Can LLM Understand Medical Imaging With Long Context?

[1]Jiahua Zhang, [2]Jinghao Liang, [3]Yiwen Cai, [3]Jingchun Ni, [2]Jianxing He
[1]Nanyang Technological University [2]Guangzhou Medical University [3]Guangdong Medical University

***Abstract*—Medical applications place stricter demands on multi-image context and high-resolution imagery, which require decisions that depend on long temporal horizons and complex examination histories. In this paper, we propose an encoder-and-LLM-frozen framework that couples frozen CT/WSI foundation encoders with a long-context LLM for spatiotemporal reasoning without fine-tuning either the visual encoders or the LLM. To mitigate noisy, low-yield visual evidence, we first use CAM-based scoring to retain informative CT slices and WSI regions, then compress encoder features with PCA and align them to the LLM embedding space via an optimal-transport calibration. The aligned vectors are injected as pseudo-tokens and interleaved with a handful of concise textual exemplars, enabling few-shot in-context inference for diagnosis, staging, and longitudinal response assessment directly within the decoder. This design (i) preserves fine-grained radiologic and histologic cues while scaling to long visual contexts, (ii) supports modular swap-in of lightweight projection strategies under different resource budgets, and (iii) approaches supervised multimodal baselines on public cancer cohorts while avoiding cross-modal fine-tuning. Overall, our results suggest a practical path toward privacy-conscious multimodal decision support that requires only inference-time preparation and alignment of embeddings.**

***Index Terms*—encoder-frozen multimodal alignment, optimal transport, class activation mapping, long-context LLM, spatiotemporal reasoning, CT, whole-slide imaging**

I. Introduction

Large Language Models (LLMs) have demonstrated remarkable versatility in leveraging test-time computation, allowing them to dynamically adapt to new tasks through prompting and in-context scheduling [1]–[3], and perform complex downstream reasoning without any additional training [4]–[7]. However, medical applications place stricter demands on multi-image context and high-resolution imagery, and often require decisions that depend on long temporal horizons and complex examination histories [8]–[10].

To preserve accuracy in long-context settings, recent studies curate training corpora that explicitly include multi-turn interactions and alternative training strategies [11]–[15] to counteract context length induced performance loss. To alleviate the computational burden, prior work [16] reduces communication overhead to improve the efficiency of multi-node distributed training. Nevertheless, accelerating the core intra-node decoding for long visual context, without materially compromising performance, remains an open problem. Long-context multimodal decoders such as leveraging Transformer–Mamba [17] pipelines to efficiently handle hundreds of visual tokens, yet they still rely heavily on learned adapters or paired supervision. Such reliance is difficult to scale in clinical settings, where datasets are often modest in size and exhibit substantial cross-center heterogeneity.

At the same time, clinical data typically exhibit complex spatial and temporal dependencies, highly heterogeneous sampling schemes across centers, and partially unreliable or low-yield spatial information. Learning systems are thus forced to reason over irregular feature scales and missing observations [18], [19]. In digital pathology, gigapixel WSIs are usually partitioned into millions of tiles, which disrupts native tissue architecture and dilutes morphological cues that are crucial for staging aggressive tumors [20], [21], thereby weakening a model's ability to reconstruct long-range spatial relationships. Similar difficulties arise in longitudinal CT follow-up: motion artefacts and irregular slice thickness or spacing complicate downstream modeling, while the limited spatial resolution of CT constrains the characterization of peritumoral microstructure. Furthermore, radiologic features of malignant tumors often strongly overlap with benign findings, and the most informative cues may appear only in a few key slices, making it challenging for generic 3D or video-style foundation models to stably extract discriminative imaging patterns.

These tensions lead to a central question: **Can we pair informative spatial and temporal information with LLM tokens for reliable spatiotemporal reasoning under long visual contexts?**

To address this question, we propose a scalable, encoder-and-LLM-frozen multi-task learning framework for medical spatiotemporal imaging features. Specifically, we leverage optimal transport (OT) to directly align expressive representations from pretrained medical foundation models to the LLM token embedding space [22]. The aligned visual vectors are injected into the embedding layer as pseudo-tokens and combined with a small number of textual exemplars, enabling the LLM to perform few-shot reasoning without updating any encoder or LLM parameters with a lightweight linear projector is trained on a small labeled set. This design preserves both local morphology and long-range spatial structure, naturally scales to large visual contexts, and circumvents expensive and fragile cross-modal fine-tuning.

