MEDICAL-4D: 4D CARDIAC MRI VISION-LANGUAGE MODELS FOR CARDIAC FUNCTION UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cardiac magnetic resonance (CMR) cine imaging provides rich 4D information for assessing cardiac function, underpinning clinical diagnosis. However, jointly modeling spatial, temporal, and inter-slice dependencies over full 4D CMR data is computationally prohibitive. Existing CMR-based diagnostic methods simplify the data via temporal downsampling or single-slice videos, discarding fine-grained cardiac motion and inter-slice anatomical dependencies essential for dynamic analysis. To address this, we propose factorized 4D representation learning with cardiac-phase and anatomical-slice positional encoding. Specifically, the factorized encoder separately performs spatial, temporal, and inter-slice modeling to enable joint spatial-temporal analysis of cardiac function. We further introduce continuous structured visual masking to capture dependencies across adjacent cardiac phases and neighboring slices. Together, these components align visual representations with clinical cardiac function assessment reports. To our knowledge, this is the first cardiac model to exploit the full 4D CMR data for continuous and complete vision-language alignment. Based on this model, we propose Medical-4D, the first cardiac function QA framework. Trained on CMR 4D clinical reports and QA pairs from three data centers, Medical-4D significantly outperforms previous state-of-the-art methods.

1 INTRODUCTION

Cardiac magnetic resonance imaging (CMR) is the gold standard for diagnosing cardiac diseases Wang et al. (2024). Short-axis cine sequences capture ventricular anatomy across slices and throughout the cardiac cycle. Based on these 4D data, comprehensive spatial and temporal information is available for assessing cardiac function, which in turn supports disease diagnosis Qiao et al. (2025). However, assessing cardiac function is time-consuming and subject to inter-observer variability Fu et al. (2026), urgently calling for dedicated question answering (QA) systems that enable automated cardiac image analysis and interactive cardiac function assessment. Yet visual QA for cardiac function diagnosis remains largely unexplored, despite the remarkable promise of conversational generative AI for empowering biomedical practitioners. In fact, the correctness of QA answers relies heavily on the underlying vision-language model. Developing dedicated CMR 4D vision-language models is thus a key step toward filling this gap in cardiac function QA.

Training such a vision-language model to achieve fine-grained alignment with clinical cardiac function assessment reports in both spatial and temporal semantics requires full 4D CMR data.Specifically, cardiac cine MRI characterizes cardiac function from two dimensions—motion and anatomy: it captures myocardial contraction and relaxation throughout the cardiac cycle, and reflects myocardial thickness across slices as well as the size and shape of the ventricular chambers. However, jointly modeling spatial, temporal, and inter-slice dependencies over full 4D CMR data makes computation prohibitively expensive. For example, a short-axis cine sequence typically contains more than 5 z-axis slices and 25–30 cardiac phases.

Existing CMR-based representation learning methods Fu et al. (2026); Nakashima et al. (2026) fail to fully exploit this 4D information; instead, they simplify the data via temporal downsampling, single-slice videos, or static 3D volumes, and adapt these reduced data to video or 3D models, rather than developing dedicated CMR-4D models. CMR-Transformer Shad et al. (2026) extracts slice-wise cardiac videos, concatenates them along the temporal dimension, and subsamples them

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to a fixed frame count for video pretraining models. This obviously loses temporal information and disrupts anatomical continuity along the z-axis. CMR-CLIP (Nakashima et al., 2026) directly and randomly selects a single frame from the cardiac phase cycle and discards all the remaining temporal frames, and then feeds the extracted 3D cardiac data into a video model to learn visual representations. CineMA (Fu et al., 2026), on the other hand, takes only two time points from the cardiac phase cycle, namely end-systole and end-diastole. The above approaches generally aim to adapt CMR-4D data to general-purpose medical vision-language models or general-purpose video models.

In contrast, we aim to develop medical 4D vision-language models specifically tailored to CMR. we propose a factorized 4D representation. The 4D encoder factorizes spatial, temporal, and inter-slice modeling to accommodate the large token space of 4D CMR. By introducing temporal attention to integrate cardiac phase features within each slice, followed by slice attention to integrate anatomical context within each phase. This factorization strategy effectively enables joint spatial-temporal learning. In addition, we introduce position encodings normalized by cardiac slice spacing and cardiac phase spacing, respectively, with encodings aligned with anatomical structure and temporal order.

Another difficulty in 4D CMR vision-language learning is that clinical cardiac function assessment reports are sparse supervision signals and are typically a summary of cardiac function compared with full 4D CMR data, which typically contains more than 300 individual images. This poses a central challenge for 4D CMR representation learning: how to design an effective strategy that encourages the vision-language model to focus on inter-frame continuity and relationships, especially continuity along the cardiac phase dimension.

Masked modeling requires models to infer missing information from context (He et al., 2022; Devlin et al., 2019). These advances have motivated the use of masking in contrastive pretraining, where jointly modeling masked images (Li et al., 2023b) and text content (Chen et al., 2022) improves the effectiveness and controllability of vision-language learning. In medical VLP, MMCLIP (Wu et al., 2026) further combines attention-guided masked image modeling with medical language modeling, highlighting the value of incorporating cross-modal interactions into masked pretraining. This masked-guided cross-modal interaction paradigm in vision-language learning inspires us to make the model focus on continuous information across images guided by cardiac function assessment reports. Therefore, we propose a structured visual masking strategy guided by cardiac function assessment reports, which encourages the model to focus on learning continuous inter-image information according to these reports.

Based on the above analysis, we introduce Medical-4D, a framework for cardiac function understanding and QA on short-axis cine CMR that combines factorized 4D representation learning with medically motivated continuous structured masked modeling and generative adaptation.

To study this framework in a clinical setting, we collected paired short-axis cine 4D cardiac imaging data and corresponding clinical cardiac function assessment reports across three data centers. These reports contain imaging findings that provide supervision for training a visual-language aligned model. In addition, based on these clinical cardiac function assessment reports, we used Qwen to construct a CMR question-answering dataset with over 30,000 pairs, including both open-ended and close-ended QA. We extensively evaluated Medical-4D on the three-center dataset, and the results show that Medical-4D substantially outperforms other medical VQA methods on the 4D CMR task.

