PAMN: Multi-phase Correlation Modeling for Contrast-Enhanced 3D Medical Image Retrieval

Haonan Tong [1,2], Ke Liu [1], Chuang Zhang[1*], Xinglin Zhang[3], Tao Chen[4], Jenq-Neng Hwang[5], Lei Li[5,6*]

[1]Inspur Electronic Information Industry Co. Ltd. Beijing, China, [2]Amap, Alibaba Group, [3]Shanghai Medical Image Insights, [4]University of Waterloo, [5]University of Washington, [6]VitaSight

***Correspondence:** zhangchuangbj@ieisystem.com; lenny.lilei.cs@gmail.com

Abstract

Contrast-enhanced 3D Medical imaging (e.g., CT, MRI) leverages phase sequences to uncover temporal dynamics vital for diagnosing tumors, lesions, and vascular issues. However, current retrieval models primarily focus on spatial features, neglecting phase-specific progression detailed in clinical reports. We present the **Phase-aware Memory Network (PAMN)**, a novel framework enhancing 3D medical image retrieval by fusing imaging phases with diagnostic text. PAMN creates rich radiological representations that enhance diagnostic accuracy by combining image details with clinical report context, rigorously tested on a novel phase-series dataset of 12,230 hospital CT scans. PAMN achieves an effective balance of performance and scalability in 3D radiology retrieval, outperforming state-of-the-art baselines through the robust fusion of spatial, temporal, and textual information.

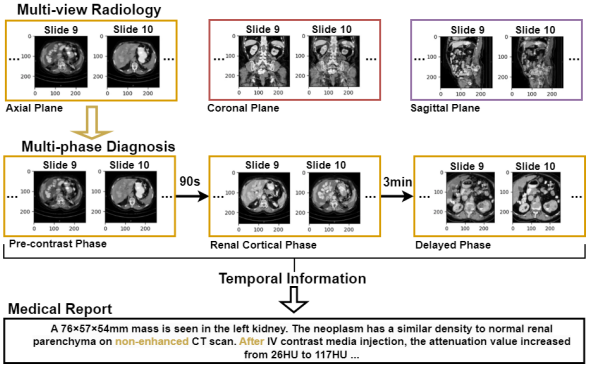

Figure 1: PAMN utilizes the multi-view and multi-phase nature of radiological imaging. It uses temporal contrast changes to align medical images with text, capturing lesion progression across sequential imaging phases for medical diagnosis.

1 Introduction

3D contrast-enhanced imaging is integral to medical diagnostics, enhancing anatomical and pathological visualization beyond non-contrast scans. This technique is particularly valuable in computed tomography (CT) and magnetic resonance imaging (MRI), where contrast agents improve tissue differentiation and vascular visualization, aiding in tumor (Pandit et al., 2025), lesion (Wei et al., 2024), and abnormality detection (Liu et al., 2024b). For comprehensive assessment, 3D contrast-enhanced imaging spans axial, sagittal, and coronal planes, following distinct phases: pre-contrast acquisition, contrast administration, and post-processing (Hsu et al., 2023). Its interpretation requires multidisciplinary collaboration among radiologists, medical physicists, and clinicians to refine diagnoses and guide decisions (Sack, 2023). While deep learning-based evaluation systems (Huang et al., 2025a; Miller et al., 2024) aid decision-making and streamline workflows, a unified approach for effectively leveraging imaging planes and phases in diagnosis is yet to be established.

Multi-view image analysis primarily addresses spatial and geometric relationships between images captured from different perspectives (Wang et al., 2015). These methods—applied in tasks such as video understanding (Li et al., 2025b; Siddiqui et al., 2024; Zhou et al., 2025), 3D rendering (Huang et al., 2025b; Liu et al., 2024a), and segmentation (Qin et al., 2023; Wang et al., 2024; Chen et al., 2025; Li, 2024)—rely on synthesizing geometric features like depth, shape, or texture to enhance scene understanding and object classification (Zhu et al., 2024; Jin et al., 2025a). Multi-scale feature aggregation leverages spatial information across multiple resolutions to extract both fine details (Yu et al., 2024; Wang et al., 2023a; Jia and Li, 2024) and broader contextual cues (Lin et al., 2023a; Cai et al., 2023), while sequential architectures—such as long short-term memory (LSTM) networks (Hong et al., 2023; Tang et al., 2024) and transformer-based models (Dong et al., 2023; Yang et al., 2023)—capture sequential (Yang et al., 2023;

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3456–3467

Peng et al., 2022) or temporal (Alkin et al., 2024; Chang et al., 2024) dependencies within image sequences. Despite these advancements, adapting these techniques to medical imaging, especially for aligning spatial and phase information with textual clinical insights in 3D contrast-enhanced imaging, remains a challenging open problem.

Diagnostic reports offer rich clinical insights from multi-phase medical images by integrating spatial, temporal, and pathological correlations. While vision-language models like BiomedCLIP (Zhang et al., 2023) and LLaVA-Med (Li et al., 2024) excel at joint visual-textual representation learning for tasks like classification and retrieval, and UniMedI (He et al., 2024) and fVLM (Shui et al., 2025) have unified multi-modal images using reports or direct 3D image-text pairing, most current methods overlook crucial temporal information. In multi-phase radiology, temporal dynamics, such as contrast agent progression, are vital for clinical decisions and often detailed in reports. Aligning these reports with specific imaging phases can significantly enhance feature representations in 3D contrast-enhanced imaging.

To address this, we introduce the Phase-aware Memory Network (PAMN), a novel framework that combines phase-specific spatial cues with temporal progression modeling of contrast enhancement, utilizing diagnostic reports to refine feature representations (Figure 2). PAMN comprises three core components:

- **Phase-aware Multiview Modeling.** PAMN organizes imaging phases and classifies anatomical planes to create phase-specific representations, ensuring anatomical consistency while capturing subtle pathological variations.
- **Phase Dynamic Modeling.** PAMN tracks the temporal progression of contrast agents across phases, differentiating tissues based on their unique enhancement patterns.
- **Text-guided Representation Alignment.** PAMN uses diagnostic reports as supervisory cues to model clinically relevant features that align with diagnostic patterns, boosting interpretability and robustness.