Our contributions are as follows:

- We introduce a cancer-centric token embedding alignment framework that maps volumetric CT and aggregated WSI features into a unified token space via OT-aligned

Corresponding author: Jianxing He. Email: jianxinghe@gzhmu.edu.cn

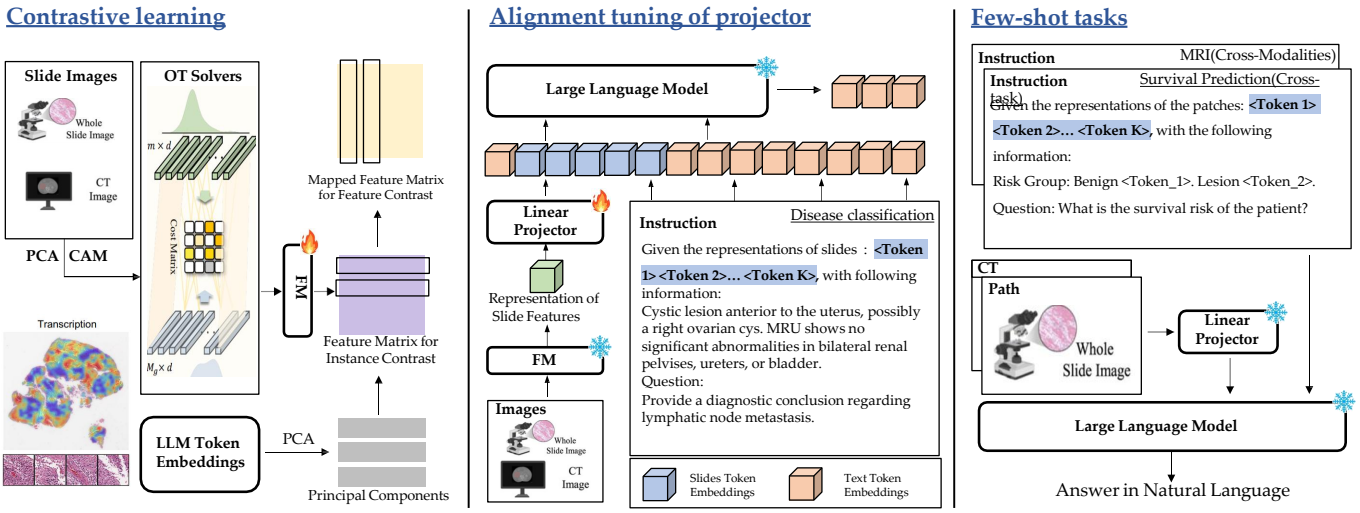

Fig. 1. We adopt optimal transport to construct an efficient pairing between image representations and LLM tokens while keeping both visual encoders and the LLM frozen.

embedding injection, where CAM-based key frame extraction filters redundant information and OT marginal constraints enforce global structure consistency across spatially dispersed diagnostic regions.

- We propose a feature-wise contrastive alignment strategy that aligns semantic dimensions rather than instances, ensuring transferability across datasets with high intra-class variance. We provide a complete mathematical formulation of the alternating OT–contrastive optimization with PCA-initialized alignment (Algorithm 1).
- We demonstrate that combining OT-aligned visual embeddings with concise, semantically rich text-level descriptions enables competitive few-shot diagnostic performance across five TCGA cancer types and a private lung cancer cohort, while keeping both visual encoders and the LLM entirely frozen.

II. Methodology

We propose an encoder-and-LLM-frozen framework that aligns long-context medical imaging modalities with LLMs through three stages: (1) CAM-based key frame extraction, (2) optimal transport matching between visual features and LLM token embeddings, and (3) feature-wise contrastive alignment. **Problem Setup.** Given a CT volume $\mathcal{V}_{CT} \in \mathbb{R}^{D\times H\times W}$ or WSI patches $\mathcal{P}_{WSI}$, pretrained encoders $f_{CT}(\cdot)$ and $f_{WSI}(\cdot)$ extract features $\mathbf{F}_{CT}, \mathbf{F}_{WSI} \in \mathbb{R}^{N\times F_V}$. Our goal is to align these visual representations with the LLM's token embedding space $\mathbb{R}^{F_L}$ for few-shot reasoning without fine-tuning.

A. CAM-based Key Frame Extraction

To filter redundant information in CT volumes and WSIs, we employ CAM on the final convolutional block of the pretrained 3D encoder. The class weights w_k are the global-average-pooled gradients of the target class score with respect to activations $\mathbf{A}_k$. The 3D CAM is $\mathbf{M}^{3D}_{CAM} = \sum_{k=1}^{K} w_k \mathbf{A}_k$, with key slices selected by:

$$\mathcal{S}_{key} = \text{TopK}\left(\left\{ s_i = \sum_{h,w} \mathbf{M}^{3D}_{CAM}(i,h,w) \right\}_{i=1}^{D'}, K_{CT} \right), \tag{1}$$

where $\mathbf{A}_k \in \mathbb{R}^{D'\times H'\times W'}$ are activation maps, w_k are class-specific weights, and s_i scores each slice's diagnostic relevance. For WSI, we use attention-based MIL [23] to identify tumor microenvironment patches, selecting top-K_{WSI} instances by attention weights. This yields compact feature sets $\mathbf{F}^{key}_{CT} \in \mathbb{R}^{K_{CT}\times F_V}$ and $\mathbf{F}^{key}_{WSI} \in \mathbb{R}^{K_{WSI}\times F_V}$.