Our contributions are threefold:**1)Factorized 4D cardiac representation.** We propose a factorized 4D representation that decomposes spatial, temporal, and inter-slice modeling, enabling the first 4D CMR-specific vision-language model to exploit the full 4D CMR data. Through normalized position encodings aligned with anatomical structure and temporal order, it encodes slice structural and temporal information with local–global positional relationships, thereby aligning with clinical cardiac function assessment reports.**2)The first cardiac function QA framework.** Building on the proposed factorized 4D representation, we present Medical-4D, the first cardiac function QA pipeline on full 4D CMR data.**3)Extensive experiments on 4D CMR QA.** We conduct extensive QA experiments on three data centers. Our method surpasses strong competitors and demonstrates cross-center generalization in the assessment of cardiac systolic and diastolic function.

2 Related Work

2.1 Spatiotemporal Modeling for Cardiac Cine MRI

Four-dimensional cardiac cine MRI (4D cine MRI) captures cardiac anatomy and motion throughout the cardiac cycle and typically contains multiple short-axis slices acquired over the full cycle Campello et al. (2021). Recent vision–language models have shown promise in cardiac MRI diagnosis, but their input construction often relies on frame subsampling or slice selection, reducing token counts and data dimensionality. Shad et al. (2026) use a video transformer with temporal subsampling and slice-wise feature flattening, which may miss brief motion changes and discard slice continuity. CMR-RD Li et al. (2026) targets 2D LGE-CMR image–text diagnosis rather than 4D or SAX cine data. CMR-CLIP (Nakashima et al., 2026) uses only the middle slice, and CineMA (Fu et al., 2026) uses only end-systole and end-diastole. Similarly, Wang et al. (2024) select representative slices and subsample frames before applying a Video Swin Transformer. These strategies inevitably discard temporal or spatial information, potentially degrading cardiac function assessment and fine-grained diagnosis.

2.2 Medical Vision-Language Models

Recent vision–language models have advanced rapidly in medical understanding, report generation, and medical visual question answering. General models adapted via medical instruction tuning and cross-modal alignment, such as Qwen Qwen Team (2026) and LLaVA Liu et al. (2023), acquire basic capabilities but often remain limited to elementary questions and lack accurate understanding of professional terminology and consistency with imaging evidence. Several medical VLMs target clinical imaging–language alignment. PanDerm Yan et al. (2025) extends VLMs to dermoscopic images, and CXR-CLIP You et al. (2023) applies vision–language contrastive learning to chest X-ray pretraining. EhrXQA Bae et al. (2023) advances question answering over chest X-ray records, while LLaVA-Med Li et al. (2023a) and LLaVA-Rad Zambrano Chaves et al. (2025) use curated instruction tuning on radiology dialogues and QA to improve image-grounded clinical reasoning and captioning. However, these models mainly use 2D or 3D medical images.

3 Method

Medical-4D consists of three stages, as illustrated in Figure 1. First, CMR 4D representation learning models spatial, temporal, and inter-slice dependencies and aligns the resulting study representation with clinical reports. Second, bidirectional cross-modal recovery introduces clinically structured masking to strengthen correspondence between phase–slice features and clinical descriptions. Finally, generative adaptation uses the learned representations for report generation and medical question answering. Additional architectural and objective details are provided in Appendix C.

3.1 CMR 4D Representation Learning

3.1.1 Cardiac-Phase and Anatomical-Slice Identity Encoding

Let $X \in \mathbb{R}^{B\times 1\times T\times Z\times H\times W}$ denote a batch of single-channel CMR cine volumes, where T and Z are the numbers of cardiac phases and anatomical slices, respectively. We rearrange each volume into independent phase–slice images and process them using a shared patch embedding layer and the pretrained MetaCLIP2 ViT-H spatial encoder. The CLS token from each image is extracted to obtain $X_{\text{spatial-CLS}} \in \mathbb{R}^{B\times T\times Z\times D}$, where $D = 1280$.

The spatial encoder captures within-frame anatomy, while the subsequent 4D encoder requires explicit information about the phase and slice associated with each frame. We therefore combine learnable temporal and slice-index embeddings, E_{time} and E_{slice}, with projections of normalized temporal and slice coordinates:

$$\begin{aligned} P_{\text{time-coord}} &= f_{\text{time}}(\tau), \\ P_{\text{z-coord}} &= f_{\text{z}}(\zeta). \end{aligned} \tag{1}$$

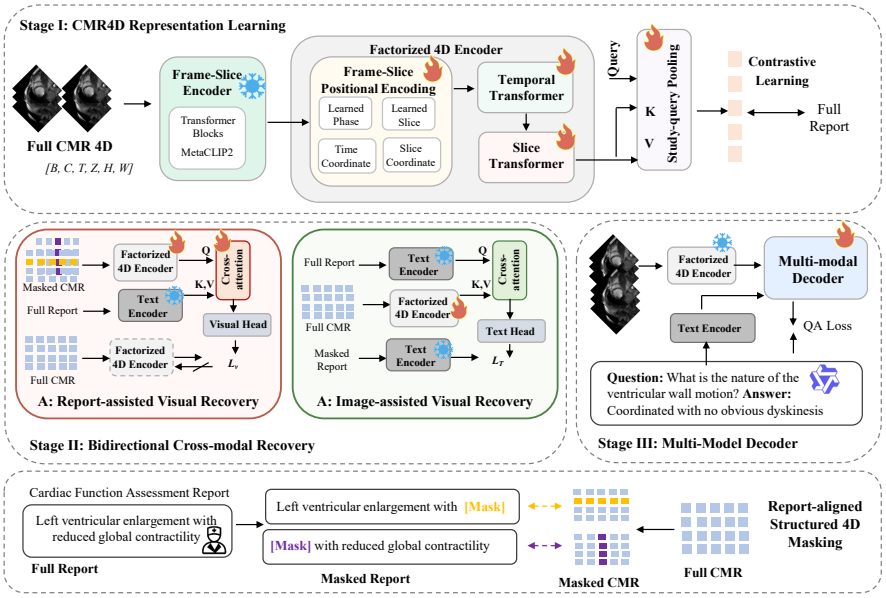

Figure 1: Overview of Medical-4D. Stage I extracts spatial features from individual phase–slice images, adds cardiac-phase and anatomical-slice identity encodings, and applies sequential temporal and slice attention. Stage II retains this alignment objective while alternating report-assisted visual recovery and image-assisted text recovery. Cross-attention connects report tokens with phase–slice features, using clinically structured visual masks and clinical phrase masking. Stage III freezes the learned encoders and trains an autoregressive decoder conditioned on the study embedding and question tokens for cardiac question answering.