2 Related Work

2.1 Multi-view Representation Learning

Multi-view representation learning (MvRL) integrates multiple perspectives into unified representations, enhancing tasks like video understanding (Siddiqui et al., 2024; Li et al., 2025a), 3D rendering (Huang et al., 2025b; Jin et al., 2025b), and segmentation (Qin et al., 2023; Chen et al., 2025). Traditional methods like Canonical Correlation Analysis (CCA) and its deep learning variants (Wang et al., 2015; Yuan et al., 2022) map views to a common subspace but struggle with high-dimensional data due to computational inefficiencies and linearity constraints (Wang et al., 2023b). Deep learning advances, including CNNs (Feng et al., 2018; Sun et al., 2020) and GNNs (Hassani and Khasahmadi, 2020; Xue et al., 2020), enable nonlinear feature fusion but face challenges with redundancy and misalignment. Contrastive learning (Lin et al., 2022; Yu et al., 2022) improves cross-view consistency by aligning shared semantics while retaining view-specific details. Beyond loss design, multi-scale feature aggregation (Yu et al., 2024; Wang et al., 2023a) enhances representation robustness (Lin et al., 2023a; Cai et al., 2023), while sequential modeling with LSTMs (Hong et al., 2023; Tang et al., 2024) and transformers (Dong et al., 2023; Yang et al., 2023, 2025) captures sequential (Peng et al., 2022) or temporal (Alkin et al., 2024; Chang et al., 2024) dependencies, benefiting tasks like 3D reconstruction and forecasting. However, medical imaging poses challenges due to heterogeneity, noise, and the need for interpretability, requiring domain-specific adaptations.

2.2 Medical Image-Text Alignment

Medical image-text alignment, integrating visual data like X-rays, MRIs, and CTs with clinical notes, enhances medical understanding and diagnosis (Zhang et al., 2024; Lu et al., 2024). Pre-trained vision-language models (VLMs) such as BiomedCLIP (Zhang et al., 2023) and LLaVA-Med (Li et al., 2024) support tasks like image retrieval and classification. However, extending this to 3D medical imaging is challenging due to limited annotated 3D volumes. While methods like knowledge distillation (Park et al., 2022) and 2D slice extraction (He et al., 2024; Lin et al., 2023b) have been used, and fVLM (Shui et al., 2025) aligned 3D images with reports using a proprietary dataset, these largely focus on spatial alignment and neglect crucial temporal dynamics like contrast agent progression. No existing public datasets address phase-series imaging, which is vital for modeling temporal contrast variations. To address this gap and enable evaluation of temporal dynamics in multi-

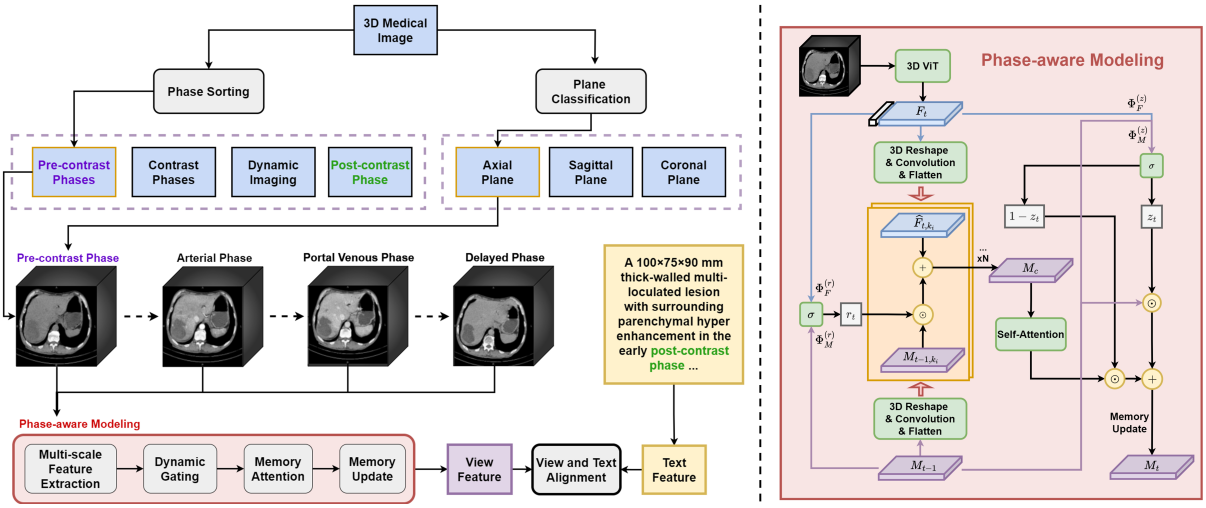

Figure 2: **Overview.** The proposed framework processes 3D medical images by first sorting imaging phases and classifying anatomical planes. The model then extracts phase-specific features from multi-phase imaging sequences, integrating them through multi-scale feature extraction, dynamic gating, memory attention, and memory update mechanisms. Finally, the extracted view features are aligned with corresponding clinical text descriptions to generate text-aligned representations, enabling improved medical image retrieval and interpretation.

phase radiology, we curated our own phase-series dataset for this study.

3 Method

In this section, we present the Phase-aware Memory Network (PAMN), which aligns multi-phase medical images with text by organizing images into phase sequences and extracting multi-scale spatial features (Figure 2).

3.1 Medical Image Organization into Phase Sequence

In our system, 3D medical images are organized into a sequence of imaging phases reflecting contrast administration and clinical needs. Plane classification ensures a uniform axial, sagittal, or coronal view across each series, enabling precise temporal comparisons. The sequence starts with pre-contrast images capturing baseline anatomy without contrast, followed by contrast-enhanced phases—arterial, portal venous, and delayed—highlighting contrast dynamics to distinguish tissues and detect abnormalities like tumors or vascular issues. Next, dynamic phases such as Mean Transit Time (MTT), Cerebral Blood Volume (CBV), and Cerebral Blood Flow (CBF) offer functional insights into circulation and perfusion, distinct from structural contrast data. Post-contrast phases conclude the sequence, providing detailed diagnostic views. This structured approach aligns with clinical protocols, maintaining spatial and temporal coherence for accurate anatomical and functional assessment.

3.2 Multi-scale Feature Extraction

Given a 3D medical image volume or a multi-phase imaging sequence, the raw image tensor is partitioned into non-overlapping patches. Let $I_t \in \mathbb{R}^{C\times D\times H\times W}$ denote the input at phase t, where C is the number of channels, and $D\times H\times W$ the spatial resolution. Feature extraction is performed using a Vision Transformer (ViT):

$$F_t = \Psi(I_t), \quad F_t \in \mathbb{R}^{N\times C'} \tag{1}$$

where $\Psi(\cdot)$ represents patch embedding and contextualization through the ViT, N is the number of patches, and C' is the embedding dimension.