B. Optimal Transport Matching

After extracting key visual instances via CAM, we align these medical imaging features with the LLM's token embedding space using optimal transport [22]. Unlike conventional attention mechanisms that compute only local pairwise similarity, OT enforces marginal constraints that encourage global trade-offs among instances, enabling the model to capture long-range structural dependencies across spatially dispersed tumor microenvironment components in WSI or key anatomical landmarks across CT slices. We solve OT via the Sinkhorn-Knopp algorithm with entropic regularization (ϵ=0.1), converging in approximately 50–100 inner iterations.

To establish a reference semantic space, we extract medical vocabulary embeddings $\mathbf{E}_{med} = \{\mathbf{e}_k = \text{LLM}_{\text{embed}}(v_k)\}_{k=1}^{M_{med}}$ from the frozen LLM's embedding layer, where v_k includes anatomical structures, disease names, diagnostic descriptors, and clinical procedures. This medical token embedding bag serves as the target distribution for cross-modal matching, providing semantically meaningful anchors in the LLM's representation space.

1) OT Formulation for Visual-to-Text Matching: Following the multiple instance learning (MIL) paradigm and inspired

by multimodal optimal transport [24], we formulate the visual feature bag (from CT or WSI) and the LLM token embedding bag as two probability distributions to be matched. For CT modality, we have the visual feature bag $\mathbf{F}_{CT}^{key} = \left\{\mathbf{f}_{CT}^{(i)}\right\}_{i=1}^{K_{CT}}$ extracted from key frames, and we wish to match it with the medical token embeddings $\mathbf{E}_{med}$.

We define probability distributions over the visual features and token embeddings:

$$\boldsymbol{\mu}_{vis} = \left[\frac{1}{K_{CT}}, \ldots, \frac{1}{K_{CT}}\right]^{\top} \in \mathbb{R}^{K_{CT}}, \tag{2}$$

$$\boldsymbol{\mu}_{text} = \left[\frac{1}{M_{med}}, \ldots, \frac{1}{M_{med}}\right]^{\top} \in \mathbb{R}^{M_{med}}. \tag{3}$$

To compute the cost matrix, we first need to project the visual features into the same dimensional space as the LLM embeddings. We use a lightweight linear transformation:

$$\hat{\mathbf{f}}_{CT}^{(i)} = \mathbf{W}_{align}\mathbf{f}_{CT}^{(i)}, \quad \mathbf{W}_{align} \in \mathbb{R}^{F_L \times F_V}, \tag{4}$$

where $\mathbf{W}_{align}$ can be initialized with PCA-derived projection matrices or learned through feature-wise contrastive learning (detailed in Sec. II-C).

The cost matrix $\mathbf{C}_{CT} \in \mathbb{R}^{K_{CT} \times M_{med}}$ captures the pairwise dissimilarity between projected visual features and token embeddings:

$$C_{CT}^{ik} = \left\|\hat{\mathbf{f}}_{CT}^{(i)} - \mathbf{e}_k\right\|_2^2 = \left\|\mathbf{W}_{align}\mathbf{f}_{CT}^{(i)} - \mathbf{e}_k\right\|_2^2. \tag{5}$$

The optimal transport problem seeks a transport plan $\mathbf{P}_{CT}^* \in \mathbb{R}^{K_{CT} \times M_{med}}$ that minimizes the total transport cost:

$$\mathbf{P}_{CT}^* = \arg\min_{\mathbf{P} \in \Pi(\boldsymbol{\mu}_{vis}, \boldsymbol{\mu}_{text})} \langle \mathbf{C}_{CT}, \mathbf{P} \rangle = \sum_{i,k} C_{CT}^{ik} P_{ik}, \tag{6}$$

where $\Pi(\boldsymbol{\mu}_{vis}, \boldsymbol{\mu}_{text})$ denotes the set of admissible transport plans with marginal constraints ensuring total mass equality, which equal to

$$\left\{\mathbf{P} \in \mathbb{R}_+^{K_{CT} \times M_{med}} \mid \mathbf{P}\mathbf{1}_{M_{med}} = \boldsymbol{\mu}_{vis}, \mathbf{P}^{\top}\mathbf{1}_{K_{CT}} = \boldsymbol{\mu}_{text}\right\} \tag{7}$$