Here, τ and ζ denote the temporal and slice coordinates, and f_{time} and f_{z} are learnable linear projections. Broadcasting over the complementary dimensions gives

$$X_0 = X_{\text{spatial-CLS}} + E_{\text{time}} + E_{\text{slice}} + P_{\text{time-coord}} + P_{\text{z-coord}}. \tag{2}$$

These encodings preserve the phase–slice organization of the visual tokens before spatiotemporal modeling.

3.1.2 Factorized 4D Spatiotemporal Visual Representation

To integrate cardiac motion and anatomical context efficiently, we factorize cross-frame modeling into temporal and slice attention. A temporal Transformer first models cardiac-phase dependencies within each anatomical slice. A slice Transformer then integrates anatomical context across slices within each cardiac phase:

$$\begin{aligned} X_{\text{time}} &= \mathcal{T}_{\text{time}}(X_0), \\ X_{\text{slice}} &= \mathcal{T}_{\text{slice}}(X_{\text{time}}), \\ X_{\text{seq}} &= \text{Flatten}_{T,Z}(X_{\text{slice}}). \end{aligned} \tag{3}$$

Both Transformers preserve the $B \times T \times Z \times D$ feature grid, and flattening produces $X_{\text{seq}} \in \mathbb{R}^{B\times(TZ)\times D}$. This sequence retains one contextual representation for each phase–slice pair.

A learnable study query q_{study} aggregates these representations through attention pooling:

$$X_{\text{study}} = \text{MHA}\left(q_{\text{study}}, X_{\text{seq}}, X_{\text{seq}}\right). \tag{4}$$

The pooled feature is normalized and projected to obtain the study-level visual embedding $v \in \mathbb{R}^{B\times 1024}$. Thus, X_{seq} supports phase–slice recovery, whereas v supports global report alignment and generative adaptation.

3.1.3 CONTRASTIVE LEARNING

The corresponding reports are encoded into study-level text embeddings $f_{\text{text}} \in \mathbb{R}^{B\times 1024}$. After ℓ_2 normalization, the image–text similarity logits are

$$S_{ij} = \lambda\, \bar{v}_i^{\mathsf{T}}\, \bar{f}_{\text{text},j}, \qquad \lambda > 0, \tag{5}$$

where λ is a learnable logit scale. Multiple augmented views may share the same patient identity. We therefore use a symmetric multi-positive contrastive objective $\mathcal{L}_{\text{CL}}$, treating image–report pairs from the same patient as positives.

To distinguish clinically similar but inconsistent descriptions, we also introduce report-level hard negatives. The hard-negative loss $\mathcal{L}_{\text{HN}}$ encourages each visual embedding to match its paired report more strongly than the corresponding perturbed reports. The Stage I objective is

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{CL}} + \beta \mathcal{L}_{\text{HN}}, \tag{6}$$

where β controls the hard-negative contribution. The complete contrastive objectives are given in Appendix C.3.

3.2 BIDIRECTIONAL CROSS-MODAL RECOVERY

Study-level contrastive alignment does not explicitly require correspondence between individual phase–slice features and clinical descriptions. We introduce two complementary recovery tasks to provide this finer-grained supervision. Stage II initializes the visual encoder from Stage I and keeps the text encoder frozen. Both recovery modules operate on phase–slice representations before study-query pooling, while the original alignment objective is retained.

3.2.1 REPORT-ASSISTED VISUAL RECOVERY

We define a visual mask $M_{\text{V}} \in \{0,1\}^{B\times T\times Z}$ over phase–slice planes. Structural findings guide contiguous slice masking within a cardiac phase, whereas functional findings guide contiguous phase masking within an anatomical slice. Sparse random plane masking is additionally applied. The masked input is

$$X^m = X \odot (1 - M_{\text{V}}), \tag{7}$$

where the mask is broadcast over the channel and spatial dimensions. Encoding X^m produces the contextual sequence X^m_{seq}.

The complete report is encoded into token representations $H_{\text{text}} \in \mathbb{R}^{B\times L\times D_{\text{text}}}$, where L is the report length and D_{text} is the text token width. A cross-attention recovery module uses masked visual features as queries and report tokens as keys and values:

$$\widehat{X}_{\text{seq}} = \mathcal{R}_{\text{V}}\left(Q = X^m_{\text{seq}},\, K = H_{\text{text}},\, V = H_{\text{text}}\right). \tag{8}$$

The shared visual encoder also processes the unmasked volume to produce X_{seq}. Its detached features serve as recovery targets. The visual loss $\mathcal{L}_{\text{V}}$ is the mean squared error over masked phase–slice positions and feature dimensions. This task encourages recovery of missing visual information using the remaining spatiotemporal context and the paired clinical report.

3.2.2 IMAGE-ASSISTED TEXT RECOVERY

In the reverse direction, we mask tokens belonging to a structural or functional clinical phrase. Let r^m denote the masked report and $M_{\text{T}} \in \{0,1\}^{B\times L}$ indicate its masked positions. The frozen text encoder produces H^m_{text}, while the complete CMR volume provides X_{seq}.

The text recovery module uses masked text features as queries and visual phase–slice features as keys and values:

$$\widehat{H}_{\text{text}} = \mathcal{R}_{\text{T}}\left(Q = H^m_{\text{text}},\, K = X_{\text{seq}},\, V = X_{\text{seq}}\right). \tag{9}$$

The recovered features are mapped to vocabulary logits using the frozen token-embedding matrix. The text loss $\mathcal{L}_{\text{T}}$ is the cross-entropy evaluated only at masked clinical-token positions.

We alternate the two recovery tasks while retaining alignment at every optimization step:

$$\mathcal{L}_{\text{Stage II}} = \begin{cases} \mathcal{L}_{\text{align}} + \gamma_{\text{V}}\mathcal{L}_{\text{V}}, & \text{visual recovery step,} \\ \mathcal{L}_{\text{align}} + \gamma_{\text{T}}\mathcal{L}_{\text{T}}, & \text{text recovery step.} \end{cases} \tag{10}$$

The weights γ_{V} and γ_{T} control the two auxiliary objectives. Their detailed formulations are provided in Appendix C.4.