Multi-scale spatial features are extracted using convolutions with kernel sizes $K = \{k_1, k_2, \ldots, k_m\}$. For each k_i, $\Phi_{F,k_i}(\cdot)$ reshapes F_t into a spatial feature map, applies the convolution, and reshapes it back to $N \times C'$. The same process applies to M_{t-1}.

$$\begin{aligned} \hat{F}_{t,k_i} &= \Phi_{F,k_i}(F_t), \\ M_{t-1,k_i} &= \Phi_{M,k_i}(M_{t-1}) \end{aligned} \tag{2}$$

where $\hat{F}_{t,k_i}, M_{t-1,k_i} \in \mathbb{R}^{N\times C'}$. This process captures diverse spatial relationships, enhancing feature robustness for memory updates and alignment.

3.3 Adaptive Gating for Memory Updates

To dynamically control memory updates, we use learnable update (z_t) and reset (r_t) gates:

$$\begin{aligned} z_t &= \sigma\big(\Phi_F^{(z)}(F_t) + \Phi_M^{(z)}(M_{t-1})\big) \odot \theta_z \\ r_t &= \sigma\big(\Phi_F^{(r)}(F_t) + \Phi_M^{(r)}(M_{t-1})\big) \odot \theta_r \end{aligned} \tag{3}$$

Here, $\sigma(\cdot)$ is the sigmoid function. Similar to Φ_{F,k_i}, the functions $\Phi_F^{(z)}(\cdot)$ and $\Phi_M^{(z)}(\cdot)$ reshape F_t and M_{t-1} into spatial feature maps, apply convolutions with distinct kernels for the update gate, and reshape back; $\Phi_F^{(r)}(\cdot)$ and $\Phi_M^{(r)}(\cdot)$ do the same for the reset gate. If an image in the phase series is missing, both z_t and r_t are set to 0 to exclude it from the update. The learnable parameters $\theta_z, \theta_r \in \mathbb{R}^N$ adaptively scale each patch, ensuring consistent feature processing across channels.

3.4 Memory Update with Attention

The model computes a candidate memory state M_c by combining multi-scale feature embeddings with past memory, modulated by the reset gate r_t:

$$M_c = \sum_{i=1}^{m} \tanh(\hat{F}_{t,k_i} + r_t \odot M_{t-1,k_i}) \tag{4}$$

The $\tanh(\cdot)$ function bounds the memory representation to prevent excessive activations, while r_t regulates the influence of past memory M_{t-1,k_i} on the update.

An attention mechanism refines M_c by emphasizing informative spatial regions. A learnable function Φ_{att}, such as a convolution, processes M_c to capture spatial dependencies, and softmax normalization yields attention weights:

$$M_c^{\text{att}} = \text{softmax}(\Phi_{\text{att}}(M_c)) \odot M_c \tag{5}$$

This prioritizes regions critical for robust feature retention and contextual understanding.

The final memory state M_t blends the previous memory M_{t-1} with the refined candidate M_c^{att}, guided by the update gate z_t:

$$M_t = (1 - z_t) \odot M_{t-1} + z_t \odot M_c^{\text{att}} \tag{6}$$

Here, z_t balances the incorporation of new information with continuity from past knowledge.

3.5 Attention Pooling and Multi-Phase Image Representation Projection

At the final phase T, an attention pooling mechanism aggregates the memory state $M_T \in \mathbb{R}^{N \times C'}$ into a compact representation for tasks like text alignment. A learnable function Φ_{pool}, such as a 1×1 convolution, computes attention scores over the N patches, and the aggregated memory $\bar{M}_T$ is:

$$\bar{M}_T = \sum_{i=1}^{N} \text{softmax}(\Phi_{\text{pool}}(M_T))_i \cdot M_{T,i} \tag{7}$$

where $M_{T,i} \in \mathbb{R}^{C'}$ is the i-th patch embedding, and $\bar{M}_T \in \mathbb{R}^{C'}$ weights each patch by its attention score, prioritizing informative regions for retrieval.

This aggregated representation $\bar{M}_T$ is then projected into a joint vision-language embedding space, yielding a normalized feature $\hat{M}$ for text comparison. This ensures scale invariance and enhances multimodal alignment for effective retrieval.

3.6 Text Representation Encoding

For cross-modal retrieval, textual descriptions of medical images are mapped into a shared embedding space with images. Given a tokenized sequence $T \in \mathbb{R}^L$, where L is the sequence length, a pretrained language model (e.g., BERT) generates contextualized features:

$$H_{\text{text}} = \text{BERT}(T) \tag{8}$$

where $H_{\text{text}} \in \mathbb{R}^{L \times C'}$ holds contextual embeddings from the final layer. The special [CLS] token embedding is projected via a learnable function and L2-normalized to yield the final text embedding $\hat{T}$ for alignment with image encodings.

3.7 Contrastive Learning for Modality Alignment

To align multi-phase medical images with their corresponding textual descriptions, we employ contrastive learning in a shared retrieval space. This framework maximizes similarity between matching image-text pairs while separating non-matching ones. Given a batch of B multi-phase image embeddings $\mathbf{M} \in \mathbb{R}^{C' \times B}$ and text embeddings $\mathbf{T} \in \mathbb{R}^{C' \times B}$, where each column represents L2-normalized embeddings (Sections 3.5 and 3.6), the similarity matrix $S \in \mathbb{R}^{B \times B}$ is computed as:

$$S = \mathbf{M}^\top \cdot \mathbf{T}, \tag{9}$$

with $S_{i,j}$ measuring the similarity between phase sequence i and diagnosis report j. The contrastive loss is:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{i,i}/\tau)}{\sum_{j=1}^{B}\exp(S_{i,j}/\tau)}, \quad (10)$$

with τ as a learnable temperature parameter. This loss maximizes similarity for matching pairs $S_{i,i}$ while minimizing it for non-matching pairs $S_{i,j}$ ($j \neq i$), leveraging in-batch negatives to enable effective cross-modal retrieval.

4 Experiments

4.1 Dataset

Hospital CT Scans. Computed Tomography (CT) images were acquired using the DISCOVERY CT750 HD FREEDOM system at Dongfang Hospital. In 2024, this study included 61,332 patient cases, with scans adhering to standardized imaging protocols for consistent, high-quality data. Anatomical regions evaluated comprised the head and brain, chest, abdomen, pelvis, spine and bones, soft tissues and vasculature, and limb joints. Post-acquisition, images were digitally archived and organized for subsequent analysis.

CT Phase Datasets. The 3D medical images were grouped into phase-series based on contrast administration and diagnostic requirements. A pre-contrast phase captured native anatomy, followed by contrast-enhanced phases to highlight dynamics, aiding detection of abnormalities such as tumors or vascular issues. The dataset, segmented by timing and contrast, yielded 12,230 phase-series samples: 7,142 with two phases, 3,451 with three, and 1,637 with four.