Similarly, for WSI modality, we solve:

$$\mathbf{P}_{WSI}^* = \arg\min_{\mathbf{P} \in \Pi(\boldsymbol{\mu}_{WSI}, \boldsymbol{\mu}_{text})} \langle \mathbf{C}_{WSI}, \mathbf{P} \rangle, \tag{8}$$

where $\mathbf{C}_{WSI}^{jk} = \left\|\mathbf{W}_{align}\mathbf{f}_{WSI}^{(j)} - \mathbf{e}_k\right\|_2^2$. The optimal transport plan $\mathbf{P}_{CT}^*$ provides a principled global matching that respects structural relationships between visual features and semantic token embeddings. We leverage this transport plan to identify the most informative visual instances for LLM reasoning. For each visual feature $\mathbf{f}_{CT}^{(i)}$, we compute its semantic alignment score by marginalizing over all token embeddings:

$$\alpha_i^{CT} = \sum_{k=1}^{M_{med}} P_{CT}^{*ik}, \tag{9}$$

which measures how strongly the i-th visual feature is matched to the medical vocabulary in the LLM's semantic space. We then select the top-$\hat{K}_{CT}$ instances with the highest alignment scores:

$$\hat{\mathbf{F}}_{CT} = \text{TopK}\left(\left\{(\alpha_i^{CT}, \mathbf{f}_{CT}^{(i)})\right\}_{i=1}^{K_{CT}}, \hat{K}_{CT}\right), \tag{10}$$

where $\hat{K}_{CT} \ll K_{CT}$ represents a compact set of the most semantically informative features. This selection is performed without any label supervision, relying solely on the geometric structure preserved by optimal transport. The same process is applied to WSI features to obtain $\hat{\mathbf{F}}_{WSI}$, yielding a unified cross-modal representation that bridges medical imaging and language modalities.

2) Micro-Batch OT for Computational Efficiency: Due to the potentially large number of visual instances (K_{CT}, K_{WSI}) and token embeddings (M_{med}), solving the current OT problem can be computationally expensive. Following the unbalanced mini-batch OT (UMBOT) strategy [24], we approximate the global OT solution by averaging solutions over multiple micro-batches.

Specifically, we partition the visual feature set into B mini-batches: $\left\{\mathbf{F}_{CT}^{(b)}\right\}_{b=1}^{B}$, where each micro-batch contains $m = K_{CT}/B$ instances. For each mini-batch b, we solve a smaller OT problem with entropic regularization (coefficient ϵ) and marginal relaxation (coefficient τ):

$$\begin{aligned}\mathcal{W}^{(b)}(\mathbf{F}_{CT}^{(b)}, \mathbf{E}_{med}) = &\min_{\mathbf{P}^{(b)} \in \Pi(\boldsymbol{\mu}_{vis}^{(b)}, \boldsymbol{\mu}_{text})} \left\langle \mathbf{C}_{CT}^{(b)}, \mathbf{P}^{(b)} \right\rangle \\ &+ \epsilon \text{KL}(\mathbf{P}^{(b)} \| \boldsymbol{\mu}_{vis}^{(b)} \otimes \boldsymbol{\mu}_{text}) \\ &+ \tau \left(D_\phi(\mathbf{P}_{vis}^{(b)} \| \boldsymbol{\mu}_{vis}^{(b)}) + D_\phi(\mathbf{P}_{text}^{(b)} \| \boldsymbol{\mu}_{text})\right),\end{aligned} \tag{11}$$

where $\mathbf{P}_{vis}^{(b)}$ and $\mathbf{P}_{text}^{(b)}$ are the marginals of $\mathbf{P}^{(b)}$, τ=0.5 is the marginal relaxation coefficient, and D_ϕ denotes the Csiszár divergence instantiated as KL divergence ($\phi(t) = t\log t - t + 1$). The unbalanced formulation allows for more robust approximation by relaxing the strict marginal constraints.

The aggregated transport plan is computed as: $\bar{\mathbf{P}}_{CT} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{P}^{*(b)}$.

This micro-batch approximation significantly reduces computational complexity from $\mathcal{O}(K_{CT}^3 \log K_{CT})$ to $\mathcal{O}(B \cdot m^2) = \mathcal{O}(K_{CT} \cdot m)$, where $m \ll K_{CT}$, making it practical for large-scale medical imaging applications.

C. Feature-wise Contrastive Alignment

To learn $\mathbf{W}_{align}$ without fine-tuning, we adopt feature-wise contrastive learning that aligns semantic dimensions rather than instances, ensuring transferability across datasets with high intra-class variance.

