# *MedJEPA*: LATENT CROSS-MODAL PREDICTION FOR MEDICAL VISION-LANGUAGE REASONING UNDER MISSING IMAGES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Vision Language Models (VLMs) have become effective tools for reasoning over medical images, yet their multimodal reasoning pathways typically assume that the image is available at inference. In many real-world clinical scenarios, images may be unavailable because they are not transmitted across clinical workflows, are withheld due to privacy or data-sharing constraints, are corrupted during transmission, or have not yet been acquired. Existing missing-modality methods rely on prompt adaptation, retrieval, or modality reconstruction, but these strategies either work around the missing input or attempt to recreate the observation itself, rather than directly recovering the latent visual representation that the downstream VLM was trained to reason over. To overcome these limitations, we propose *MedJEPA*, a latent cross-modal prediction framework inspired by the Joint-Embedding Predictive Architecture (JEPA). Given only the clinical text and the corresponding question, *MedJEPA* predicts the patch-level visual representations that a frozen medical VLM would extract from the missing image, without synthesizing pixels.The predictor is trained with a JEPA-style masked latent objective in CLIP feature space, using the frozen image representation as a stop-gradient target. A masking curriculum spans partial to complete visual absence, with patch annealing progressively removing real visual tokens and iterative self-refinement improving the latent prediction from previous estimates. At inference time, the predicted visual tokens are projected through the frozen multimodal projector of the VLM and injected into its visual stream. A confidence-gated cascade then progressively evaluates sparse and dense visual-token injection, reverting to text-only reasoning when the predicted visual signal fails to improve the model's confidence. Across PMC-VQA and ScienceQA, *MedJEPA* consistently improves image-absent accuracy, from 70.8% to 73.4% on PMC-VQA and from 68.8% to 72.3% on ScienceQA, yielding gains of 2.6 and 3.5 percentage points, respectively. Ablations confirm the benefits of latent prediction, anti-collapse regularization, and self-refinement. These results show that latent cross-modal prediction enables medical VLMs to reason more reliably when images are unavailable.

## 1 INTRODUCTION

Vision-language models (VLMs) have emerged as powerful tools for medical reasoning by jointly interpreting clinical language and visual evidence from radiological and biomedical images (Li et al., 2023; Moor et al., 2023). However, their multimodal reasoning capability is fundamentally dependent on the availability of the corresponding image at inference time. In real-world clinical workflows, this assumption can fail due to incomplete transmission, privacy constraints, transmission corruption, or delayed image acquisition (Towbin et al., 2016). When the image is unavailable, the VLM must instead reason from the clinical text alone, potentially losing visual evidence that would otherwise inform its decision. This raises a fundamental question: ***Can a medical VLM recover useful visual information when the image itself is unavailable?***

Recent studies have shown that text-only models can do surprisingly well on medical visual question answering (VQA), emphasizing the importance of explicitly evaluating image-absent settings (Kahl et al., 2024). More broadly, the missing-modality literature has demonstrated that multimodal Trans-



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.

formers can be substantially affected by the absence of an input modality (Ma et al., 2022a). Existing approaches have explored several ways to mitigate this problem, including missing-modality-aware prompting (Lee et al., 2023a), retrieval-based strategies, and explicit modality reconstruction or imputation. While these approaches attempt to compensate for the missing modality, they generally do not directly target the latent visual representations consumed by the downstream VLM. More recent work has begun to explore the prediction of missing representations (Ke et al., 2025; Chen & Guo, 2026), suggesting that modality completion can be formulated directly in representation space. This motivates a more targeted question: ***Rather than reconstructing what the model cannot observe, can we recover the representation it actually needs for reasoning?***

To address this question, we introduce *MedJEPA*, a cross-modal joint-embedding predictive architecture that formulates the missing image representation as a latent prediction target. Given only clinical text and a question, *MedJEPA* predicts the patch-level visual tokens that a frozen CLIP encoder (Radford et al., 2021) would have extracted from the absent image. These tokens are then passed to a frozen VLM through its original multimodal projector and injected into its visual stream, thereby preserving the downstream reasoning pathway without requiring modification of the VLM. Unlike pixel-space reconstruction, which requires recovering appearance-level details that may be irrelevant to the clinical question, JEPA-style learning directly predicts representations in a pretrained feature space (Assran et al., 2023; 2025). *MedJEPA* extends this idea to the missing-modality inference by learning to predict an entirely absent visual representation from the available clinical text.

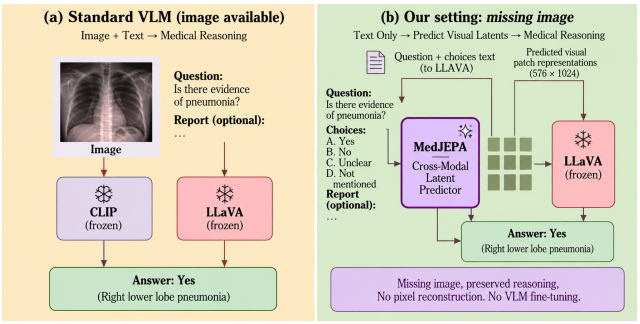
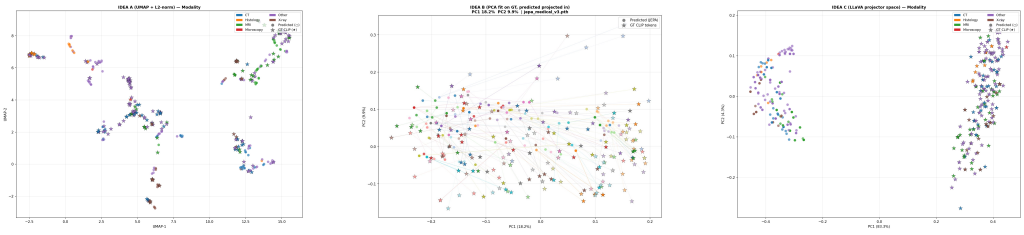



Figure 1: **MedJEPA overview.** (a) A standard VLM jointly reasons over image and text using frozen CLIP and LLaVA. (b) MedJEPA predicts visual representations from clinical text and injects them into the frozen VLM.

To make latent prediction robust to complete image absence at deployment, *MedJEPA* is trained under progressively increasing levels of visual absence using deployment-aligned masking, patch annealing, and iterative latent self-refinement. However, predicted visual representations may also introduce erroneous or spurious evidence when the latent prediction is uncertain. To address this issue, we employ a confidence-gated inference cascade that progressively evaluates sparse and dense visual-token injection and falls back to text-only reasoning when the injected representation fails to improve the model's confidence margin. This design separates visual representation recovery from downstream medical reasoning: *MedJEPA* is responsible for recovering the missing visual signal, while the frozen VLM remains responsible for multimodal reasoning and answer generation. For a detailed discussion of related work, see Appendix A.4.

Our contributions are summarized as follows:

1. **Latent formulation of missing-image medical VQA.** We formulate missing-image medical VQA as a latent visual representation prediction problem, where the objective is to predict the visual representation required by a frozen VLM rather than reconstructing the missing image in pixel space.
2. We introduce **MedJEPA**, a cross-modal JEPA framework that learns to predict missing visual representations using deployment-aligned masking, patch annealing, and iterative latent self-refinement.
3. **Confidence-gated latent injection.** We develop a confidence-gated inference cascade that selectively injects predicted visual tokens through sparse and dense configurations and falls back to text-only reasoning when the predicted visual signal does not improve the model's confidence margin.
4. **Comprehensive evaluation under missing vision.** We evaluate *MedJEPA* on PMC-VQA (Zhang et al., 2024) and ScienceQA (Lu et al., 2022) under image-absent settings and conduct ablation studies to quantify the contributions of latent prediction, anti-collapse regularization, and iterative self-refinement.