To illustrate the utility of multi-phase imaging, Figure 3 presents a representative case comparing non-contrast and delayed contrast-enhanced phases. The wedge-shaped hypodense lesion in the right hepatic lobe shows persistent hypoenhancement, suggesting heterogeneous fatty infiltration or perfusion alterations. A dense focus in the left distal ureter, consistent with a calculus, is more conspicuous post-contrast due to improved delineation of urinary structures. Subtle renal micro-calcifications are also visible across phases. These observations highlight how contrast-enhanced imaging enhances diagnostic accuracy in evaluating hepatic perfusion and urinary obstruction.

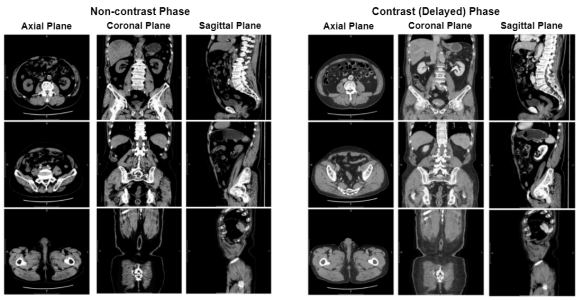

Figure 3: Comparison of non-contrast and delayed contrast-enhanced CT images highlights a wedge-shaped hepatic lesion with persistent hypoenhancement and improved visualization of a left ureteral stone with upstream dilation, underscoring the diagnostic value of contrast enhancement in characterizing hepatic perfusion and urinary obstruction.

4.2 Results

We compare the retrieval performance of PAMN variants (with and without temporal data) against 2D and 3D baselines: PMC-CLIP (Lin et al., 2023b) and a self-implemented 3D vanilla baseline for 3D medical image retrieval.

4.2.1 Quantitative Performance on Phase-Series Dataset

Table 1 presents the performance of three models on the Phase-Series dataset: 3D Vanilla Baseline, 3D Baseline ST, and PAMN. The 3D Vanilla Baseline is a self-implemented model for 3D medical image retrieval, focusing on pairwise image-text alignment without specialized phase-series handling. 3D Baseline ST, a variant of this baseline, simplifies the input by sampling a single image from the phase series, representing a single temporal snapshot. In contrast, PAMN, our proposed model, dynamically models phase feature interactions, leveraging multi-phase information to enhance retrieval performance. PAMN consistently outperforms both variants of the 3D vanilla baseline across all test sizes and retrieval tasks.

For example, in the image retrieval task with 100 test samples, the standard 3D Vanilla Baseline achieves an IR R@10 of 30.77%, while 3D Baseline ST scores 62.00%. PAMN outperforms both by reaching an IR R@10 of 76.00%—this represents an improvement of 45.23 percentage points over 3D Vanilla Baseline and 14.00 points over 3D Baseline ST. A similar pattern is observed in the text retrieval task, where the TR R@10 scores are 31.62% for 3D Vanilla Baseline, 62.00% for 3D Baseline ST, and 73.00% for PAMN.

Table 1: Comparison of PMC-CLIP, 3D Vanilla Baseline, and PAMN on Phase-Series and Caption Datasets.

Test Samples	Metric	Phase-Series			Caption				
		3D Baseline	3D Baseline ST	PAMN	PMC-CLIP	3D Baseline L	3D Baseline	PAMN w/o T	PAMN w/ T
100	IR R@1	7.69	18.00	24.00	9.00	32.00	64.00	32.00	64.00
	IR R@5	17.52	42.00	60.00	28.00	63.00	95.00	63.00	97.00
	IR R@10	30.77	62.00	76.00	45.00	73.00	99.00	78.00	99.00
	TR R@1	4.70	15.00	21.00	18.00	30.00	70.00	31.00	73.00
	TR R@5	20.51	42.00	58.00	47.00	68.00	95.00	65.00	97.00
	TR R@10	31.62	62.00	73.00	59.00	80.00	98.00	80.00	98.00
500	IR R@1	2.67	6.20	7.20	4.40	14.60	39.60	13.80	52.00
	IR R@5	7.14	19.60	25.00	12.80	34.40	76.20	34.40	82.40
	IR R@10	12.47	30.20	39.00	18.80	47.80	87.20	46.40	90.00
	TR R@1	1.98	6.00	7.20	7.60	14.20	40.40	15.00	48.80
	TR R@5	7.82	19.20	24.60	20.20	38.00	74.20	33.80	81.20
	TR R@10	11.44	28.60	38.00	31.00	52.20	87.00	49.00	88.80
1000	IR R@1	1.57	3.60	5.10	1.90	9.00	27.30	8.10	39.10
	IR R@5	4.00	11.50	17.30	7.60	24.60	61.10	23.40	72.00
	IR R@10	7.05	19.90	25.70	12.10	34.80	76.10	33.00	84.30
	TR R@1	1.39	3.30	5.60	4.60	9.20	26.60	8.90	37.20
	TR R@5	4.18	11.60	17.00	13.00	26.70	61.80	24.00	72.40
	TR R@10	7.66	18.80	27.00	19.80	38.40	75.30	35.50	84.00
2000	IR R@1	1.04	2.40	3.35	1.15	8.10	19.10	8.00	30.75
	IR R@5	2.49	7.70	11.85	4.35	20.20	47.45	19.05	61.70
	IR R@10	4.43	11.70	17.95	7.60	28.40	62.25	27.80	75.80
	TR R@1	0.80	1.95	3.45	3.15	8.50	18.45	7.55	29.70
	TR R@5	2.72	6.96	11.25	8.55	21.75	47.30	19.80	61.70
	TR R@10	4.67	11.60	17.30	13.55	30.95	62.15	28.75	74.70

As the number of test samples increases, overall retrieval performance declines—a reflection of the challenges inherent in large-scale retrieval. Nevertheless, PAMN demonstrates a more controlled degradation. For example, at 500 test samples, the TR R@10 drops to 11.44% for 3D Vanilla Baseline and 28.60% for 3D Baseline ST, while PAMN still manages 38.00%. At 1000 samples, PAMN attains an IR R@10 of 25.70% and a TR R@10 of 27.00%, significantly outperforming 3D Vanilla Baseline's 7.05% and 7.66%, with 3D Baseline ST achieving intermediate scores of 19.90% and 18.80%. Even at the largest scale of 2000 samples, PAMN yields 17.95% (IR R@10) and 17.30% (TR R@10), compared to 4.43% and 4.67% for 3D Vanilla Baseline, with 3D Baseline ST reaching 11.70% and 11.60%.