We first extract principal components from LLM token embeddings $\mathbf{E}_{LLM} \in \mathbb{R}^{N_t \times F_L}$ to obtain semantic coordinate axes:

$$\mathbf{C} = \text{PCA}\left(\mathbf{E}_{LLM}, P\right) \in \mathbb{R}^{P \times F_L}, \tag{12}$$

retaining 95% variance. Visual features are projected to this subspace: $\widetilde{\mathbf{F}}_{visual} = \mathbf{F}_{visual} \times \mathbf{C}^{\top}$. We then align feature

dimensions (columns $\boldsymbol{\phi}_i$) with text embeddings from medical reports ($\boldsymbol{\psi}_i$) with $\mathcal{L}_{fea}$ equal to:

$$\frac{1}{P}\sum_{i=1}^{P}\log\frac{\exp\left(\theta\left(\boldsymbol{\phi}_i,\boldsymbol{\psi}_i\right)/\tau\right)}{\sum_{j=1}^{P}\left[\exp\left(\theta\left(\boldsymbol{\phi}_i,\boldsymbol{\phi}_j\right)/\tau\right)+\exp\left(\theta\left(\boldsymbol{\phi}_i,\boldsymbol{\psi}_j\right)/\tau\right)\right]}, \tag{13}$$

where θ denotes cosine similarity and τ is a temperature parameter. The alternating optimization proceeds as follows: (i) initialize $\mathbf{W}_{align} \leftarrow \mathbf{C}$; (ii) fix $\mathbf{W}_{align}$ and solve OT (Eq. 6) to obtain $\mathbf{P}^*$ and select informative instances; (iii) fix the selected instances and update $\mathbf{W}_{align}$ by one gradient step on $\mathcal{L}_{fea}$; repeat (ii)–(iii) for 3–5 outer iterations. Because PCA initialization places visual features near the LLM semantic subspace from the start, convergence is fast. No gradients pass through the visual encoders or LLM at any point.

D. Alignment Tuning and Zero-Shot Inference

After obtaining aligned visual features, we train a lightweight linear projector $\mathbf{W}_{proj} \in \mathbb{R}^{K_{token}\times\hat{K}}$ to aggregate selected instances into a fixed number of pseudo-tokens (typically $K_{token} \in \{16, 64\}$):

$$\mathbf{E}_{visual} = \mathbf{W}_{proj}\,\widetilde{\mathbf{F}}_{visual} \times \mathbf{C} \in \mathbb{R}^{K_{token}\times F_L}. \tag{14}$$

The projector is trained on a small labeled dataset from the source domain by minimizing cross-entropy loss:

$$\mathcal{L}_{proj} = -\sum_{i=1}^{N_{train}} \log P_{LLM}\Big(y_i \mid [\text{Inst.}] \oplus [\mathbf{E}^{(i)}_{visual}] \oplus [\text{Ctx}_i]\Big), \tag{15}$$

while keeping visual encoders and LLM entirely frozen. Only $K_{token} \times \hat{K}$ parameters are updated, enabling convergence even with limited labeled data. For zero-shot inference on unseen datasets or tasks, these pseudo-tokens are injected into unified instruction templates: Prompt = [Task Instruction] $\oplus$ $[\mathbf{E}_{visual}]$ $\oplus$ [Clinical Context], enabling the LLM to perform medical reasoning across node-level (e.g., diagnosis prediction) and edge-level (e.g., survival analysis) tasks without further parameter updates. The complete pipeline is summarized in Algorithm 1.

Algorithm 1 OT Matching & Feature-wise Contrastive Alignment

Require: Key visual features $\mathbf{F}^{key} \in \mathbb{R}^{K\times F_V}$, medical token embeddings $\mathbf{E}_{med} \in \mathbb{R}^{M_{med}\times F_L}$, LLM embeddings $\mathbf{E}_{LLM} \in \mathbb{R}^{N_t\times F_L}$, entropic regularization ϵ, marginal relaxation τ, mini-batch count B, outer iterations T=3–5
Ensure: Aligned visual features $\widetilde{\mathbf{F}} \in \mathbb{R}^{\hat{K}\times P}$