(a) PCA in LLaVA projector space. (b) PCA fitted on ground-truth tokens. (c) L2-normalized UMAP.

Figure 2: Representation analysis of MedJEPA predictions and ground-truth CLIP visual tokens. (a) PCA in the downstream LLaVA projector space reveals a global distributional gap between predicted and ground-truth representations. (b) PCA fitted only on ground-truth tokens shows substantial geometric overlap between the two representations. (c) L2-normalized UMAP reveals coherent modality-level structure across predicted and ground-truth tokens.

## 2 Our Framework: *MedJEPA*

*MedJEPA* addresses medical vision-language reasoning when the image is unavailable at inference. The key idea is to recover the *visual representation* required by a frozen VLM rather than reconstructing the missing image itself. We first introduce the latent interface used by the downstream VLM and the JEPA principle underlying our predictor, then formulate missing-image reasoning as cross-modal latent prediction. *MedJEPA* subsequently learns this mapping under deployment-aligned visual masking and finally injects the predicted representation through a confidence-gated cascade. The complete framework of training and inference of *MedJEPA* is shown in Figure 3.

### 2.1 Preliminaries

**VLM latent interface.** Modern vision-language models typically encode an input image into a sequence of visual representations and project these representations into the language model's embedding space, enabling them to be incorporated into subsequent multimodal reasoning (Radford et al., 2021; Liu et al., 2024). For an image $I$, a frozen visual encoder produces:

$$Z = E_{\mathrm{V}}(I) \in \mathbb{R}^{N \times d}, \tag{1}$$

where $N$ denotes the number of spatial visual tokens and $d$ their feature dimension. In our setting, the frozen CLIP-ViT-L/14336 encoder produces $N = 576$ tokens on a $24 \times 24$ grid with $d = 1024$. A frozen multimodal projector then maps these representations into the language-model space, preserving the interface through which the VLM performs multimodal reasoning (Liu et al., 2024). This interface is central to MedJEPA: the missing image can be unavailable, while its representation space remains the target required by the downstream model.

**Joint-embedding predictive learning.** JEPAs learn by predicting the representation of an unobserved target from available contextual information, avoiding explicit reconstruction of the raw observation (Assran et al., 2023). Given a context representation $C$ and a target representation $Z$, a predictor learns the following:

$$\widehat{Z} = G_\phi(C), \tag{2}$$

with the target treated as a fixed supervisory signal. This formulation allows the predictor to focus on structure represented in the target space rather than appearance-level details. MedJEPA extends this principle from within-modality prediction to *cross-modal* prediction: the context is text, while the target is the visual representation of an image that is entirely absent at inference. Additional background on Vision–Language Models, missing-modality learning, and Joint-Embedding Predictive Architectures is provided in Appendix A.5.

### 2.2 Motivation

A key question is whether the missing image should be reconstructed, or whether its useful visual information can be recovered directly in latent space. We study this through a representation

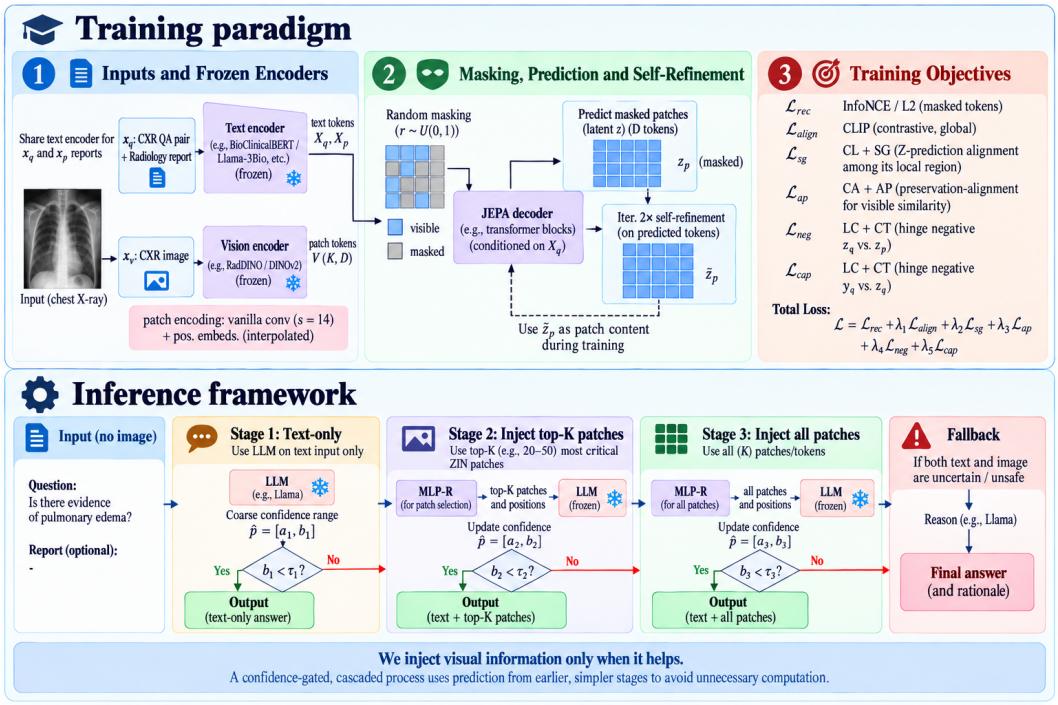



Figure 3: **Overview of MedJEPA.** During training, the model learns to predict masked CLIP visual representations from available clinical text and optional visible visual patches using deployment-aligned masking, patch annealing, and iterative self-refinement. At inference, the image is unavailable; MedJEPA predicts the complete visual representation from text alone and progressively injects sparse or dense predicted visual tokens into the frozen VLM using a confidence-gated cascade, with text-only fallback when the injected representation does not improve confidence.

analysis of MedJEPA on held-out PMC-VQA samples. Figure 2(a) shows a clear shift between predicted and ground-truth tokens in the LLaVA projector space. However, when PCA is fitted on the ground-truth representations, Figure 2(b) shows that the two representations share substantial geometric structure. Figure 2(c) further shows similar modality-level structure across the predicted and ground-truth tokens. These results suggest that the predicted representation does not need to match the exact missing image representation to capture useful visual information. Based on this observation, we formulate missing-image reasoning as a *latent prediction* problem rather than image reconstruction. MedJEPA therefore learns to predict a text-conditioned visual representation in the feature space used by the downstream VLM, with the goal of recovering visual structure that is useful for reasoning.

### 2.3 Problem Formulation

We consider medical visual question answering with paired image–text inputs, where each sample consists of an image $I$, textual context $t$, question $q$, and answer $y$, i.e., $\mathcal{D} = \{(I_i, t_i, q_i, y_i)\}_{i=1}^{N}$. In the standard multimodal setting, the model predicts the answer from both modalities, $p_\theta(y \mid I, t, q)$. Our target setting corresponds to complete image absence at inference, where only $(t, q)$ remain available and the model must predict $y$ without access to $I$. Let the frozen visual encoder $E_{\text{CLIP}}$ map the missing image to its visual representation $Z = E_{\text{CLIP}}(I) \in \mathbb{R}^{N_p \times d}$. The objective of MedJEPA is to infer this missing visual representation from the available textual context and use it to support the original multimodal reasoning pathway, without observing or reconstructing the image itself at inference.