Despite these improvements, overall recall values remain relatively low, likely due to the limited number of training samples. This data dependency suggests that both the 3D vanilla baseline and PAMN could benefit from larger-scale training or pretraining on more diverse multimodal medical datasets. In Section 4.3.1, we further demonstrate that PAMN adheres to data scaling laws, supporting our hypothesis that increased data availability leads to improved model performance.

4.2.2 Quantitative Performance on General Caption Dataset

We further evaluate the generalizability of PAMN for 3D medical image retrieval, regardless of phase-series availability (Table 1). We compare PAMN against PMC-CLIP, a 2D-based medical image analysis model, and the 3D vanilla baseline, the state-of-the-art for 3D medical image retrieval. Additionally, we assess the impact of patch size variations in the 3D vanilla baseline by comparing 3D Baseline L (8×32×32) and 3D Baseline (4×16×16) as baselines for retrieval performance at different levels of spatial granularity. Our study focuses on two PAMN variations: PAMN w/o T, trained on a dataset without phase series, and PAMN w/ T, which incorporates phase series during training.

The results reveal that PMC-CLIP underperforms significantly compared to 3D retrieval models, underscoring the fundamental differences between 2D and 3D medical image retrieval. Unlike natural images, where 2D features often suffice for analysis, medical images contain volumetric structures that require models capable of extracting spatial features across multiple slices. PMC-CLIP lacks spatial encoding, leading to a substantial drop in retrieval accuracy. For example, at 100 test samples, PMC-CLIP achieves an IR R@1 of only 9.00%, while both 3D Baseline (4×16×16) and PAMN w/ T reach 64.00%. Similarly, at 1000 test samples, PMC-CLIP's IR R@1 drops to 1.90%, whereas 3D Baseline (4×16×16) achieves 27.30% and PAMN w/ T further improves to 39.10%. These results confirm that retrieval in 3D medical imaging necessitates models that

can process volumetric information, as 2D-based retrieval fails to capture depth and structural continuity.

Beyond demonstrating the necessity of spatial information, our experiments show that PAMN w/o T generalizes effectively to datasets that do not include temporal phase series, indicating that it is not dependent on temporal variations to achieve strong performance. When trained on a dataset without phase information, PAMN w/o T achieves retrieval performance comparable to 3D Baseline L (8×32×32), suggesting that it can successfully converge on static 3D datasets without requiring explicit temporal cues. At 500 test samples, PAMN w/o T attains IR R@1 = 13.80%, IR R@5 = 34.40%, and IR R@10 = 46.40%, closely matching 3D Baseline L (8×32×32), which achieves 14.60%, 34.40%, and 47.80%, respectively. The similarity in performance suggests that PAMN effectively learns spatial representations without phase data, making it well-suited for single-phase datasets where temporal variation is unavailable.

Although PAMN w/o T performs well on static datasets, our results indicate that incorporating phase series during training significantly enhances retrieval performance, as demonstrated by PAMN w/ T. When trained with both the baseline dataset and additional phase series, PAMN w/ T consistently outperforms 3D Baseline (4×16×16), showing that temporal diversity improves retrieval performance. The benefits of phase series are particularly evident in larger test sets, where additional temporal variations enable the model to learn richer feature representations, enhancing generalization. For example, at 500 test samples, PAMN w/ T achieves IR R@1 = 52.00%, IR R@5 = 82.40%, and IR R@10 = 90.00%, whereas 3D Baseline (4×16×16) reaches only 39.60%, 76.20%, and 87.20%, respectively. A similar trend appears at 1000 test samples, where PAMN w/ T achieves IR R@1 = 39.10%, significantly surpassing 3D Baseline (4×16×16), which only reaches 27.30%. These findings suggest that phase series contribute additional diversity to the dataset, improving the model's ability to differentiate similar images and leading to better retrieval outcomes.

4.3 Ablation Study

In this section, we evaluate the critical factors affecting PAMN's retrieval performance through a series of controlled experiments, analyzing the influence of dataset scale, feature aggregation strategies, and phase-series length on retrieval accuracy.

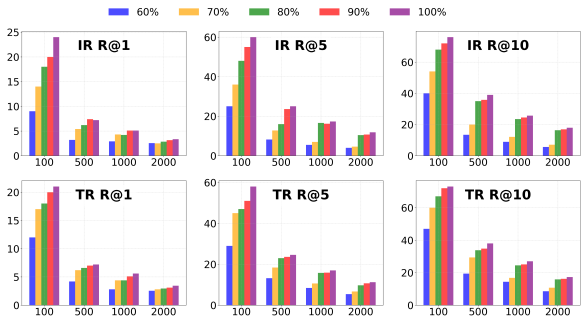

Figure 4: Impact of dataset size on retrieval performance.

4.3.1 Data Scale Law

We evaluate the impact of dataset size on retrieval performance by training PAMN on five different scales (60%–100%) of the full phase-series dataset. As shown in Figure 4, retrieval accuracy improves consistently with larger training data, confirming the strong dependence on the scale of the dataset. For the 100-sample test set, IR R@10 improves from 40.00% (60% data) to 76.00% (full data), while TR R@10 increases from 47.00% to 73.00%.

In particular, PAMN matches or exceeds 3D Baseline ST on the training scale 80%, with PAMN achieving IR R@10 of 23.50% and TR R@10 of 24.50% compared to 3D Baseline ST's 19.90% and 18.80%, respectively, on the test set of 1000 samples. These findings reinforce the data-intensive nature of retrieval models, validating our hypothesis from Section 4.2.1 that increased data availability enhances model performance and supports PAMN's scalability under larger training regimes.

4.3.2 Pooling Aggregation

We assess several fixed pooling aggregation strategies to underscore the importance of integrating phase knowledge. As shown in Figure 5, even these naive methods—mean (MEAN), max (MAX), and attention-based (ATTENTION) pooling—that statically aggregate phase features lead to significant improvements over the baseline 3D Vanilla Baseline model, which lacks any explicit phase feature integration.

On the 100-sample test set, the baseline model achieves an IR R@10 of 30.77% and a TR R@10 of 31.62%. Fixed pooling methods improve these metrics notably: IR R@10 increases to 51.00% (MEAN) and 58.00% (MAX), while TR R@10 reaches 63.00% (MEAN) and 72.00% (MAX).

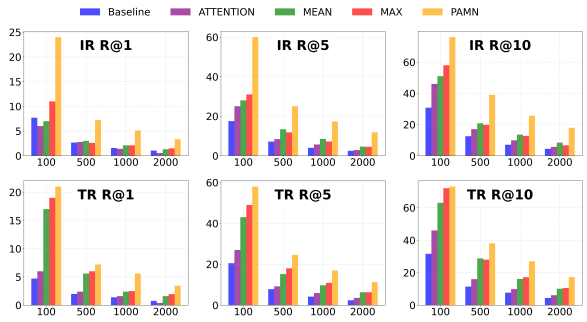

Figure 5: Impact of feature aggregation strategies on retrieval performance.