Compute PCA: $\mathbf{C} = \text{PCA}(\mathbf{E}_{LLM}, P) \in \mathbb{R}^{P\times F_L}$
Initialize $\mathbf{W}_{align} \leftarrow \mathbf{C}$
for $t = 1, \ldots, T$ **do**
  // OT-based instance selection (fix $\mathbf{W}_{align}$)
  for $i = 1, \ldots, K$ **do**
    $\hat{\mathbf{f}}^{(i)} \leftarrow \mathbf{W}_{align}\,\mathbf{f}^{(i)}$
    for $k = 1, \ldots, M_{med}$ **do**
      $C^{ik} \leftarrow \|\hat{\mathbf{f}}^{(i)} - \mathbf{e}_k\|_2^2$
    end for
  end for
  Partition $\mathbf{F}^{key}$ into B mini-batches $\{\mathbf{F}^{(b)}\}_{b=1}^{B}$
  for $b = 1, \ldots, B$ **do**
    Solve unbalanced OT via Sinkhorn (ϵ, τ, KL) $\rightarrow$ $\mathbf{P}^{*(b)}$
  end for
  Aggregate: $\bar{\mathbf{P}} \leftarrow \frac{1}{B}\sum_{b=1}^{B}\mathbf{P}^{*(b)}$
  Alignment scores: $\alpha_i \leftarrow \sum_k \bar{P}_{ik}$, $\forall i$
  $\hat{\mathbf{F}} \leftarrow \text{TopK}(\{(\alpha_i, \mathbf{f}^{(i)})\}_{i=1}^{K},\ \hat{K})$
  // Feature-wise contrastive update (fix instances)
  $\widetilde{\mathbf{F}} \leftarrow \hat{\mathbf{F}} \times \mathbf{C}^{\top}$ *// PCA projection*
  $\mathbf{W}_{align} \leftarrow \mathbf{W}_{align} - \eta\,\nabla_{\mathbf{W}_{align}}\mathcal{L}_{fea}$ (Eq. 13)
end for
return $\widetilde{\mathbf{F}}$, $\mathbf{W}_{align}$

III. Experiment

A. Experimental Setup

We explore the feasibility of directly embedding foundation model representations without projector training. Specifically, representations for CT data were extracted using a pre-trained 3D ResNet [25], while patch-level features for WSI were extracted using a pre-trained Transformer encoder [26], where the models are fine-tuned on clinical datasets.

a) Dataset: We construct a multi-center pan-cancer dataset emphasizing cross-modal alignment for evaluation. The public cohort is derived from TCGA [27], comprising histopathology whole-slide images (WSI) and computed tomography (CT) across five cancer types (BLCA, BRCA, UCEC, KIRC, LUAD) with 2,731 patients, where LUAD and UCEC are used for self-supervised pretraining along with partial data from clinical ACC data. To further validate cross-site generalization, we additionally collect a private lung cancer cohort from the First Affiliated Hospital of Guangdong Medical University, containing 506 patients with paired WSI and CT scans.

b) Competing Methods: We compare against four categories of baselines: (i) *Unimodal methods*: Cox-PH, DeepSurv for CT-based survival analysis, 3D-ResNet and Radiomic features for pathology-based prediction; (ii) *Domain adaptation approaches*: Dom-Adv and Deep CORAL that explicitly address distribution shift between modalities; (iii) *Multimodal LLM methods*: LLaVA-Med [28], which adapts visual instruction tuning for biomedical images; LongLLaVA [29], which extends context length for processing long visual sequences; and Mantis [30], which enables interleaved multi-image reasoning; (iv) *Advanced domain adaptation*: DANN and MMD-Trans that employ adversarial training and maximum mean discrepancy for cross-modal alignment. All methods are evaluated under identical few-shot settings ($k \in \{3, 5, 10\}$) to ensure fair comparison.

B. Results and Discussion

RQ1: How effective is our framework in cross-modal and cross-site transfer?

TABLE I

C-INDEX (MEAN ± STD) PERFORMANCE FOR COMPARISON OF MULTIMODAL AND UNIMODAL METHODS, WITH SAME PRETRAINING DATA USED AND 5 FOLD SAMPLES USED FOR EACH METHODS FINETUNING SETTING.