### 2.4 CROSS-MODAL LATENT PREDICTION

Inspired by the context–target–predictor formulation of JEPA (Assran et al., 2023), we adapt latent predictive learning to the missing-image setting by separating the available and target modalities. Unlike conventional JEPA, where contextual and target representations originate from the same modality, ***MedJEPA*** uses the available clinical text as context and the CLIP representation of the unavailable image as the prediction target. The frozen PMC-LLaMA embeddings are therefore contextualized by a trainable TextContextEncoder, while any visible CLIP patches available during training are encoded with spatial positional information and used only as auxiliary visual context. A Transformer predictor then maps these contextual representations to the masked visual tokens:

$$\widehat{Z}_{\mathcal{M}} = G_\phi\left(C_{\text{text}}, C_{\text{patch}}, \mathcal{M}\right). \tag{3}$$

We instantiate the predictor with 576 learnable spatial queries, one for each position in the $24 \times 24$ CLIP patch grid, so that the predicted representation preserves the spatial structure expected by the downstream VLM. At complete visual absence, $C_{\text{patch}} = \varnothing$, and the same predictor reduces to pure text-conditioned prediction of the full $576 \times 1024$ visual representation. The resulting latent is therefore not an independently generated visual feature, but an estimate of the representation that the frozen VLM would have received from the missing image.

### 2.5 DEPLOYMENT-ALIGNED TRAINING

A key challenge in cross-modal latent prediction is the discrepancy between training and deployment: the predictor may observe visual context during training, whereas no visual tokens are available when the image is missing. We therefore train ***MedJEPA*** to progressively transition from partially observed visual inputs to complete visual absence. This training strategy combines masked latent prediction with a curriculum over visual context and iterative prediction refinement.

**Dynamic masking.** Following the masked-prediction principle underlying MAE and JEPA (He et al., 2022; Assran et al., 2023), we randomly mask visual tokens during training, sampling $\rho \sim \mathcal{U}(0.50, 1.00)$ at each step. This exposes the predictor to a continuum of missing-vision conditions, including the fully masked case $\rho = 1$, which exactly matches the deployment setting. Unlike standard within-modality masking, the masked tokens here represent the visual state that must ultimately be inferred from the available text.

**Patch annealing.** Although complete masking is required at deployment, providing a small amount of real visual context early in training offers spatial scaffolding for learning the correspondence between textual context and visual patch locations. We therefore initially retain up to 50% of the ground-truth patches and cosine-anneal their availability to zero over the first 80% of training. The predictor consequently ends training under the same fully image-absent condition encountered at inference.

**Latent self-refinement.** We further adapt the latent-prediction paradigm to allow the predictor to improve its own visual estimate. The first pass produces the complete visual representation from text alone,

$$\widehat{Z}^{(1)} = G_\phi\left(C_{\text{text}}, \varnothing, \mathcal{M}_{\text{all}}\right), \tag{4}$$

which is detached and re-encoded as visual context for a second prediction:

$$\widehat{Z}^{(2)}_{\mathcal{M}} = G_\phi\left(C_{\text{text}}, F_{\text{patch}}\left(\text{sg}[\widehat{Z}^{(1)}]\right), \mathcal{M}\right). \tag{5}$$

The refinement prediction is supervised against the same frozen visual target, encouraging the model to improve an initial text-conditioned estimate rather than repeatedly predicting from identical context. The same refinement mechanism is used at inference, providing consistency between training and deployment.

**Latent prediction objective.** We train the predictor directly in the frozen CLIP representation space using normalized, stop-gradient targets. The objective combines token-level prediction and refinement with global text–vision alignment and explicit variance regularization:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + 0.5\mathcal{L}_{\text{refine}} + 0.20\mathcal{L}_{\text{sig}} + 0.20\mathcal{L}_{\text{align}} + 0.30\mathcal{L}_{\text{var}}. \tag{6}$$

The SigLIP-style term provides global cross-modal supervision (Zhai et al., 2023), while the variance term adapts the anti-collapse principle of VICReg to preserve diversity across predicted spatial

Table 1: Performance comparison under full-image and image-absent settings. MedJEPA denotes our latent visual representation prediction framework. The Δ column reports the absolute improvement in percentage points over the corresponding text-only setting. The results in green are models where *MedJEPA* was used during inference , the red and blue rows shows various noise bounds.

| Model | Full-image acc. (%) | Text-only acc. (%) | + MedJEPA (%) | Δ (pp) |
|---|---|---|---|---|
| GLM-5.3-Flash | 85.4 | 90.2 | – | – |
| Qwen3-VL-235B | 89.22 | 90.80 | – | – |
| DeepSeek-V4.1-Flash | 87.72 | 90.3 | – | – |
| Qwen3-VL-8B | 88.20 | 90.08 | – | – |
| LLaMA-4-Maverick | 87.42 | 82.80 | – | – |
| **Phi-3.5-Vision-Instruct** | 86.92 | 88.8 | **90.4** | ↑ **1.6** |
| **LLaVA-1.5-13B** | 88.42 | 87.7 | **89.6** | ↑ **2.1** |
| **MedLLAVA-7B** *(primary)* | 69.69 | 70.8 | **73.4** | ↑ **2.8** |
| *Noise / bounds — MedLLAVA-7B backbone* | | | | |
| Random chance | – | 25.0 | – | – |
| Random visual tokens (B4) | – | 26.3 | – | – |
| GT CLIP-L tokens (B3, ceiling) | **74.6** | – | – | – |

tokens (Bardes et al., 2022). Together, these objectives enforce local latent fidelity while maintaining global correspondence and spatial structure.

### 2.6 Confidence-Gated Latent Injection

At inference, the image is unavailable and *MedJEPA* first obtains the refined latent prediction $\widehat{Z}$, which is mapped through the frozen multimodal projector and supplied to the downstream VLM. Since the text-conditioned representation is inherently uncertain, we do not automatically replace the missing image with all predicted tokens. Instead, visual evidence is introduced progressively according to the VLM's confidence. Let

$$m(v) = p(a^{(1)} \mid t, q, v) - p(a^{(2)} \mid t, q, v), \tag{7}$$

denote the probability margin between the two most likely answer choices under visual information $v$. **Stage 1: Text-only.** We first evaluate the frozen VLM with $v = \varnothing$. If $m(\varnothing) > \tau_1$, with $\tau_1 = 0.5$, the text-only prediction is retained and no latent visual tokens are injected. **Stage 2: Sparse latent injection.** For cases that remain uncertain, we rank the predicted patches by their refinement consistency and inject only the $K = 144$ most stable tokens, corresponding to 25% of the visual grid. The sparse prediction is retained when $m(\widehat{Z}_{\text{top}K}) > m(\varnothing)$. **Stage 3: Dense latent injection.** If sparse injection does not improve the text-only confidence, we provide the complete predicted representation $\widehat{Z}$ comprising all 576 visual tokens. The dense prediction is retained only when $m(\widehat{Z}) > m(\varnothing)$; otherwise, the system falls back to the Stage-1 text-only prediction. The cascade therefore treats the predicted latent representation as additional visual evidence rather than a mandatory replacement for the missing image, progressively increasing visual intervention only when it improves the downstream VLM's confidence. Additional architectural, optimization, objective, and inference details are provided in Appendix A.7.

## 3 Experiments

### 3.1 Settings

**Dataset.** We evaluate *MedJEPA* on two complementary multimodal reasoning benchmarks, PMC-VQA (Zhang et al., 2023) and ScienceQA (Lu et al., 2022). PMC-VQA spans diverse biomedical domains, including predominantly radiology alongside pathology, microscopy, signals, and biomedical illustrations, with medical images paired with questions, four answer choices, and figure captions. ScienceQA covers natural science, social science, and language science across elementary-to-high-school curricula, with each instance containing a question, four answer choices, a textual hint, and

an optional image. We use the figure caption and hint as the textual conditioning inputs, respectively, while the corresponding images are used as visual targets during training and withheld at inference.