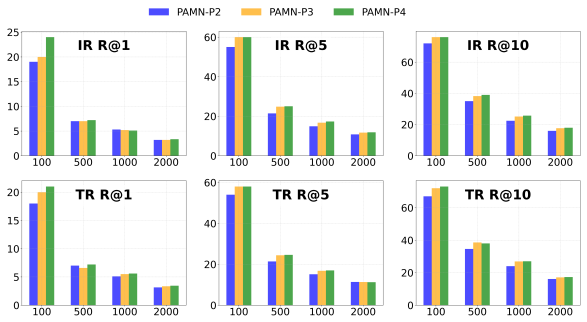

Figure 6: Retrieval performance of PAMN with varying phase-series lengths.

Although ATTENTION pooling offers moderate gains (46.00% for both IR and TR R@10), the benefits of aggregating phase features are clear.

More importantly, PAMN—which dynamically models phase feature interactions—outperforms all fixed pooling strategies across test sizes. On the 100-sample test set, PAMN records an IR R@10 of 76.00% and a TR R@10 of 73.00%. In large-scale retrieval (2000-sample test set), fixed pooling methods yield IR R@10 values of 8.40% (MEAN), 6.70% (MAX), and 5.65% (ATTENTION), whereas PAMN achieves 17.95%. Similarly, for TR R@10, PAMN attains 17.30% compared to 10.15% (MEAN), 10.60% (MAX), and 6.10% (ATTENTION).

These results reveal two critical insights. First, integrating phase knowledge via even simple aggregation boosts recall significantly over phase-agnostic baselines. Second, PAMN's adaptive modeling of phase interactions consistently surpasses fixed strategies, ensuring robust retrieval across scales. This underscores the necessity of dynamic, phase-aware approaches for advancing diagnostic performance.

4.3.3 Adaptive Series Length

We evaluated PAMN's robustness to varying phase-series lengths, testing a single trained model on datasets with 2, 3, or 4 imaging phases. In the 2000-sample set, for instance, there were 1219 samples for 2-phase, 495 for 3-phase, and 286 for 4-phase evaluations. Figure 6 demonstrates that increasing the number of phases generally improves retrieval performance, highlighting the benefit of richer temporal representations. For the 100-sample test set, the impact of additional phases was clear: IR R@1 increased from 19.00% for PAMN-P2 to 24.00% for PAMN-P4, while TR R@10 rose from 67.00% to 73.00%. These improvements show that incorporating more phase information can significantly boost recall.

However, in larger test sets, these gains tended to saturate. For example, in the 2000-sample set, IR R@10 improved by 1.70 percentage points from PAMN-P2 to PAMN-P3, but only by an additional 0.30 percentage points for PAMN-P4. A similar pattern was observed with TR R@10 and in the 1000-sample test set. This saturation likely results from the reduced sample size for the 4-phase configuration combined with the inherent challenges of large-scale retrieval. As noted in Section 4.2.2, PAMN without phase training defaults to baseline spatial features. The results here further confirm that adding phases consistently enhances recall, though gains depend on data distribution, underscoring PAMN's reliance on robust, balanced datasets for optimal temporal modeling.

5 Conclusion

This study introduces the Phase-Aware Memory Network (PAMN), a novel framework enhancing 3D medical image retrieval by integrating multi-view spatial alignment, multi-phase temporal dynamics, and diagnostic text context. Unlike conventional methods, PAMN captures contrast-phase progression, aligning it with clinical reports to produce richer, clinically relevant representations. Experiments on a 12,230-sample CT dataset show PAMN outperforming 2D and 3D baselines, with up to 45% recall@10 improvement and robust scalability. Ablations confirm the value of temporal cues and dynamic feature fusion. PAMN's adaptability to single-phase datasets highlights its versatility, paving the way for intelligent, context-driven radiological analysis, with potential for automated diagnostics and broader clinical applications.

6 Limitations

Despite PAMN's promising performance, this study faces limitations primarily stemming from the relatively limited scale and diversity of available multi-phase medical imaging data. The dataset curated from hospital CT scans, although extensive, may not comprehensively capture the full variability encountered across different institutions or imaging protocols, potentially restricting the model's generalizability. Furthermore, the performance gains demonstrated by PAMN highlight its data-dependent nature, suggesting that future research would benefit from larger-scale and multi-institutional datasets encompassing broader anatomical contexts and varied clinical scenarios. Nevertheless, this work establishes a robust framework for phase-aware image retrieval and lays critical groundwork for future advancements in multi-phase medical retrieval and automated clinical report generation.

7 Acknowledgments

This work was supported by Shangdong Provincial Natural Science Foundation (Grant No. ZR2023LZH014).

References

Benedikt Alkin, Maximilian Beck, Korbinian Pöppel, Sepp Hochreiter, and Johannes Brandstetter. 2024. Vision-lstm: xlstm as generic vision backbone. *arXiv preprint arXiv:2406.04303*.

Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. 2023. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17302–17313.

Ming Chang, Xishan Zhang, Rui Zhang, Zhipeng Zhao, Guanhua He, and Shaoli Liu. 2024. Recurrentbev: A long-term temporal fusion framework for multi-view 3d detection. In *European Conference on Computer Vision*, pages 131–147. Springer.

Mu Chen, Liulei Li, Wenguan Wang, Ruijie Quan, and Yi Yang. 2025. General and task-oriented video segmentation. In *European Conference on Computer Vision*, pages 72–92. Springer.

Haotian Dong, Enhui Ma, Lubo Wang, Miaohui Wang, Wuyuan Xie, Qing Guo, Ping Li, Lingyu Liang, Kairui Yang, and Di Lin. 2023. Cvsformer: Cross-view synthesis transformer for semantic scene completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8874–8883.

Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. 2018. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 264–272.

Kaveh Hassani and Amir Hosein Khasahmadi. 2020. Contrastive multi-view representation learning on graphs. In *International conference on machine learning*, pages 4116–4126. PMLR.

Xiaoxuan He, Yifan Yang, Xinyang Jiang, Xufang Luo, Haoji Hu, Siyun Zhao, Dongsheng Li, Yuqing Yang, and Lili Qiu. 2024. Unified medical image pre-training in language-guided common semantic space. In *European Conference on Computer Vision*, pages 123–139. Springer.

Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2023. 3d concept learning and reasoning from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9202–9212.