3.3 Generative Adaptation

Stage III transfers the alignment encoders learned in Stage II and keeps them frozen. A trainable autoregressive decoder adapts their representations to report generation and medical question answering. The visual encoder produces $v = \mathcal{E}_{\text{visual}}(X)$, which is projected into a single visual memory token. For QA, the frozen text encoder processes the question q, including answer options when available, to produce contextual question tokens:

$$\begin{aligned} M_{\text{visual}} &= W_{\text{visual}} v, \\ M_{\text{question}} &= W_{\text{question}}\big(\mathcal{E}_{\text{text}}(q)\big), \\ M &= \text{Concat}\left(M_{\text{question}}, M_{\text{visual}}\right). \end{aligned} \tag{11}$$

Here, $M \in \mathbb{R}^{B\times(L_q+1)\times d}$, where L_q is the question length and $d = 1024$ is the decoder hidden dimension. For report generation, the question input is empty and the clinical report is the target.

Conditioned on M, the causal decoder predicts each target token from its preceding tokens:

$$p(y_\ell \mid y_{<\ell}, q, X) = \text{Decoder}(M, y_{<\ell})_{y_\ell}, \tag{12}$$

where the subscript denotes the predicted probability assigned to y_ℓ. Training uses teacher forcing and combines the QA and report-generation cross-entropy losses:

$$\mathcal{L}_{\text{gen}} = \lambda_{\text{QA}}\mathcal{L}_{\text{QA}} + \lambda_{\text{cap}}\mathcal{L}_{\text{cap}}. \tag{13}$$

Only the decoder and its associated input projections, positional parameters, and normalization layers are optimized in this stage.

4 Datasets

4.1 Real CMR Datasets

This study was conducted on a cardiac MRI dataset collected from three clinical centers. Each data pair consists of short-axis cine MRI acquired using a segmented balanced steady-state free precession (bSSFP) sequence during breath-holding, together with the corresponding physician assessment report on cardiac function. A total of 6264 subjects were enrolled, including 155 from data center 3, 2756 from data center 1, and 3353 from data center 2. The MRI systems and acquisition parameters differed across the three centers.

4.2 Constructing the Cardiac Function Question Answering Dataset

The construction of the cardiac function question answering dataset was based on the clinical cardiac function diagnostic reports provided by the three data centers described above. Each subject has a corresponding physician's cardiac function assessment report, as shown in Figure 2. A cardiac function assessment report generally includes assessments of cardiac structure and systolic/diastolic function, such as myocardial thickness, ventricular cavity size, and ventricular wall systolic or diastolic performance. To accurately evaluate cardiac structural and systolic/diastolic performance, we used Qwen 3.8-27B Qwen Team (2026) to extract information from each subject's clinical cardiac function assessment report, obtaining cardiac structure descriptions and myocardial wall systolic/diastolic descriptions, which we define as the structure part and the wall motion part, respectively. Subsequently, we used Qwen 3.8-27B Qwen Team (2026) to generate corresponding close-ended and open-ended questions for the structure part and the wall motion part, respectively. As shown in Figure 2, close-ended questions are multiple-choice questions, typically with 4 to 7 options, whereas open-ended questions have no fixed answers. All answer GTs are derived from the subject's clinical cardiac function assessment report. We constructed multiple questions for each subject. After a rough manual screening, the numbers of questions for structure in the open-ended and close-ended settings were 13,000 and 13,120, respectively, while those for wall motion were 10,214 and 11,239, respectively.

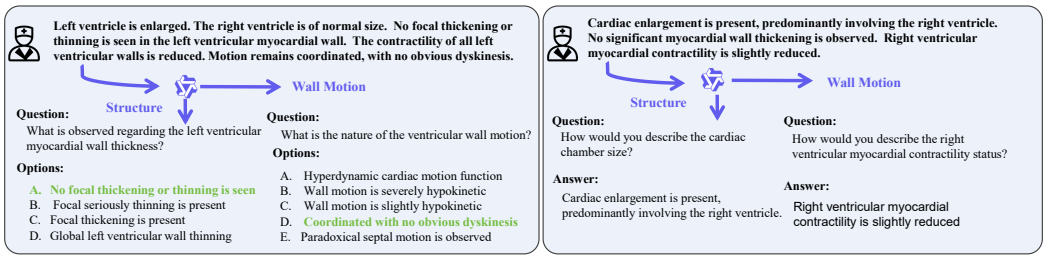

Figure 2: Workflow of constructing the cardiac function question answering dataset from clinical cardiac function reports.

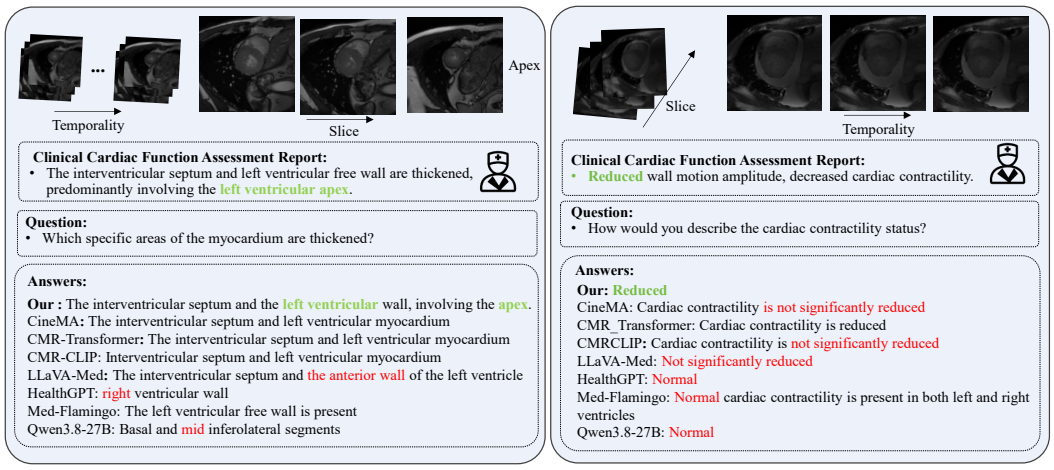

Figure 3: Example comparison of cardiac function assessment QA. Left: structure-related question. Right: wall motion-related question.