**Implementation Details.** We implement *MedJEPA* with a four-layer Pre-LN Transformer TextContextEncoder using 16 attention heads, a $4\times$ FFN expansion, and a hidden dimension of 1024, preceded by a linear projection from the frozen PMC-LLaMA 7B token embeddings (4096→1024). The PatchContextEncoder augments visible CLIP-ViT-L/14336 (Radford et al., 2021) patch tokens with learnable spatial positional embeddings; no additional projection is required since the CLIP-L feature dimension matches the 1024-dimensional hidden space. The JEPAPredictor is a four-layer Pre-LN Transformer decoder with 8 attention heads, a $2\times$ FFN expansion, and 576 learnable query embeddings, producing 1024-dimensional visual tokens that are directly compatible with LLaVA-1.5's `multi_modal_projector`. The resulting architecture contains approximately 85M trainable parameters. We train all learnable components with AdamW using a learning rate of $1 \times 10^{-4}$, weight decay of 0.05, $\beta = (0.9, 0.95)$, cosine learning-rate decay with 400 warm-up steps, and gradient clipping at 1.0. Training uses a batch size of 1 with 8-step gradient accumulation, yielding an effective batch size of 8, while the CLIP-ViT-L/14336 encoder and PMC-LLaMA embedding layer remain frozen throughout (Wu et al., 2024). The visual mask ratio is sampled uniformly from $[0.50, 1.00]$ at each training step, while the proportion of visible ground-truth patches is cosine-annealed from 0.50 to 0 over the first 80% of training.

Table 2: Results on ScienceVQA under the out-of-distribution (OOD) setting. *MedJEPA* denotes our latent visual representation prediction framework. The Δ column reports the absolute improvement in percentage points over the corresponding text-only setting.

| Model | Full-image acc. (%) | Text-only acc. (%) | + MedJEPA (%) | Δ (pp) |
|---|---|---|---|---|
| GLM-5.3-Flash | 98.44 | 91.01 | – | – |
| Qwen3-VL-235B | 96.00 | 82.44 | – | – |
| DeepSeek-V4.1-Flash | 98.67 | 90.30 | – | – |
| Qwen3-VL-8B | 88.67 | 82.00 | – | – |
| LLaMA-4-Maverick | 74.88 | 75.58 | – | – |
| Phi-3.5-Vision-Instruct | 95.11 | 85.80 | 88.4 | ↑ **2.6** |
| LLaVA-1.5-13B | 79.56 | 76.70 | 79.6 | ↑ **2.9** |
| **MedLLaVA-7B (primary)** | **58.67** | **68.80** | **72.3** | ↑ **3.5** |

Table 3: Ablation of loss components.

| Setting | Acc. (%) | Δ |
|---|---|---|
| **v2: (L1 + Refinement)** | **70.46** | **0** |
| L2: (L1 + VICReg) | 69.30 | ↓ **1.16** |
| L3: (L1 + Alignment) | 70.40 | ↓ **0.06** |
| **v3: (all terms)** | **73.20** | ↑ **2.74** |

**Baselines.** In our evaluation, we benchmark MedJEPA across a diverse set of multimodal models, including GLM-5.3-Flash (Zeng et al., 2026), Qwen3 VL-235 (Bai et al., 2025), DeepSeek-V4.1-Flash (Xu et al., 2026), Qwen2-VL-7B-Instruct (Bai et al., 2025), LLaMA-3.2-11B-Vision-Instruct (Meta, 2025), Phi-4-Vision-Instruct (Abdin et al., 2024), LLaVA-1.5-13B (Liu et al., 2024), and Med-LLaVA-7B (Li et al., 2023). We additionally consider controlled baselines, including uniform random selection, question text-only reasoning , and caption+question text-only reasoning to isolate the contribution of available textual context. To assess the role of visual representations, we further evaluate random visual tokens and ground-truth CLIP-L/14@336 tokens, representing the unstructured-token floor and the upper bound of the token-injection pipeline, respectively.All ablations are conducted using the MedLLaVA-7B model. Details of the models used and the evaluation protocol are provided in Appendix A.9.

**Metrics.** We report MCQ accuracy, measured as the fraction of test questions for which the model's predicted answer letter (A/B/C/D) matches the ground-truth label. For PMC-VQA, ground-truth labels are recovered by matching the annotated answer text to the four answer-choice columns, while for ScienceQA they are obtained directly from the integer answer index.

### 3.2 Main Results

**Effectiveness of MedJEPA.** Table 1 shows that *MedJEPA* consistently improves image-absent reasoning across the evaluated VLMs. On the primary MedLLAVA-7B backbone, *MedJEPA* increases accuracy from 70.8% to 73.4%, a gain of 2.8 percentage points. Similar improvements are observed for Phi-3.5-Vision-Instruct (88.8%→90.4%) and LLaVA-1.5-13B (87.7%→89.6%), demonstrating that the latent prediction framework transferred across different VLM architectures. The random

visual-token baseline remains close to chance (26.3% vs. 25.0%), indicating that gains are not due to arbitrary token injection, while ground-truth CLIP tokens reach 74.6%, placing *MedJEPA* only 1.2 percentage points below the visual-token ceiling on MedLLAVA-7B.

**Effects of Different Losses.** Table 3 shows that the complete loss formulation provides the strongest performance. The v2 baseline achieves 70.46% accuracy, while using VICReg or alignment alone slightly reduces performance, by 1.16 and 0.06 percentage points, respectively. In contrast, combining refinement, SigLIP, alignment, and VICReg in the full objective improves accuracy to 73.20%, a gain of 2.74 points over the baseline. This result indicates that the auxiliary objectives are most effective when jointly optimized, providing complementary supervision for latent prediction.

**Architecture Analysis.** The results in Table 4 show that the choice of visual target layer has a noticeable effect on downstream performance. The penultimate CLIP layer provides 70.46% accuracy, while the earlier layer ($-3$) gives a comparable 70.72%; in contrast, the final layer drops substantially to 66.42%, despite its higher representation-level cosine similarity. This indicates that stronger target alignment does not necessarily yield more useful patch representations for downstream reasoning. Replacing the PMC-LLaMA text embeddings with the CLIP text encoder also slightly reduces accuracy (70.46%→70.28%), supporting the use of the medical language representation for textual conditioning.

Table 4: Architecture ablations on the first 5,000 PMC-VQA test examples. Both ablations use the v2 training recipe with the full training set and 20,000 training steps. $\Delta$ is relative to the v2 default configuration (R1 layer $-2$, 70.46%).

| Abl. | Setting | Acc. | Δ | Cos. |
|---|---|---|---|---|
| R1 | CLIP $-1$ | 66.42 | ↓ **4.04** | 0.6544 |
| | $-2$ **(default)** | **70.46** | **0** | **0.5439** |
| | $-3$ † | 70.72 | ↑ **0.26** | 0.5320 |
| | $-4$ | 69.76 | ↓ **0.70** | 0.5177 |
| T2 | **Ours** PMC-LLaMA 7B | **70.46** | **0** | – |
| | CLIP-L | 70.28 | ↓ **0.18** | – |

**Performance on Out of Domain Setting.** Table 2 shows that *MedJEPA* consistently recovers useful visual information under distribution shift. On the primary MedLLAVA-7B backbone, image-absent accuracy improves from 68.8% to 72.3% (+3.5 points), while Phi-3.5-Vision-Instruct and LLaVA-1.5-13B gain 2.6 and 2.9 points, respectively. The consistent improvements across all evaluated backbones, together with the larger gain on MedLLAVA-7B than in the PMC-VQA setting, indicate that latent visual prediction remains beneficial even when the reasoning task changes to ScienceVQA. Additional experimental results for different hyperparameter settings, including VICReg, SigLIP, and alignment weights, are provided in the Appendix A.10.