Jessica C Hsu, Zhongmin Tang, Olga E Eremina, Alexandros Marios Sofias, Twan Lammers, Jonathan F Lovell, Cristina Zavaleta, Weibo Cai, and David P Cormode. 2023. Nanomaterial-based contrast agents. *Nature Reviews Methods Primers*, 3(1):30.

Hongyan Huang, Junyang Mo, Zhiguang Ding, Xuehua Peng, Ruihao Liu, Danping Zhuang, Yuzhong Zhang, Genwen Hu, Bingsheng Huang, and Yingwei Qiu. 2025a. Deep learning to simulate contrast-enhanced mri for evaluating suspected prostate cancer. *Radiology*, 314(1):e240238.

Zanming Huang, Jimuyang Zhang, and Eshed Ohn-Bar. 2025b. Neural volumetric world models for autonomous driving. In *European Conference on Computer Vision*, pages 195–213. Springer.

Sen Jia and Lei Li. 2024. Adaptive masking enhances visual grounding. *arXiv preprint arXiv:2410.03161*.

Can Jin, Tianjin Huang, Yihua Zhang, Mykola Pechenizkiy, Sijia Liu, Shiwei Liu, and Tianlong Chen. 2025a. Visual prompting upgrades neural network sparsification: A data-model perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4111–4119.

Can Jin, Ying Li, Mingyu Zhao, Shiyu Zhao, Zhenting Wang, Xiaoxiao He, Ligong Han, Tong Che, and Dimitris N Metaxas. 2025b. Lor-vp: Low-rank visual prompting for efficient vision model adaptation. In *The Thirteenth International Conference on Learning Representations*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Lei Li. 2024. Cpseg: Finer-grained image semantic segmentation via chain-of-thought language prompting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 513–522.

Lei Li, Sen Jia, Jianhao Wang, Zhaochong An, Jiaang Li, Jenq-Neng Hwang, and Serge Belongie. 2025a. Chatmotion: A multimodal multi-agent for human motion analysis. *arXiv preprint arXiv:2502.18180*.

Lei Li, Sen Jia, Jianhao Wang, Zhongyu Jiang, Feng Zhou, Ju Dai, Tianfang Zhang, Zongkai Wu, and Jenq-Neng Hwang. 2025b. Human motion instruction tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17582–17591.

Weifeng Lin, Ziheng Wu, Jiayu Chen, Jun Huang, and Lianwen Jin. 2023a. Scale-aware modulation meet transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6015–6026.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461.

Libin Liu, Shen Chen, Sen Jia, Jingzhe Shi, Zhongyu Jiang, Can Jin, Wu Zongkai, Jenq-Neng Hwang, and Lei Li. 2024a. Graph canvas for controllable 3d scene generation. *arXiv preprint arXiv:2412.00091*.

Yuxia Liu, Duyang Gao, Yuanyuan He, Jing Ma, Suet Yen Chong, Xinyi Qi, Hui Jun Ting, Zichao Luo, Zhigao Yi, Jingyu Tang, and 1 others. 2024b. Single-point mutated lanmodulin as a high-performance mri contrast agent for vascular and kidney imaging. *Nature Communications*, 15(1):9834.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, and 1 others. 2024. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874.

Robert JH Miller, Aditya Killekar, Aakash Shanbhag, Bryan Bednarski, Anna M Michalowska, Terrence D Ruddy, Andrew J Einstein, David E Newby, Mark Lemley, Konrad Pieszko, and 1 others. 2024. Predicting mortality from ai cardiac volumes mass and coronary calcium on chest computed tomography. *Nature Communications*, 15(1):2747.

AS Pandit, A Keenlyside, DZ Khan, G Reischer, MA Kamal, N Yoh, Z Jaunmuktane, A Borg, NL Dorward, SE Baldeweg, and 1 others. 2025. Mapping pituitary neuroendocrine tumors: An annotated mri dataset profiling tumor and carotid characteristics. *Scientific Data*, 12(1):80.

Sangjoon Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, Chang Min Park, and Jong Chul Ye. 2022. Self-evolving vision transformer for chest x-ray diagnosis through knowledge distillation. *Nature communications*, 13(1):3848.

Kebin Peng, Rifatul Islam, John Quarles, and Kevin Desai. 2022. Tmvnet: Using transformers for multi-view voxel-based 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 222–230.

Zheyun Qin, Xiankai Lu, Xiushan Nie, Dongfang Liu, Yilong Yin, and Wenguan Wang. 2023. Coarse-to-fine video instance segmentation with factorized conditional appearance flows. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1192–1208.

Ingolf Sack. 2023. Magnetic resonance elastography from fundamental soft-tissue mechanics to diagnostic imaging. *Nature Reviews Physics*, 5(1):25–42.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and 1 others. 2025. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. *arXiv preprint arXiv:2501.14548*.

Nyle Siddiqui, Praveen Tirupattur, and Mubarak Shah. 2024. Dvanet: Disentangling view and action features for multi-view action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4873–4881.

Kai Sun, Jiangshe Zhang, Junmin Liu, Ruixuan Yu, and Zengjie Song. 2020. Drcnn: Dynamic routing convolutional neural network for multi-view 3d object recognition. *IEEE Transactions on Image Processing*, 30:868–877.

Yujin Tang, Peijie Dong, Zhenheng Tang, Xiaowen Chu, and Junwei Liang. 2024. Vmrnn: Integrating vision mamba and lstm for efficient and accurate spatiotemporal forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5663–5673.

Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. 2023a. Omni aggregation networks for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22387.

Ren Wang, Haoliang Sun, Yuling Ma, Xiaoming Xi, and Yilong Yin. 2023b. Metaviewer: Towards a unified multi-view representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11590–11599.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR.

Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. 2024. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*.

Yi Wei, Meiyi Yang, Meng Zhang, Feifei Gao, Ning Zhang, Fubi Hu, Xiao Zhang, Shasha Zhang, Zixing Huang, Lifeng Xu, and 1 others. 2024. Focal liver lesion diagnosis with deep learning and multistage ct imaging. *Nature communications*, 15(1):7040.

Fei Xue, Xin Wu, Shaojun Cai, and Junqiu Wang. 2020. Learning multi-view camera relocalization with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11372–11381. IEEE.

Chen Yang, Yangfan He, Aaron Xuxiang Tian, Dong Chen, Jianhui Wang, Tianyu Shi, Arsalan Heydarian, and Pei Liu. 2025. Wcdt: World-centric diffusion transformer for traffic scene generation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6566–6572. IEEE.

Liying Yang, Zhenwei Zhu, Xuxin Lin, Jian Nong, and Yanyan Liang. 2023. Long-range grouping transformer for multi-view 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18257–18267.

En Yu, Zhuoling Li, and Shoudong Han. 2022. Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8834–8843.