5 Experiment

5.1 Baselines

We used seven comparison methods: the general-purpose model Qwen3.8-27B Qwen Team (2026); the general medical models LLaVA-Med Moor et al. (2023); Li et al. (2023a), HealthGPT-Pro-8B Lin et al. (2025), and Med-Flamingo Moor et al. (2023); and the cardiac MRI-specific models CineMA Fu et al. (2026), CMR-CLIP Nakashima et al. (2026), and CMR-Transformer Shad et al. (2026). Except for Qwen3.8-27B Qwen Team (2026), all methods were fine-tuned or LoRA fine-tuned on our cardiac MRI dataset and cardiac function report–question dataset to ensure fair comparison. To match each method's design or pretrained model, we used different input forms: a single cardiac image, all slices at one cardiac cycle time point (3D), a single slice, or a cardiac video concatenating multiple slices (2D+T).

5.2 Qualitative Comparison

Figure 3 shows the visualization results for 4D cardiac function assessment. The left panel presents the assessment results for structure-related cardiac function questions, while the right panel presents the assessment results for wall motion. In the left panel, the Clinical Cardiac Function Assessment Report indicates that three regions—the interventricular septum, the left ventricular wall, and the apex—are thickened. These three thickened regions appear across multiple individual cardiac images along the slice dimension. Therefore, relying only on a single cardiac image or a cardiac video from one slice will inevitably miss some regions. For example, CMR-Transformer and CMR-CLIP, which can only adapt to cardiac video data from a single slice, both miss the diagnosis at the apex. In contrast, single-image-based methods such as HealthGPT and Med-Flamingo produce completely incorrect diagnoses. The right panel presents the wall motion assessment results. According to the Clinical Cardiac Function Assessment Report, the cardiac contractility status is Reduced, and only our method provides the correct answer. The assessment of systolic and diastolic cardiac function

depends on the entire cardiac cycle, and the differences between functional states are subtle and difficult to distinguish. Although methods such as CMR-Transformer and CMR-CLIP learn from cardiac data along the slice dimension, CMR-Transformer downsamples the cardiac cycle (a standard heartbeat cycle contains 25 or 30 frames) to 16 frames in order to adapt to certain released general video models, which inevitably affects cardiac wall motion diagnosis. Methods such as LLaVA-Med, HealthGPT, and CineMA completely discard the temporal information of the heart, which inevitably leads to completely incorrect assessments of wall motion-related cardiac function.

5.3 Quantitative Comparison

Table 1: Comparative analysis on the center 1 and center 2 datasets. Token recall for open-ended questions; accuracy for close-ended questions.

Method	Input	Close		Open	
		Structure	Wall Motion	Structure	Wall Motion
Med-flamingoMoor et al. (2023)	2D	72.84	35.71	58.92	55.63
Llava-med Li et al. (2023a)	2D	79.46	27.84	59.87	52.91
HealthGPT-Pro-8B	3D	77.56	36.48	40.75	49.82
CineMA Fu et al. (2026)	3D	**91.98**	61.92	65.88	55.74
Qwen 3.8-27B Qwen Team (2026)	2D+T	54.28	26.31	49.41	42.26
CMR-CLIP Nakashima et al. (2026)	2D+T	87.93	84.84	62.97	64.88
CMR-Transformer Shad et al. (2026)	2D+T	89.84	85.64	64.72	65.06
Ours (Medical-4D)	**4D**	91.72	**90.39**	**66.41**	**65.85**

Table 2: Evaluation on Center 3 (unseen during training). Token recall for open-ended questions; accuracy for close-ended questions.

Method	Input	Close		Open	
		Structure	Wall Motion	Structure	Wall Motion
Med-flamingoMoor et al. (2023)	2D	73.57	26.93	60.78	57.87
Llava-med Li et al. (2023a)	2D	80.20	28.99	61.48	54.55
HealthGPT-Pro-8B	3D	78.34	47.19	42.26	41.77
CineMA Fu et al. (2026)	3D	90.72	63.64	65.39	57.59
Qwen 3.8-27B Qwen Team (2026)	2D+T	55.07	36.82	45.78	43.81
CMR-CLIP Nakashima et al. (2026)	2D+T	88.66	89.07	64.54	46.25
CMR-Transformer Shad et al. (2026)	2D+T	**91.97**	88.43	65.44	63.61
Ours(Medical-4D)	**4D**	91.35	**90.13**	**65.81**	**64.65**

This section presents a quantitative QA comparison experiment based on cardiac function diagnostic reports from three data centers. To ensure fairness, except for Qwen 3.8-27B Qwen Team (2026), all methods are trained or LoRA fine-tuned on the center 1 and center 2 datasets. Specifically, 80% of the center 1 and center 2 datasets are randomly selected as the training set, and the remaining data are used as the test set. Table 1 presents the quantitative comparison results on open-ended and close-ended questions. To adapt to the pretrained models of CMR-Transformer and CMR-CLIP, we randomly select a cardiac video from a single slice, i.e., 2D+T. Table 1 shows that our method significantly outperforms the comparison methods in wall motion assessment. Although CMR-Transformer preserves cardiac temporal information, it downsamples heartbeat frames to adapt to the video model, resulting in poor wall motion performance. In contrast, the vision-language models based on 3D and 2D single images, which are widely used in medical models, discard temporal information and thus cannot be effectively applied to cardiac function assessment.

To assess the generalization of our method, we directly evaluated it on the Center 3 dataset without training. The results are reported in Table 2, which show that our method maintains stable performance on the cross-center dataset.

5.4 Ablation Study

We conduct two groups of ablation experiments to evaluate the contribution of the factorized 4D encoder and the bidirectional cross-modal recovery objectives. All experiments use the same MetaCLIP2 initialization, data split, report tokenizer, hard-negative reports, and optimization settings.

5.4.1 Factorized 4D Spatiotemporal Encoding

We evaluate the contribution of factorized 4D spatiotemporal encoding through three cumulative settings. The baseline independently encodes each phase–slice image using the MetaCLIP2 spatial encoder and aggregates the resulting features through study-query pooling. We first add the slice encoder to model anatomical-slice dependencies, and then incorporate the temporal encoder to capture cardiac-phase dependencies, yielding the complete factorized 4D encoder. The results are reported in Table 3.

Table 3: Ablation of factorized 4D spatiotemporal encoding.

