72, no. 11, pp. 2149–2163, 2023.

[3] Y. Zhang, J. Yang, J. Tian *et al.*, "Modality-aware mutual learning for multi-modal medical image segmentation," in *de Bruijne, M., et al. (eds.) International Conference on Medical Image Computing and Computer-Assisted Intervention. LNCS*, vol. 12902. Springer, Cham, 2021, pp. 589–599.

[4] X. Lin, X. Zhou, T. Tong, X. Nie, L. Wang, H. Zheng, J. Li, E. Xue, S. Chen, M. Zheng *et al.*, "A super-resolution guided network for improving automated thyroid nodule segmentation," *Computer Methods and Programs in Biomedicine*, vol. 227, p. 107186, 2022.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[6] H. Lin, C. Xu, and J. Qin, "Taming vision-language models for medical image analysis: A comprehensive review," 2025. [Online]. Available: https://arxiv.org/abs/2506.18378

[7] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[8] L. Wu, J. Zhuang, and H. Chen, "Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 873–22 882.

[9] F. Bai, Y. Du, T. Huang, M. Q.-H. Meng, and B. Zhao, "M3d: Advancing 3d medical image analysis with multi-modal large language models," *arXiv preprint arXiv:2404.00578*, 2024.

[10] X. Zhou, Z. Li, Y. Xue, S. Chen, M. Zheng, C. Chen, Y. Yu, X. Nie, X. Lin, L. Wang *et al.*, "Cuss-net: a cascaded unsupervised-based strategy and supervised network for biomedical image diagnosis and segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 5, pp. 2444–2455, 2023.

[11] X. Wang, H. Ying, X. Xu, X. Cai, and M. Zhang, "Transliver: A hybrid transformer model for multi-phase liver lesion classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 329–338.

[12] Z. Li, X. Zhou, and T. Tong, "Ug-net: unsupervised-guided network for biomedical image segmentation and classification," in *International Conference on Artificial Neural Networks*. Springer, 2023, pp. 197–208.

[13] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri *et al.*, "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[14] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar, "Med-flamingo: a multimodal medical few-shot learner," in *Machine Learning for Health (ML4H)*. PMLR, 2023, pp. 353–367.

[15] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[16] I. E. Hamamci, S. Er, C. Wang, F. Almas, A. G. Simsek, S. N. Esirgun, I. Doga, O. F. Durugol, W. Dai, M. Xu *et al.*, "Developing generalist foundation models from a multimodal dataset for 3d computed tomography," *arXiv preprint arXiv:2403.17834*, 2024.

[17] J. Liu, Y. Zhang, J.-N. Chen, J. Xiao, Y. Lu, B. A Landman, Y. Yuan, A. Yuille, Y. Tang, and Z. Zhou, "Clip-driven universal model for organ segmentation and tumor detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 152–21 164.

[18] C. Li, H. Zhu, R. I. Sultan, H. B. Ebadian, P. Khanduri, C. Indrin, K. Thind, and D. Zhu, "Mulmodseg: Enhancing unpaired multi-modal medical image segmentation with modality-conditioned text embedding and alternating training," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

[19] C. Li, Y. Li, W. Wang, Y. Li, X. Gao, and Y. Zhang, "Gt4-o: Modality-driven heterogeneous graph learning for all-modal biomedical representation," in *European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 321–338.

[20] Y. Xu, F. Zhou, C. Zhao, Y. Wang, C. Yang, and H. Chen, "Distilled prompt learning for incomplete multimodal survival prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 5102–5111.

[21] X. Liang, X. Li, F. Li, J. Jiang, Q. Dong, W. Wang, K. Wang, S. Dong, G. Luo, and S. Li, "Medfilip: medical fine-grained language-image pre-training," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[22] M. Lou, H. Ying, X. Liu, H.-Y. Zhou, Y. Zhang, and Y. Yu, "Sdr-former: A siamese dual-resolution transformer for liver lesion classification using 3d multi-phase imaging," *Neural Networks*, p. 107228, 2025.

[23] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[24] W. Ma, C. Chen, Y. Gong, N. Y. Chan, M. Jiang, C. H.-K. Mak, J. M. Abrigo, and Q. Dou, "Causal effect estimation on imaging and clinical data for treatment decision support of aneurysmal subarachnoid hemorrhage," *IEEE Transactions on Medical Imaging*, 2024.

[25] J. Yu, Q. Hu, M. Jiang, Y. Wang, C. T. Wong, J. Wang, H. Zhang, and Q. Dou, "Language-enhanced local-global aggregation network for multi-organ trauma detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 393–403.

[26] P. Wang, H. Zhang, and Y. Yuan, "Mcpl: Multi-modal collaborative prompt learning for medical vision-language model," *IEEE Transactions on Medical Imaging*, 2024.

[27] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.