Qian Yu, Xiaoqi Zhao, Youwei Pang, Lihe Zhang, and Huchuan Lu. 2024. Multi-view aggregation network for dichotomous image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3921–3930.

Yun-Hao Yuan, Jin Li, Yun Li, Jipeng Qiang, Yi Zhu, Xiaobo Shen, and Jianping Gou. 2022. Learning canonical f-correlation projection for compact multiview representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19260–19269.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, and 1 others. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. 2025. Reagent-v: A reward-driven multi-agent framework for video understanding. *arXiv preprint arXiv:2506.01300*.

Ye Zhu, Yu Wu, Nicu Sebe, and Yan Yan. 2024. Vision+ x: A survey on multimodal learning in the light of data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

A Implementation Details

Input 3D phase images are normalized via Min-Max Normalization and resized to $32 \times 256 \times 256$. A 3D Vision Transformer (ViT) with 12 layers and a patch size of $8 \times 32 \times 32$ extracts phase embeddings, producing a 257×768 output (1 [CLS] token and 256 patch tokens, each 768-dimensional). After PAMN encoding, the phase-series tokens are processed by a pretrained BERT model with 12 layers. Text inputs are capped at 512 tokens, with the [CLS] token serving as the global feature for alignment. Training occurs on an Inspur NF5468M6 server over 50 epochs with a batch size of 32, using 8 NVIDIA A100 GPUs and DeepSpeed for bf16 mixed-precision training. We use the AdamW optimizer with weight decay, an initial learning rate of 10^{-4}, a warm-up phase, and a cosine decay schedule. Retrieval performance is evaluated using Recall@k (R@1, R@5, R@10), where R@1 measures top-1 accuracy and R@5/R@10 assess broader retrieval effectiveness.

B Attention on Phase Series

PAMN fuses multi-phase medical images with diagnostic text via a tiered attention framework—local (within phases), sequential (across phases), and global—to build a unified diagnostic representation. For instance, in the arterial phase, PAMN captures vascular enhancement outlining liver abscesses, while later phases reveal mucosal hyperenhancement and fat stranding linked to colitis, tracking contrast progression and subtle tissue shifts (Figure 7).

A global attention mechanism integrates these phase-specific insights, aligning spatial abnormalities with clinical report semantics. The resulting attention maps spotlight key features—liver lesions, colonic edema, pleural and pericardial effusions—enhancing diagnostic focus.

PAMN's visual features are aligned with text features extracted by a language model, emphasizing key terms—such as peripherally enhancing, abscess, circumferential hyperenhancement,

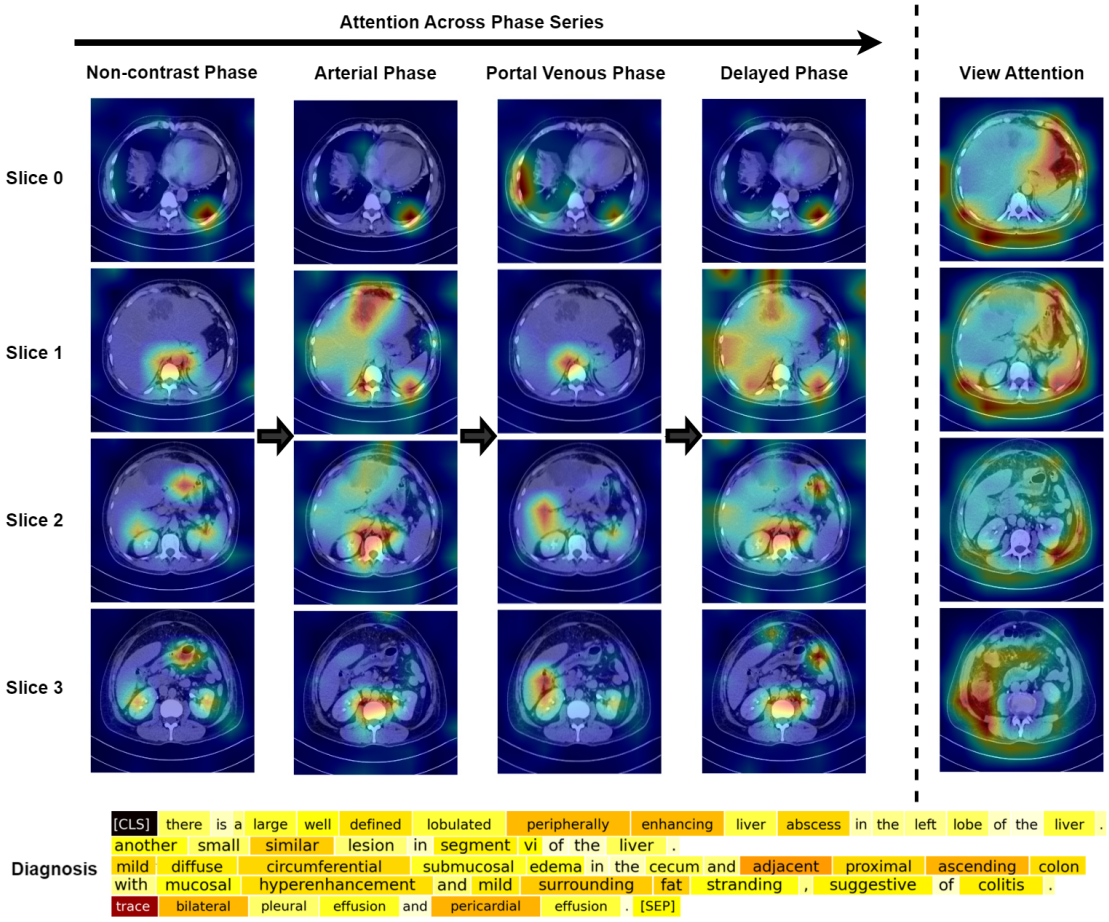

Figure 7: **Feature Alignment.** Attention maps from PAMN across contrast phases for representative axial slices (rows), with the final column (View Attention) integrating multi-phase focus. The corresponding diagnostic text highlights key terms (e.g., peripherally enhancing) attended by the language model, with warmer heatmap colors indicating higher attention weights.

fat stranding, and pericardial effusion—that are attended to by the language model (see Figure 7). These terms directly correspond to imaging characteristics observed in specific contrast phases: in the arterial phase, peripherally enhancing and abscess reflect the contrast-defined rim of a lesion; in later phases, circumferential hyperenhancement and fat stranding signify inflammatory changes tied to colitis; and pericardial effusion is identified through the global integration of multi-phase data. This precise alignment of semantic and imaging cues amplifies diagnostic accuracy and enhances the interpretability of complex 3D medical image analysis.