Method	Close		Open	
	Structure	Wall Motion	Structure	Wall Motion
MetaCLIP2 + study-query pooling	87.42	87.96	59.15	59.28
+Slice encoder	90.00	88.54	65.78	62.05
+Temporal encoder	91.72	90.39	66.41	65.85

5.4.2 Bidirectional Cross-modal Recovery and Clinically Structured Masking

We evaluate the contribution of clinically structured masking and bidirectional cross-modal recovery through four progressively enhanced settings. The Stage I alignment model is used as the baseline and is trained with the original alignment objective. We first add report-assisted visual recovery with a slice-only mask, which evaluates whether report information can improve the recovery of anatomical-slice features. We then extend the visual mask to both the slice and temporal dimensions to assess the contribution of cardiac-phase continuity. Finally, we add image-assisted text recovery to the slice-and-temporal visual recovery setting, forming the complete bidirectional recovery model. All variants are initialized from the same Stage I checkpoint and use the same training schedule and downstream decoder protocol. Table 4 reports the results on close- and open-setting structure and wall-motion questions.

Table 4: Ablation of bidirectional cross-modal recovery and clinically structured masking.

Method	Close		Open	
	Structure	Wall Motion	Structure	Wall Motion
Stage I alignment	91.60	89.62	66.00	65.43
+ Report-assisted Visual Recovery (slice-mask)	91.69	89.83	66.21	65.48
+ Report-assisted Visual Recovery (slice + temporal mask)	91.70	90.34	66.40	65.83
+ Bidirectional Recovery	91.72	90.39	66.41	65.85

6 Conclusion

We propose Medical-4D, the first vision–language framework for cardiac function understanding and question answering from full 4D CMR cine imaging. Its factorized encoder jointly integrates spatial, temporal, and inter-slice information. Clinically structured masking and bidirectional cross-modal recovery further enhance the framework's performance on cardiac structure and wall motion assessment. Experimental results show that, compared with general vision-language models, medical vision-language QA models, and CMRI-specific models, Medical-4D achieves competitive performance on both open-ended and close-ended questions, especially in wall motion cardiac function assessment.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for generate synthetic data sets, implement methods, assist with translation. We have not used generative AI tools for propose or refine hypotheses, assist in the writing of proofs, design or provide feedback on research methodology or experiments, clean and reformat dataset, support qualitative and thematic data analysis, interpret results, create or modify scientific figures or images, and help develop theoretical models or conceptual frameworks are not applicable to this work. Additionally, we used generative AI tools for suggest experimental parameters, create or edit software code, summarize or analyse existing literature. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Below, we share a list of common research subtasks along with the necessity of disclosing them in the paper manuscript.

In the paper writing section, we utilized DeepSeek and ChatGPT to translate and refine papers. During the research process, we employed DeepSeek and ChatGPT 5.5 to assist in conducting literature reviews and organizing experimental ideas. In the experimental section, we used the locally deployed Qwen 3.8-27B to generate corresponding QA based on manually curated diagnostic content. In the comparative experiment, we employed Codex-ChatGPT 5.5 to assist in implementing certain experimental code.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

If authors feel that their paper submission raises questions regarding the Code of Ethics, they are encouraged to include a paragraph of Ethics Statement (at the end of the main text before references) to address potential concerns where appropriate. Topics include, but are not limited to, studies that involve human subjects, practices to data set releases, potentially harmful insights, methodologies and applications, potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues (e.g., IRB, documentation, research ethics). This statement should not be more than 1 page.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

It is important that the work published in ICLR is reproducible. Authors are strongly encouraged to include a paragraph-long Reproducibility Statement at the end of the main text (before references) to discuss the efforts that have been made to ensure reproducibility. This paragraph should not itself describe details needed for reproducing the results, but rather reference the parts of the main paper, appendix, and supplemental materials that will help with reproducibility. For example, for novel models or algorithms, a link to an anonymous downloadable source code can be submitted as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims can be included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps can be provided in the supplementary materials. Each of the above are examples of things that can be referenced in the reproducibility statement.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36:3867–3880, 2023.

Victor M. Campello et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021. doi: 10.1109/TMI.2021.3060083.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 679–689. Springer, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Yunguan Fu, Wenjia Bai, Weixi Yi, Charlotte Manisty, Anish N Bhuva, Thomas A Treibel, James C Moon, Matthew J Clarkson, Rhodri Huw Davies, and Yipeng Hu. Development and validation of a versatile foundation model for cine cardiac magnetic resonance image analysis. *Communications Medicine*, 2026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564, 2023a.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23390–23400. IEEE, 2023b.

Yansong Li, Zhongxi Qiu, Yun Tian, Zheng Jinyu, and Shuo Li. Cmr-rd: Long-tailed adaptive vlm for explainable cmr diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7003–7013, 2026.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh (eds.), *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 10 Dec 2023. URL `https://proceedings.mlr.press/v225/moor23a.html`.

Makiya Nakashima, Jielin Qiu, Peide Huang, Jihye Lee, Po-hao Chen, Richard Grimm, Christopher Nguyen, Byung-Hak Kim, Ding Zhao, Deborah Kwon, et al. Contrastive language image pre-training for a cardiac magnetic resonance image embedding with zero-shot capabilities. *Nature Communications*, 17(1):4416, 2026.

Mengyun Qiao, Kathryn A McGurk, Shuo Wang, Paul M Matthews, Declan P O'Regan, and Wenjia Bai. A personalized time-resolved 3d mesh generative model for unveiling normal heart dynamics. *Nature Machine Intelligence*, 7(5):800–811, 2025.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Rohan Shad, Cyril Zakka, Dhamanpreet Kaur, Mrudang Mathur, Robyn Fong, Joseph Cho, Ross Warren Filice, John Mongan, Kimberly Kallianos, Nishith Khandwala, et al. A generalizable deep learning system for cardiac mri. *Nature Biomedical Engineering*, pp. 1–16, 2026.

Yan-Ran Wang, Kai Yang, Yi Wen, Pengcheng Wang, Yuepeng Hu, Yongfan Lai, Yufeng Wang, Kankan Zhao, Siyi Tang, Angela Zhang, et al. Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. *Nature Medicine*, 30 (5):1471–1480, 2024.

Biao Wu, Yutong Xie, Zeyu Zhang, Vu Minh Hieu Phan, Qi Chen, Ling Chen, and Qi Wu. Mmclip: Cross-modal attention masked modelling for medical language-image pre-training. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2440–2455, 2026.

Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 31(8):2691–2702, 2025.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 101–111. Springer, 2023.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025.