### 3.3 Ablations

We have done ablations on different hyperparameters namely inference refinement , mask ratio , refinement loss weight and varience loss weight as shown in Figure 4 a, , Figure 4 b , 4 c and 4 d respectively. More ablations are shown in the Appendix section A.10. **Impact of Inference Refinement.** Increasing the number of refinement passes improves performance up to two passes, after which the gains largely saturate. Accuracy rises from 69.58% without refinement to 69.82% at the default two passes, with only marginal improvement at three and four passes. This suggests that iterative self-refinement provides useful correction early, while additional passes offer limited benefit. **Impact of Mask Ratio Range.** Performance remains relatively stable across masking schedules, with the default $0.50$–$1.00$ range providing a strong trade-off between partial and complete visual absence. Increasing the minimum mask ratio to $0.75$ yields the highest accuracy (69.88%), whereas a fixed 75% mask performs worse, indicating that exposure to varying levels of missing vision is more beneficial than training at a single masking severity. **Impact of Refinement-Loss Weight.** The refinement-loss sweep shows that performance is best around the default $\lambda_{\text{refine}} = 0.50$, with larger weights consistently degrading accuracy. Increasing the weight to 0.75 or 1.00 reduces performance to 69.72%, suggesting that overly emphasizing the refinement objective can introduce noise rather than improving the latent representation.

**Impact of Variance-Loss Weight.** Performance remains relatively stable around the default variance-loss weight, but overly large weights lead to degradation. The default $\lambda_{\text{var}} = 0.15$ main-

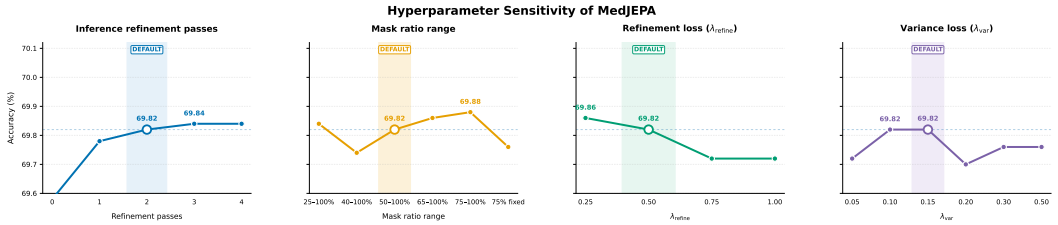



Figure 4: Hyperparameter sensitivity of MedJEPA across four key settings: inference refinement passes, mask ratio range, refinement-loss weight $\lambda_{\text{refine}}$, and variance-loss weight $\lambda_{\text{var}}$. The highlighted point in each panel denotes the default configuration, while the curves show performance across the corresponding hyperparameter sweep.

tains 69.82% accuracy, whereas increasing it to 0.20 lowers performance to 69.70%. This indicates that anti-collapse regularization is useful when applied moderately, but excessive variance constraints can interfere with downstream reasoning.

### 3.4 Image Retrieval Diagnostic

To further demonstrate the effectiveness of MedJEPA, we perform an auxiliary image-retrieval diagnostic. For each test caption, the JEPAPredictor generates 576 visual tokens using text alone, which are mean-pooled across spatial positions to obtain a 1024-dimensional query embedding. We construct a gallery of 1,000 test images by similarly mean-pooling their CLIP-ViT-L/14336 patch representations and rank the gallery using cosine similarity, returning the top-15 images for each query. We observe the retrievals exhibit a consistent modality-level correspondence across all queries: CT angiography captions retrieve axial CT images, MRI captions retrieve T1, T2, and STIR sequences, endoscopy captions retrieve endoscopic images, and histology captions retrieve tissue-stain images throughout the top-15 results. This behavior emerges without explicit modality supervision, showing that the predictor places its text-conditioned representations in modality-consistent regions of the pretrained CLIP space, consistent with the structure observed in our representation analysis. The diagnostic therefore clarifies the nature of the learned latent signal: MedJEPA recovers *what kind of visual evidence* the text is likely to describe rather than *which specific image* is missing. Such coarse visual structure can still provide useful context to the downstream VLM, while the absence of exact-image recovery highlights the uncertainty inherent in latent completion and motivates our confidence-gated cascade, which injects predicted visual tokens only when they provide additional evidence beyond text-only reasoning. The corresponding image associated with this shown in the Appendix section A.8 .

## 4 Conclusion

We presented *MedJEPA*, a latent cross-modal prediction framework for medical vision-language reasoning when the image is unavailable at inference time. Rather than reconstructing pixels or adapting the VLM to ignore the missing modality, MedJEPA predicts the patch-level visual representation that the frozen VLM would otherwise receive, using only the available clinical text and question. Deployment-aligned masking, latent self-refinement, and confidence-gated visual-token injection enable the predictor to recover useful visual structure while limiting the impact of uncertain latent estimates. Across PMC-VQA and ScienceQA, MedJEPA consistently improves image-absent reasoning over text-only inference and remains effective across different VLM backbones and under task transfer, while representation analyses show that the predicted tokens capture structured visual and modality-level information despite not recovering the exact missing image. These findings support a broader view of missing-modality reasoning: when the original observation is unavailable, recovering a downstream-useful latent representation can be more practical than reconstructing the observation itself. More broadly, MedJEPA demonstrates that a frozen medical VLM can be augmented with a learned latent completion module without modifying its underlying reasoning pathway, providing a general approach for recovering useful multimodal reasoning capability under incomplete visual input.

## REFERENCES


Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Variance-invariance-covariance regularization for self-supervised learning. *ICLR, Vicreg*, 1(2), 2022.

Taixi Chen and Nancy Guo. Learning from reliable latent prompts for visual recognition with missing modalities. *arXiv preprint arXiv:2606.30597*, 2026.

Renjie Gu, Kaichen Zhou, Yan Luo, and Mengyu Wang. Geoworld-vlm: Geometry from world models for vision-language models. *arXiv preprint arXiv:2605.16713*, 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.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017, 2023.

Kim-Celine Kahl, Selen Erkan, Jeremias Traub, Carsten T Lüth, Klaus Maier-Hein, Lena Maier-Hein, and Paul F Jaeger. Sure-vqa: Systematic understanding of robustness evaluation in medical vqa tasks. *arXiv preprint arXiv:2411.19688*, 2024.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25864–25873. IEEE, 2025.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14943–14952, June 2023a.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14943–14952. IEEE, 2023b.

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, 2023.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18177–18186, June 2022a.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18156–18165. IEEE, 2022b.

Jing Meng, Zhenfang Zhu, Qiang Lu, Hongli Pei, Kefeng Li, Yuzhi Ren, Meng Li, Xiaowen Sun, and Dawei Zhao. A prompt-driven framework for compensation and fusion in multimodal sentiment analysis with missing modalities. *Knowledge-Based Systems*, pp. 115438, 2026.

AI Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2025.

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 *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023.

Avishka Perera, Kumal Hewagamage, Saeedha Nazar, Kavishka Abeywardana, Hasitha Gallella, Ranga Rodrigo, and Mohamed Afham. Crossjepa: Cross-modal joint-embedding predictive architecture for efficient 3d representation learning from 2d images. *arXiv preprint arXiv:2511.18424*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. Vla-jepa: Enhancing vision-language-action model with latent world model. In *European Conference on Computer Vision*, pp. 478–497. Springer, 2026.

Alexander J Towbin, Christopher J Roth, Mark Bronkalla, and Dawn Cram. Workflow challenges of enterprise imaging: Himss-siim collaborative white paper. *Journal of digital imaging*, 29(5): 574–582, 2016.