Model	Patho	CT	BLCA ($N = 373$)	UCEC ($N = 480$)	BRCA ($N = 511$)	GBMLGG ($N = 569$)	KIRC ($N = 247$)	LUNG ($N = 506$)
Cox-PH		✓	0.605 ± 0.028	0.610 ± 0.067	0.663 ± 0.047	0.649 ± 0.019	0.597 ± 0.053	0.685 ± 0.041
DeepSurv		✓	0.644 ± 0.038	0.632 ± 0.070	0.642 ± 0.044	0.654 ± 0.021	0.623 ± 0.028	0.673 ± 0.039
3D-ResNet	✓		0.585 ± 0.055	0.595 ± 0.072	0.644 ± 0.042	0.604 ± 0.032	0.606 ± 0.068	0.679 ± 0.045
Radiomic	✓		0.491 ± 0.048	0.511 ± 0.050	0.583 ± 0.066	0.670 ± 0.036	0.535 ± 0.057	0.618 ± 0.058
Dom-Adv	✓	✓	0.622 ± 0.031	0.638 ± 0.049	0.681 ± 0.039	0.720 ± 0.024	0.633 ± 0.038	0.715 ± 0.035
Deep CORAL	✓	✓	0.658 ± 0.039	0.645 ± 0.038	0.635 ± 0.050	0.720 ± 0.031	0.645 ± 0.034	0.683 ± 0.042
LLaVA-Med	✓	✓	0.645 ± 0.037	0.655 ± 0.042	0.695 ± 0.048	0.785 ± 0.029	0.660 ± 0.041	0.728 ± 0.040
LongLLaVA	✓	✓	0.668 ± 0.039	0.685 ± 0.038	0.708 ± 0.046	0.778 ± 0.032	0.688 ± 0.037	0.745 ± 0.038
Mantis	✓	✓	0.692 ± 0.041	0.715 ± 0.041	0.718 ± 0.049	0.770 ± 0.034	0.712 ± 0.039	0.758 ± 0.041
DANN	✓	✓	0.708 ± 0.040	0.734 ± 0.039	0.729 ± 0.051	0.766 ± 0.035	0.727 ± 0.035	0.768 ± 0.043
MMD-Trans	✓	✓	0.715 ± 0.041	0.720 ± 0.040	0.748 ± 0.052	0.797 ± 0.033	0.753 ± 0.036	0.785 ± 0.044
Ours	✓	✓	0.831 ± 0.033	0.819 ± 0.013	0.849 ± 0.047	0.823 ± 0.032	0.833 ± 0.045	0.894 ± 0.038

TABLE II

ZERO-SHOT TRAINING-FREE METHODS PERFORMANCE ACROSS FEW-SHOT SETTINGS IN 5 FOLD.

Dataset	Text-only ICL		Project Methods				Segmentation Ablation		Full
	Llama3-8B	Llama3-70B	Ran-Noi	Zero-Pad	Ran-Pro	OT-PCA	Ran-Sel	GT-Label	OT-CAM
CT-Survival Prediction (C-index)									
BLCA	0.452±0.033	0.482±0.029	0.475±0.025	0.543±0.024	0.581±0.024	0.613±0.022	0.631±0.026	0.618±0.025	**0.658±0.028**
BRCA	0.443±0.035	0.473±0.030	0.466±0.027	0.593±0.026	0.572±0.026	0.604±0.024	0.620±0.028	0.607±0.027	**0.649±0.031**
KIRC	0.439±0.037	0.469±0.033	0.462±0.029	0.567±0.028	0.568±0.028	0.625±0.026	0.614±0.030	0.598±0.029	**0.645±0.034**
WSI-Survival Prediction (C-index)									
BLCA	0.467±0.031	0.497±0.027	0.496±0.024	0.601±0.025	0.608±0.022	0.623±0.020	0.658±0.024	0.642±0.023	**0.689±0.026**
BRCA	0.458±0.033	0.488±0.029	0.487±0.026	0.592±0.027	0.599±0.024	0.614±0.022	0.651±0.026	0.638±0.025	**0.682±0.029**
KIRC	0.454±0.035	0.484±0.031	0.483±0.027	0.588±0.029	0.595±0.026	0.610±0.024	0.645±0.028	0.631±0.027	**0.678±0.032**
WSI-Subtype Classification (ACC)									
TCGA	0.543±0.028	0.573±0.024	0.581±0.021	0.664±0.020	0.693±0.020	0.701±0.019	0.721±0.022	0.708±0.021	**0.754±0.024**
Clinical	0.530±0.032	0.560±0.028	0.568±0.025	0.671±0.024	0.680±0.024	0.688±0.023	0.712±0.027	0.697±0.026	**0.743±0.030**

Table I presents the performance comparison across five TCGA cancer types and one private clinical cohort. Multimodal LLM-based approaches represent a promising direction by leveraging pretrained language models for cross-modal reasoning. These methods outperform basic domain adaptation baselines, indicating that LLM's semantic understanding provides valuable inductive bias for medical image interpretation. Nevertheless, their performance falls short of advanced domain adaptation methods (DANN, MMD-Trans) and our approach, suggesting that direct visual instruction tuning alone is insufficient to bridge the substantial modality gap between histopathology and radiology images. The limitation stems from their reliance on generic visual-language alignment, which lacks the structured geometric correspondence that our OT-based framework explicitly establishes.

RQ2: What is the contribution of feature-wise contrastive learning to few-shot capability?

Table II reveals the critical role of our feature-wise contrastive alignment strategy through systematic ablation. Replacing the PCA-derived semantic axes with random noise (Ran-Noi) consistently degrades performance across all tasks and datasets, highlighting the central importance of structured geometry induced by LLM token embeddings.