A Appendix

A.1 Data Acquisition Parameters

- For Center 1, the images were acquired using two MRI systems. The first system was a MAGNETOM Vida scanner (Siemens Healthineers), with the following imaging parameters: acquisition matrix = 256 $\times$ 208–256, slice thickness = 8 mm, and TR/TE = 33.24/1.22 ms. The second system was an Ingenia Elition X scanner (Philips), with the following imaging parameters: acquisition matrix = 240 $\times$ 224–240, slice thickness = 8 mm, and TR/TE = 2.7818/1.391 ms.
- For Center 2, the images were acquired on a Skyra scanner (Siemens). The imaging parameters were as follows: acquisition matrix = 208 $\times$ 196–210, slice thickness = 8 mm, and TR/TE = 43.96/1.38 ms.
- For Center 3, the images were acquired on an Ingenia Elition X scanner (Philips). The imaging parameters were as follows: acquisition matrix = 288 $\times$ 288, slice thickness = 8 mm, and TR/TE = 3.1096/1.555 ms. The reported TR value corresponds to the temporal resolution for Siemens cine acquisitions.

B Implementation details for Medical-4D

B.1 Implementation Details

We initialize the spatial and text encoders from pretrained MetaCLIP2 ViT-H and use CMR inputs with $T = 30$ cardiac phases, $Z = 7$ anatomical slices, and a spatial resolution of 224×224.

The temporal and slice Transformers each contain two layers. All stages use AdamW and bfloat16 mixed-precision training. Stage I is trained for 50 epochs on eight NVIDIA A100 GPUs, with a learning rate of 10^{-5}, weight decay of 0.1, and cosine learning-rate decay following 100 warm-up steps. We use a per-GPU batch size of one and GradCache over eight mini-batches, yielding an effective contrastive batch size of 64. The pretrained text encoder and spatial Transformer blocks remain frozen. Each study uses two hard-negative reports with $\beta = 0.5$. Stage II continues training from Stage I for 10 epochs on eight A100 GPUs, retaining the same learning rate, optimizer settings, and effective contrastive batch size. Report-assisted visual recovery and image-assisted text recovery are alternated, with loss weights $\gamma_{\mathrm{V}} = 1.0$ and $\gamma_{\mathrm{T}} = 0.5$, respectively. Structured masking uses contiguous blocks of 3–5 slices or 10–15 cardiac phases, supplemented by a random plane-masking ratio of 0.02. Each recovery module uses one cross-attention layer with a hidden width of 256 and eight attention heads. Stage III is trained for 15 epochs on four A100 GPUs, using a per-GPU batch size of 24, a learning rate of 1.2×10^{-5}, weight decay of 0.01, and 100 warm-up steps. The alignment encoders remain frozen, while the generative decoder and its associated projections are optimized. The decoder contains four layers with 16 attention heads and a hidden dimension of $d = 1024$.

C Additional Method Details

This appendix supplements Section 3 with feature dimensions, complete training objectives, and implementation details. All symbols follow the main paper.

C.1 Spatial Features and Identity Encodings

The input tensor $X \in \mathbb{R}^{B\times 1\times T\times Z\times H\times W}$ is rearranged into BTZ single-channel images. In our configuration, $T = 30$, $Z = 7$, and $H = W = 224$. Each image is tokenized into a 16×16 spatial grid, yielding 256 patch tokens:

$$F_{\text{tokens}} \in \mathbb{R}^{BTZ\times 256\times D}.$$

These tokens are processed by the MetaCLIP2 ViT-H spatial encoder:

$$H_{\text{spatial}} = \mathcal{E}_{\text{spatial}}^{\text{MetaCLIP2}}(F_{\text{tokens}}).$$

Extracting and reshaping the frame-level CLS tokens produces $X_{\text{spatial-CLS}} \in \mathbb{R}^{B\times T\times Z\times D}$, with $D = 1280$.

The spatial positional embeddings of MetaCLIP2 encode patch locations within each image. The additional identity encodings in Equation equation 2 instead identify the phase and slice associated with each frame-level CLS token. Their dimensions are

$$\begin{aligned} E_{\text{time}} &\in \mathbb{R}^{T\times D}, & E_{\text{slice}} &\in \mathbb{R}^{Z\times D}, \\ \tau &\in \mathbb{R}^{B\times T}, & \zeta &\in \mathbb{R}^{B\times Z}, \\ P_{\text{time-coord}} &\in \mathbb{R}^{B\times T\times D}, & P_{\text{z-coord}} &\in \mathbb{R}^{B\times Z\times D}. \end{aligned}$$

The scalar normalized coordinates are independently mapped to D-dimensional vectors by linear projections. Temporal terms are broadcast across slices, and slice terms are broadcast across phases.

C.2 Factorized Attention and Study Pooling

Both $\mathcal{T}_{\text{time}}$ and $\mathcal{T}_{\text{slice}}$ contain two Transformer layers. For each slice z, temporal attention processes

$$X_0[:, :, z, :] \in \mathbb{R}^{B\times T\times D}.$$

For each phase t, slice attention subsequently processes

$$X_{\text{time}}[:, t, :, :] \in \mathbb{R}^{B\times Z\times D}.$$

Thus, the order of computation is temporal attention followed by slice attention. The cross-plane attention cost scales as $\mathcal{O}(ZT^2+TZ^2)$, compared with $\mathcal{O}((TZ)^2)$ for joint attention over the phase–slice grid, omitting shared feature-width and batch factors. This comparison concerns cross-plane attention rather than the full spatial encoder.

Flattening preserves a one-to-one correspondence between sequence positions and phase–slice pairs. The learnable query $q_{\text{study}} \in \mathbb{R}^{1\times D}$ is broadcast across the batch for Equation equation 4. We denote the squeezed pooling output by f_{pool}, and its study-normalized representation by f_{study}:

$$f_{\text{pool}}, f_{\text{study}} \in \mathbb{R}^{B\times D}.$$

Subsequent visual normalization and projection produce $v \in \mathbb{R}^{B\times 1024}$.

Throughout both the main paper and this appendix, $D = 1280$ denotes the visual token width. Contrastive similarities use the projected embedding v and the report embedding f_{text}, both of width 1024. The contextual text-token width D_{text} is distinct from this shared contrastive embedding dimension.