Yunbo Wang and Yuxin Peng. Mars: Learning modality-agnostic representation for scalable cross-media retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4765–4777, 2021.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. Deepseek-v4. 1-flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):12113–12132, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering, 2024. *URL https://arxiv. org/abs/2305.10415*, 40, 2024.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

# A APPENDIX

The Appendix provides supplementary material, including reproducibility statement A.1, use of large language models A.2, ethics statement A.3 , related works A.4, background details A.5, limitations A.6, additional method details A.7, image retrieval analysis A.8, and Additional experimental results A.10 .

## A.1 REPRODUCIBILITY STATEMENT

We have taken careful steps to ensure the reproducibility of our work. At this stage, we have created an anonymous repository (`https://anonymous.4open.science/r/MedJEPA-9C45/README.md`) containing sample benchmark data and the code for our main method. Upon acceptance, we plan to release the complete implementation, benchmark details, and a dedicated project page to facilitate full reproducibility and further research.

## A.2 USE OF LARGE LANGUAGE MODELS

We employed LLMs as supplementary tools during manuscript preparation, primarily for language refinement and editorial assistance, as well as for literature search and discovery of relevant prior work. They did not contribute to the formulation of the research problem, methodological design, experimental execution, analysis, or scientific conclusions, all of which were carried out and validated by the authors. Additional details on their use are provided in the corresponding sections of the manuscript.

## A.3 ETHICS STATEMENT

MedJEPA is evaluated entirely on publicly available benchmark datasets, including PMC-VQA and ScienceQA, and does not involve the collection of new patient data, interaction with human subjects, or prospective clinical experimentation. The medical images and associated questions used in our experiments are treated solely as research data for retrospective model evaluation, and the original data sources and usage conditions are respected. MedJEPA is intended as a research framework for robust multimodal reasoning under missing inputs and should not be interpreted as a clinical diagnostic system without further validation, calibration, appropriate clinical oversight, and compliance with applicable privacy and regulatory requirements.

## A.4 RELATED WORK

### A.4.1 MISSING-MODALITY LEARNING

Multimodal models experience substantial performance degradation when one or more input modalities are unavailable, with the extent of this degradation depending strongly on the underlying modality-fusion mechanism (Ma et al., 2022b). Existing approaches have primarily addressed missing modalities through missing-modality-aware prompting and lightweight adaptation (Lee et al., 2023b), retrieval or generation of complementary information (Yao et al., 2022), and reconstruction or prediction of missing observations or latent representations (Meng et al., 2026; Wang & Peng, 2021). While these methods provide effective mechanisms for compensating for missing inputs, most have been developed primarily for classification or representation learning rather than generative reasoning with frozen VLMs. This distinction motivates *MedJEPA*, which directly predicts the latent visual representation required by an existing multimodal reasoning pathway, rather than modifying the model to ignore the missing modality or compensating for its absence indirectly.

### A.4.2 LATENT COMPLETION FOR VISION-LANGUAGE MODELS

Recent work has explored recovering missing visual information directly in latent space rather than reconstructing images in pixel space. The Latent Imagination Module (LIM), for example, predicts text-conditioned visual embeddings and injects them into a frozen VLM, demonstrating the potential of latent completion to mitigate performance degradation under missing vision (Gu et al., 2026). In contrast, Latent World Recovery (LWR) models different modalities as partial observations of a shared latent state and adapts downstream representations according to the modalities that remain

available (Xu et al., 2023). These studies establish latent-space completion as a promising direction for handling missing modalities. *MedJEPA* differs in that it preserves the original multimodal reasoning pathway and explicitly predicts the spatial, patch-level visual representation expected by the frozen VLM from the missing image. Rather than adapting downstream representations to accommodate the available modalities, *MedJEPA* focuses on recovering the missing visual signal itself to support downstream medical reasoning.

### A.4.3 JEPA AND LATENT PREDICTIVE REPRESENTATIONS

Joint-Embedding Predictive Architectures (JEPAs) learn by predicting representations of partially observed inputs rather than reconstructing their raw observations. I-JEPA and V-JEPA establish this principle for spatial and spatiotemporal visual representations, respectively (Assran et al., 2023), while V-JEPA 2 extends latent predictive learning toward visual understanding and prediction (Assran et al., 2025). Related latent world models demonstrate that pretrained representations can serve as compact predictive states without pixel-level reconstruction (Zhou et al., 2024), while Causal-JEPA and VLA-JEPA show how latent prediction and masking can encode task-relevant structure (Perera et al., 2025; Sun et al., 2026). *MedJEPA* extends this predictive paradigm from within-modality representation learning to cross-modal latent completion. Rather than predicting a representation of a partially observed input from the same modality, *MedJEPA* predicts the visual representation of an entirely unobserved image from the available clinical text and question. This formulation enables JEPA-style latent prediction to serve as a mechanism for recovering missing visual evidence while retaining the original downstream multimodal reasoning pathway.

### A.4.4 MEDICAL VISION-LANGUAGE REASONING AND ROBUSTNESS

Medical VLMs such as LLaVA-Med and Med-Flamingo have demonstrated strong capabilities in medical image understanding and question answering (Li et al., 2023; Moor et al., 2023). However, recent work shows that medical VQA performance can be influenced by language-only cues and dataset artifacts, motivating explicit no-image baselines and robustness-oriented evaluation. Other methods improve medical reasoning and reliability while retaining access to visual inputs during inference (Singhal et al., 2025; Ikezogwo et al., 2023). These studies motivate evaluating missing-modality methods based on their ability to preserve downstream medical reasoning. In this setting, *MedJEPA* addresses the complementary problem of image absence by replacing the unavailable visual input with a predicted latent representation while preserving the VLM's original multimodal reasoning pathway.

## A.5 BACKGROUND

**Vision–Language Models (VLMs).** VLMs extend language models with visual encoders, enabling multimodal reasoning by jointly processing images and textual inputs (Liu et al., 2023). In the medical domain, VLMs have been applied to visual question answering, clinical image interpretation, and report generation across radiology and other biomedical imaging modalities (Li et al., 2023). Let $\mathcal{I}$ denote the image space and $\mathcal{X}$ the space of textual inputs. A VLM $M_\theta$ maps an image $I \in \mathcal{I}$ and textual context $x \in \mathcal{X}$ to an output distribution:

$$p_\theta(y \mid I, x). \tag{8}$$

For an autoregressive output sequence $y = \{y_1, \ldots, y_T\}$, this distribution can be factorized as

$$p_\theta(y \mid I, x) = \prod_{t=1}^{T} p_\theta(y_t \mid I, x, y_{<t}), \tag{9}$$

where $y_{<t}$ denotes the previously generated tokens. In typical multimodal VLMs, the image is first transformed into a sequence of visual representations that are projected into the language model's embedding space and jointly processed with the textual input. This design makes the visual representation part of the model's native reasoning pathway, but also creates a failure mode when the image is unavailable at inference.

**Missing-Modality Learning.** Multimodal models are generally trained under the assumption that the required modalities are simultaneously available. When one modality is missing, the model must

instead estimate the output using only the remaining observations (Ma et al., 2022b). Let $(I, x, y)$ denote an image–text–answer triple, and let $m \in \{0, 1\}$ indicate whether the image is available. Conventional multimodal reasoning corresponds to

$$p_\theta(y \mid I, x), \qquad m = 1, \tag{10}$$

whereas the missing-image setting requires

$$p_\theta(y \mid x), \qquad m = 0. \tag{11}$$

Existing missing-modality strategies commonly adapt the model to the missing input, retrieve complementary information, or reconstruct the absent modality. These approaches either modify the downstream reasoning pathway or attempt to recreate information at the observation level. For medical VLMs, however, the relevant question is not necessarily how to reproduce the missing image itself, but whether useful visual structure can be recovered in the representation space consumed by the downstream model. This motivates treating the missing visual modality as a latent prediction problem rather than an observation-reconstruction problem.