The progression from Ran-Noi → Zero-Pad → Ran-Pro → OT-PCA illuminates the mechanism underlying our approach. Random noise injection provides no meaningful semantic structure, forcing the LLM to rely heavily on label priors during in-context learning. Zero-padding preserves the original feature geometry but fails to establish correspondence with the LLM's semantic space. Random projection maintains some geometric relationships in expectation but lacks explicit alignment with semantically meaningful directions. Our OT-PCA method explicitly maps visual features onto the principal components of LLM token embeddings, ensuring that each feature dimension corresponds to an interpretable semantic direction.

RQ3: How do projection schemes, key frame selection, and self-supervised strategies affect performance?

The comparison among different visual-to-LLM projection strategies in Table II demonstrates that projector design—especially whether it is OT-aligned and parameter-free—has a substantial impact on framework performance. Simple dimension-matching approaches such as Zero-Pad and PCA provide baseline functionality but fail to establish meaningful semantic correspondence with the LLM's representation space.

These trends reveal that untrained, activation-free projectors with appropriate alignment can outperform more complex schemes that once the visual features are aligned with the principal semantic axes of LLM token embeddings, the projector's role is primarily to resolve dimensional mismatch while preserving the established geometric correspondence.

IV. Conclusion

Our approach enables in-context learning for cancer diagnosis tasks, demonstrating competitive performance while eliminating the need for costly cross-modal fine-tuning.

References

[1] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024.

[2] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang, "Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model," *arXiv preprint arXiv:2401.16420*, 2024.

[3] Shannon Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag, "Learning to decode collaboratively with multiple language models," 2024.

[4] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al., "Mobilevlm v2: Faster and stronger baseline for vision language model," *arXiv preprint arXiv:2402.03766*, 2024.

[5] Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu, "Appagent: Multimodal agents as smartphone users," *arXiv preprint arXiv:2312.13771*, 2023.

[6] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua, "Next-gpt: Any-to-any multimodal llm," *arXiv preprint arXiv:2309.05519*, 2023.

[7] Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, et al., "Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale," *arXiv preprint arXiv:2406.19280*, 2024.

[8] Junyang Wang, Haiyang Xu, Haitao Jia, Xi Zhang, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang, "Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration," *arXiv preprint arXiv:2406.01014*, 2024.

[9] Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, Jiadai Sun, Xinyue Yang, Yu Yang, Zehan Qi, Shuntian Yao, Xueqiao Sun, Siyi Cheng, Qinkai Zheng, Hao Yu, Hanchen Zhang, Wenyi Hong, Ming Ding, Lihang Pan, Xiaotao Gu, Aohan Zeng, Zhengxiao Du, Chan Hee Song, Yu Su, Yuxiao Dong, and Jie Tang, "Visualagentbench: Towards large multimodal models as visual foundation agents," 2024.

[10] Penghao Wu and Saining Xie, "V*: Guided visual search as a core mechanism in multimodal llms," 2023.

[11] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang, "Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output," 2024.

[12] Tiancheng Zhao, Qianqian Zhang, Kyusong Lee, Peng Liu, Lu Zhang, Chunxin Fang, Jiajia Liao, Kelei Jiang, Yibo Ma, and Ruochen Xu, "Omchat: A recipe to train multimodal language models with strong long context and video understanding," 2024.

[13] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu, "Long context transfer from language to vision," 2024.

[14] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[15] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li, "Llava-next: A strong zero-shot video understanding model," April 2024.

[16] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han, "Longvila: Scaling long-context visual language models for long videos," 2024.

[17] Xidong Wang, Dingjie Song, Shunian Chen, Junyin Chen, Zhenyang Cai, Chen Zhang, Lichao Sun, and Benyou Wang, "Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture," 2025.

[18] A. Weers et al., "From pixels to histopathology: A graph-based framework for interpretable whole slide image analysis," *arXiv preprint arXiv:2503.11846*, 2025.

[19] E. Bullmore and O. Sporns, "Complex brain networks: Graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[20] S. Brussee et al., "Graph neural networks in histopathology: Emerging trends and future directions," *Medical Image Analysis*, 2024, Early Access.

[21] D. Ahmedt-Aristizabal et al., "A survey on graph-based deep learning for computational histopathology," *Pattern Recognition*, vol. 129, pp. 108740, 2021.

[22] Luis Caicedo Torres, Luiz Manella Pereira, and M. Hadi Amini, "A survey on optimal transport for machine learning: Theory and applications," 2021.

[23] M. Y. Lu et al., "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[24] Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty, "Unbalanced minibatch optimal transport; applications to domain adaptation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3186–3197.

[25] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[26] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo, "Swin transformer v2: Scaling up capacity and resolution," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[27] The Cancer Genome Atlas Research Network, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[28] 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," 2023.

[29] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang, "Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture," 2024.

[30] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhu Chen, "Mantis: Interleaved multi-image instruction tuning," *Transactions on Machine Learning Research*, vol. 2024, 2024.