C.3 Multi-positive Alignment and Hard Negatives

The normalized embeddings in Equation equation 5 are

$$\bar{v}_i = \frac{v_i}{\|v_i\|_2}, \qquad \bar{f}_{\text{text},j} = \frac{f_{\text{text},j}}{\|f_{\text{text},j}\|_2}.$$

Training augmentations include slice-order reversal and periodic chunking along the phase dimension. Augmented views from the same patient retain the corresponding physician report. Let c_i and c_j denote the patient identities associated with the visual and textual samples. The positive-pair mask is

$$M_{ij} = \mathbb{I}(c_i = c_j).$$

For the effective contrastive batch of size B, the directional losses are

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\sum_{j=1}^{B} M_{ij}\exp(S_{ij})}{\sum_{j=1}^{B}\exp(S_{ij})}, \tag{14}$$

$$\mathcal{L}_{\text{t2i}} = -\frac{1}{B}\sum_{j=1}^{B} \log \frac{\sum_{i=1}^{B} M_{ij}\exp(S_{ij})}{\sum_{i=1}^{B}\exp(S_{ij})}. \tag{15}$$

Their average defines

$$\mathcal{L}_{\text{CL}} = \frac{1}{2}\left(\mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}\right).$$

For each study i, hard-negative reports contain clinically inconsistent modifications to the matched report. Their projected text embeddings are denoted by $f_{\text{neg},i,m} \in \mathbb{R}^{1024}$. The matched and hard-negative logits are

$$s_i^+ = \lambda\, \bar{v}_i^{\mathsf{T}} \bar{f}_{\text{text},i}, \qquad s_{i,m}^- = \lambda\, \bar{v}_i^{\mathsf{T}} \bar{f}_{\text{neg},i,m}.$$

With M candidate hard negatives per study, the loss is

$$\mathcal{L}_{\text{HN}} = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \log \frac{\exp(s_i^+)}{\exp(s_i^+) + \sum_{m=1}^{M} R_{i,m}\exp(s_{i,m}^-)}. \tag{16}$$

Here, $R_{i,m}$ indicates a valid hard negative, and $\mathcal{I}$ contains studies with at least one valid negative. In this expression, the scalar M counts candidate negatives; the decoder memory M in Equation equation 11 is a separate, stage-specific tensor. We use two hard-negative reports per study and set $\beta = 0.5$.

C.4 Structured Masking and Recovery Objectives

Visual masking and targets. The visual mask acts on complete input planes rather than individual spatial patches. A structural mask covers a contiguous slice segment within a phase, and a functional mask covers a contiguous phase segment within a slice. The configured block lengths are 3–5 slices and 10–15 phases, respectively, with an additional random plane-masking ratio of 0.02. Mask selection is restricted to valid phase–slice positions.

The unmasked and masked volumes share the same visual encoder. For recovery supervision, the unmasked features are detached, so the target branch does not receive gradients from $\mathcal{L}_{\text{V}}$. This

stop-gradient operation does not prevent the shared encoder from being trained by the alignment objective or the masked-input branch.

Let

$$|M_{\mathrm{V}}| = \sum_{b,t,z} M_{\mathrm{V}}^{b,t,z}.$$

For a nonempty visual mask, the mean squared recovery loss is

$$\mathcal{L}_{\mathrm{V}} = \frac{1}{D|M_{\mathrm{V}}|} \sum_{b,t,z} M_{\mathrm{V}}^{b,t,z} \left\| \widehat{X}_{\mathrm{seq}}^{b,t,z} - \mathrm{sg}(X_{\mathrm{seq}}^{b,t,z}) \right\|_2^2. \tag{17}$$

Although X_{seq} is stored as a flattened sequence, the superscripts b, t, z refer to its corresponding phase–slice position. The factor D explicitly averages the squared error over feature channels.

Clinical text recovery. A structural or functional phrase is selected from the report, and its corresponding token positions define M_{T}. The masked report supplies the queries in Equation equation 9; the unmasked visual sequence supplies the keys and values. The recovered text features are mapped to the token-embedding space and scored against the frozen vocabulary embedding matrix.

Let $r_{b,l}$ denote the original token and

$$|M_{\mathrm{T}}| = \sum_{b,l} M_{\mathrm{T}}^{b,l}.$$

The masked-token objective is

$$\mathcal{L}_{\mathrm{T}} = -\frac{1}{|M_{\mathrm{T}}|} \sum_{b,l} M_{\mathrm{T}}^{b,l} \log p\left(r_{b,l} \mid \widehat{H}_{\mathrm{text}}^{b,l}\right). \tag{18}$$

Unmasked tokens and padding positions do not contribute to this loss.

Cross-attention and optimization. Each recovery module contains one cross-attention layer with width 256 and eight attention heads. Learned projections map visual and textual features into the common attention space and map recovered features back to the required output space. The modules use X_{seq} or X_{seq}^{m}, rather than the pooled study embedding v.

Stage II alternates the two auxiliary objectives according to Equation equation 10, with $\gamma_{\mathrm{V}} = 1.0$ and $\gamma_{\mathrm{T}} = 0.5$. The report-alignment objective remains active in both cases. The text encoder and pretrained spatial Transformer blocks are frozen; the trainable visual components and recovery modules are optimized during this stage.

C.5 Generative Decoder and Token-level Supervision

Stage III transfers the Stage II alignment encoders and discards the recovery heads. The decoder receives one projected visual token and the contextual question tokens:

$$M_{\mathrm{visual}} \in \mathbb{R}^{B\times 1\times d}, \qquad M_{\mathrm{question}} \in \mathbb{R}^{B\times L_q\times d},$$

where $d = 1024$. It contains four causal Transformer layers with 16 attention heads. The maximum question and target sequence lengths are 77 and 128 tokens, respectively.

Training uses teacher forcing: the target sequence is shifted so that y_ℓ is predicted from $y_{<\ell}$ and the memory M. Let $\mathcal{M}_{\mathrm{QA}}$ and $\mathcal{M}_{\mathrm{cap}}$ indicate valid target-token positions for QA and report-generation samples, respectively. For $k \in \{\mathrm{QA}, \mathrm{cap}\}$, the task-specific loss is

$$\mathcal{L}_k = -\frac{\sum_{b,\ell} \mathcal{M}_k^{b,\ell} \log p(y_{b,\ell} \mid y_{b,<\ell}, q_b, X_b)}{\sum_{b,\ell} \mathcal{M}_k^{b,\ell}}. \tag{19}$$

Padding positions are excluded. For report-generation samples, q_b is empty. The two losses are combined using Equation equation 13, with $\lambda_{\mathrm{QA}} = \lambda_{\mathrm{cap}} = 1$. The alignment encoders and tied vocabulary embedding remain frozen, while the decoder and its associated trainable projections, positional parameters, and normalization layers are updated.