**Joint-Embedding Predictive Architectures (JEPAs).** Joint-Embedding Predictive Architectures (JEPAs) learn to predict a target representation from a contextual representation rather than reconstructing the target observation directly (Assran et al., 2023; Bardes et al., 2021). Let $E(\cdot)$ denote a frozen representation encoder that maps an observation $x$ into a latent representation $z = E(x)$. Given a context $c$, a predictor $G_\phi$ estimates the target representation as

$$\hat{z} = G_\phi(c), \tag{12}$$

and is trained by minimizing a distance between the prediction and the target representation:

$$\mathcal{L}_{\text{JEPA}} = d(\hat{z}, \text{sg}(z)), \tag{13}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $d(\cdot, \cdot)$ measures representation-level discrepancy. Because the objective operates in latent space, the predictor is not required to reproduce pixel-level details that may be irrelevant to downstream tasks. Spatial or masked variants further allow the model to predict unobserved representations from available context, making JEPAs a natural framework for latent prediction under incomplete observations. MedJEPA extends this principle across modalities: instead of predicting a missing region within the same modality, it predicts the spatial visual representation of an entirely absent image from the available clinical text and question, allowing the recovered latent representation to be injected directly into the frozen VLM's original visual reasoning pathway.

### A.6 Limitations

Despite these results, several limitations remain. First, the evaluation is limited to two benchmarks, PMC-VQA and ScienceQA, and the available textual conditioning is provided in the form of figure captions or task hints; real clinical workflows may contain substantially noisier, incomplete, or differently structured textual context. Broader evaluation across additional medical modalities, institutions, clinical tasks, and naturally occurring missing-image scenarios is therefore needed to establish generalization beyond the benchmark setting.

Second, MedJEPA cannot recover visual information that has no correlate in the available text. Its prediction should therefore be interpreted as a *text-conditioned visual prior*, rather than an exact estimate of the original image representation. This intrinsic ambiguity is also reflected in our retrieval analysis: the predicted representations exhibit modality-level structure but do not reliably identify the exact source image. Consequently, performance can remain limited when the answer depends on fine-grained visual details that are not described by the textual context.

Third, the learned representation is tied to the visual feature space consumed by the underlying VLM. Our predictor targets frozen CLIP-ViT-L/14336 representations and injects them through the corresponding multimodal projector, which makes the current implementation dependent on this visual representation interface and limits direct portability to VLMs with substantially different visual encoders or tokenization schemes. Extending latent prediction across heterogeneous visual representation spaces remains an important direction.

Finally, confidence-gated injection provides a conservative mechanism for suppressing potentially harmful latent estimates, but the cascade relies on the downstream VLM's confidence margin as a

proxy for whether the injected representation is useful. As model confidence is not equivalent to correctness, the gate cannot guarantee that an accepted latent completion is clinically or factually correct. In addition, iterative refinement and multi-stage injection introduce extra computation relative to direct text-only inference. Future work should therefore investigate calibrated uncertainty estimates, more efficient latent prediction, and verification mechanisms that assess the reliability of predicted visual evidence independently of the VLM's own confidence.

### A.7 Additional Method Details

This section provides implementation details omitted from the main paper.

**Architecture Details.**

MedJEPA uses the frozen CLIP-ViT-L/14336 visual encoder and the frozen PMC-LLaMA 7B embedding layer. The CLIP encoder produces 576 visual tokens of dimension 1024. The visual target is extracted from the **[final target layer; e.g., penultimate** $(-2)$ **layer]** of CLIP. The PMC-LLaMA embedding layer produces 4096-dimensional token embeddings, which are projected to the 1024-dimensional MedJEPA space before entering the TextContextEncoder (Radford et al., 2021; Wu et al., 2024).

The TextContextEncoder consists of four Pre-LN Transformer layers with 16 attention heads, a $4\times$ feed-forward expansion, and hidden dimension 1024. The PatchContextEncoder adds learnable two-dimensional positional embeddings to visible CLIP patch tokens and applies no additional feature projection. The JEPAPredictor is a four-layer Pre-LN Transformer decoder with 8 attention heads, a $2\times$ feed-forward expansion, and 576 learnable spatial query embeddings. The complete set of trainable MedJEPA parameters contains approximately 85M parameters. The CLIP encoder, PMC-LLaMA embedding layer, and downstream VLM remain frozen throughout training.

**Optimization Details.**

All trainable MedJEPA parameters are optimized jointly with AdamW using a learning rate of $1 \times 10^{-4}$, weight decay of 0.05, and $\beta = (0.9, 0.95)$. We use cosine learning-rate decay with 400 warm-up steps and gradient clipping with a maximum norm of 1.0. Training uses a batch size of 1 with eight gradient-accumulation steps, corresponding to an effective batch size of 8.

**Latent Objective Details.**

The visual target is $L_2$ normalized before computing representation-level losses and is detached from gradient computation:

$$Z^* = \mathrm{sg}\left[\frac{Z}{\|Z\|_2}\right]. \tag{14}$$

The primary prediction loss is computed using Smooth-L1 between the normalized predicted and target tokens:

$$\mathcal{L}_{\mathrm{pred}} = \mathrm{SmoothL1}\left(\widehat{Z}^*, Z^*\right). \tag{15}$$

The refinement loss applies the same objective to the second-pass prediction and is evaluated on the masked positions:

$$\mathcal{L}_{\mathrm{refine}} = \mathrm{SmoothL1}\left(\widehat{Z}^{(2)*}_{\mathcal{M}}, Z^*_{\mathcal{M}}\right). \tag{16}$$

For global alignment, textual and visual representations are mean-pooled over their token dimensions. The SigLIP-style objective is computed from the paired text–visual representations, with a rolling memory bank of 128 previous visual embeddings providing additional negative examples (Zhai et al., 2023). The direct alignment term is

$$\mathcal{L}_{\mathrm{align}} = 1 - \cos\left(\overline{C}_{\mathrm{text}}, \overline{Z}\right). \tag{17}$$

To prevent collapse across spatial positions, we use a VICReg-style variance penalty (Bardes et al., 2021):

$$\mathcal{L}_{\mathrm{var}} = \frac{1}{d}\sum_{j=1}^{d} \max\left(0,\, \gamma - \mathrm{Std}(\widehat{Z}_{:,j})\right), \tag{18}$$

where $\gamma = 0.5$.

The weighted combination used in all experiments is the objective reported in the main paper:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + 0.5\mathcal{L}_{\text{refine}} + 0.20\mathcal{L}_{\text{sig}} + 0.20\mathcal{L}_{\text{align}} + 0.30\mathcal{L}_{\text{var}}. \tag{19}$$

**Refinement and Inference Details.**

MedJEPA uses two refinement passes at inference. The first pass generates the complete visual representation from text alone:

$$\widehat{Z}^{(1)} = G_\phi\left(C_{\text{text}}, \varnothing, \mathcal{M}_{\text{all}}\right). \tag{20}$$

The prediction is detached and passed through the PatchContextEncoder before the second pass:

$$C^{(1)}_{\text{patch}} = F_{\text{patch}}\left(\text{sg}[\widehat{Z}^{(1)}]\right). \tag{21}$$

The second pass then produces the refined representation conditioned on the text and the previous latent estimate.

For sparse injection, we measure the stability of each predicted patch between the two refinement passes:

$$s_i = \frac{1}{1 + \left\|\widehat{z}^{(2)}_i - \widehat{z}^{(1)}_i\right\|_2^2}. \tag{22}$$

The $K = 144$ highest-scoring patches are selected for sparse injection.

For LLaVA-1.5, the predicted 1024-dimensional visual tokens are mapped through the frozen multimodal projector:

$$P_{\text{mm}} : \mathbb{R}^{1024} \to \mathbb{R}^{4096}. \tag{23}$$

The resulting tokens are supplied at the model's visual-token position.

### A.8 Retrieval Task Analysis

Figure 5 illustrates representative retrieval results, with the corresponding retrieved image highlighted by the red bounding box.

### A.9 Models Used

We evaluate MedJEPA across a diverse set of VLMs to assess whether latent visual prediction remains effective across different model families and scales. The evaluated models include both general-purpose and medical VLMs, covering recent large-scale multimodal systems as well as the backbone used for our primary experiments.

**General-purpose VLMs.** We include *GLM-5.3-Flash*, *Qwen3-VL-235B*, *DeepSeek-V4.1-Flash*, *Qwen3-VL-8B*, and *LLaMA-4-Maverick* as contemporary general-purpose VLMs. These models provide a broader reference for image-absent reasoning and allow us to examine whether the missing-image setting is specific to medical VLMs or reflects a more general limitation of multimodal reasoning. For these models, we report full-image and text-only performance as reference points; MedJEPA is not applied to models for which the required visual representation interface is not available in our implementation.

**Medical and open-weight VLMs.** For direct evaluation with MedJEPA, we consider *Phi-3.5-Vision-Instruct*, *LLaVA-1.5-13B*, and *MedLLaVA-7B*. These models provide complementary architectural and domain settings for evaluating latent visual representation prediction. In particular, *MedLLaVA-7B* serves as our primary backbone because its visual-token interface is compatible with the CLIP-ViT-L/14@336 representations predicted by MedJEPA, enabling direct replacement of the unavailable visual tokens without modifying the underlying VLM. Unless otherwise stated, the MedJEPA results reported in the main experiments therefore use MedLLaVA-7B as the primary backbone.

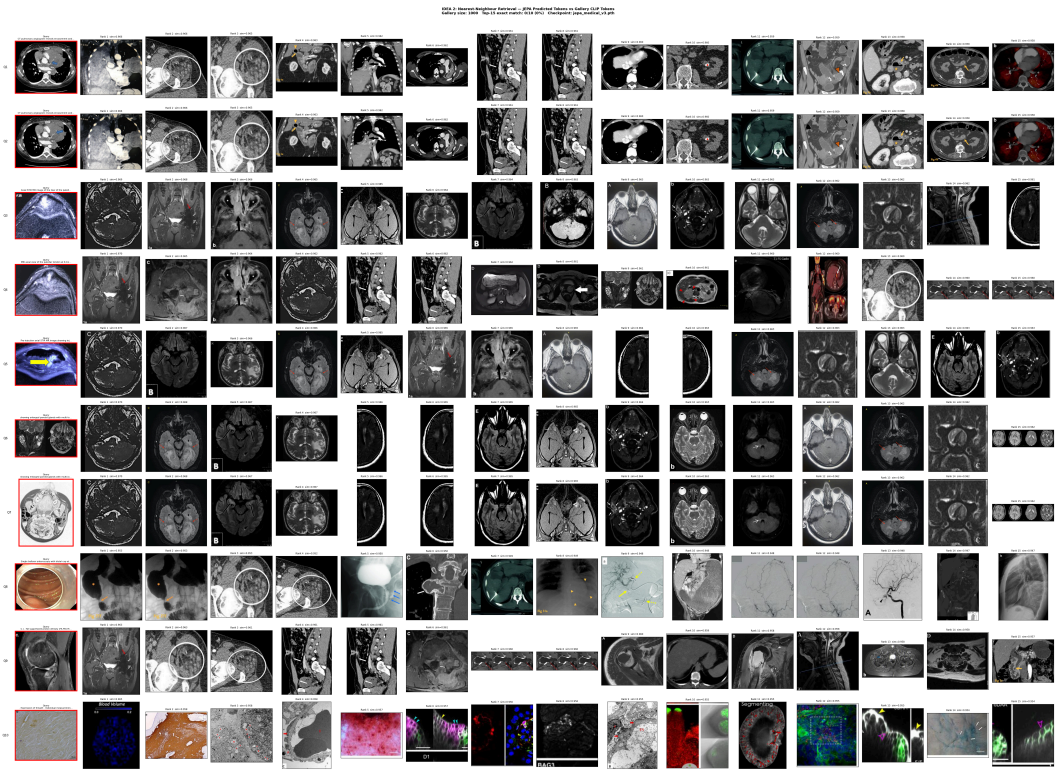


Figure 5: Text-conditioned visual representations are used to retrieve images from a gallery of 1,000 test images using cosine similarity. For each query, the top-15 retrieved images exhibit consistent modality-level correspondence with the textual description, including CT, MRI, endoscopy, and histology, despite the absence of the image and explicit modality supervision. The results indicate that MedJEPA captures coarse visual structure and modality-specific information in the pretrained CLIP space, while not reliably recovering the exact missing image.

**Evaluation protocol.** For each backbone, we first evaluate the model in the full-image setting, where the original image is available, and in the text-only setting, where the visual input is removed. We then apply MedJEPA to the models for which compatible visual-token injection is supported and evaluate the resulting image-absent predictions. This setup separates three quantities: the performance ceiling obtained with the original image, the degradation caused by removing the image, and the amount of visual reasoning capability recovered by the predicted latent representation. As shown in Table 1, MedJEPA improves image-absent performance for all three directly evaluated backbones, demonstrating that the approach is not restricted to a single VLM architecture.

### A.10 Addtional Results

The results on different hyperpamaters like VICReg, SigLIP and alignment weight are shown in Table 5 , 7 and 6 respectively .

Table 5: Sensitivity to the VICReg variance threshold $\gamma$ on PMC-VQA test examples. The default configuration is highlighted in blue.

| $\gamma$ | **Acc. (%)** | **Val. Cos-Sim** |
|---|---|---|
| **0.5 (default)** | **69.82** | **0.5404** |
| 1.0 | 69.74 | 0.5406 |
| 1.5$^{\dagger}$ | 69.88 | 0.5399 |
| 2.0 | 69.62 | 0.5403 |
| 3.0 | 69.80 | 0.5404 |

$^{\dagger}$ Above-default setting.

Table 6: Sensitivity to the alignment-loss weight $\lambda_{\text{align}}$ on PMC-VQA test examples. The default configuration is highlighted in blue.

| $\lambda_{\text{align}}$ | **Acc. (%)** | **Val. Cos-Sim** |
|---|---|---|
| 0.00$^{\dagger\ddagger}$ | 69.84 | 0.5393 |
| 0.05$^{\dagger}$ | 69.90 | 0.5394 |
| **0.10 (default)** | **69.82** | **0.5404** |
| 0.20 | 69.68 | 0.5402 |
| 0.30 | 69.78 | 0.5411 |

$^{\dagger}$ Above-default setting. $^{\ddagger}$ Zero-weight setting.

Table 7: Sensitivity to the SigLIP-loss weight $\lambda_{\text{siglip}}$ on PMC-VQA test examples. The default configuration is highlighted in blue.

| $\lambda_{\text{siglip}}$ | **Acc. (%)** | **Val. Cos-Sim** |
|---|---|---|
| 0.00 | 69.78 | 0.5364 |
| 0.10$^{\dagger}$ | 69.84 | 0.5412 |
| **0.20 (default)** | **69.82** | **0.5404** |
| 0.30$^{\dagger}$ | 69.90 | 0.5398 |
| 0.40$^{\dagger}$ | 69.92 | 0.5381 |

$^{\dagger}$ Above-default setting.