Process Reward Model as Cross-Task Learner. An Extension of Vision-Language Model to Zero-Shot Medical Reasoning

Jiahua Zhang
Nanyang Technological University
Singapore

Yidong Tian
Nanyang Technological University
Singapore

Jinghao Liang
Guangzhou University
Guangzhou, China

Dianhan Lin
Shantou University
Shantou, China

Yiwen Cai
Guangdong Medical University
Dongguan, China

Yuanqing Liu
Guangdong Medical University
Dongguan, China

Zishan Huang
Guangzhou Medical University
Guangzhou, China

Jingchun Ni
Guangdong Medical University
Dongguan, China

Jianxing He✉
Guangzhou Medical University
Guangzhou, China
jianxinghe@gzhmu.edu.cn

Abstract

Leveraging vision large language models (VLMs) for long frame medical imaging is hindered by context capacity constraints and the prohibitive cost of per task finetuning. In this paper, we propose MedTRACE that decouples lightweight process reward model training from a fully zero-shot inference pipeline, enabling a frozen VLM to generalize to unseen medical imaging tasks without any gradient updates at deployment. Evaluated on cross task generalization spanning volumetric CT and histopathology WSI benchmarks, MedTRACE achieves competitive performance compared to modality specific finetuning approaches and MIL baselines.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Process Reward Model, Vision-language model, Medical Imaging

ACM Reference Format:
Jiahua Zhang, Yidong Tian, Jinghao Liang, Dianhan Lin, Yiwen Cai, Yuanqing Liu, Zishan Huang, Jingchun Ni, and Jianxing He. 2026. Process Reward Model as Cross-Task Learner. An Extension of Vision-Language Model to Zero-Shot Medical Reasoning. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835873

1 Introduction

Long frame medical imaging sequences such as multi slice CT volumes and gigapixel whole-slide histopathology images (WSIs) comprising hundreds of axial slices and patches constitute the primary evidence for clinical diagnosis, staging, and prognosis estimation [17, 39]. Unlike single image snapshots, these sequences encode rich spatial and morphological information about disease progression that is essential for accurate clinical reasoning. Vision large language models (VLMs) have recently demonstrated remarkable capabilities in medical image interpretation [15], yet their practical deployment on long-frame data faces a fundamental architectural bottleneck: the number of visual tokens a VLM can process is bounded by its context capacity, making it infeasible to ingest an entire CT volume or WSI simultaneously [26]. More critically, clinical demands evolve continuously with new diagnostic tasks, imaging protocols, and disease contexts emerge far more rapidly than labeled datasets can be curated. However, existing VLM-based approaches typically require task specific finetuning with substantial annotated data for every new application scenario [7, 18].

Recently, Process Reward Models (PRMs) [22, 42] have emerged as a powerful paradigm for enhancing reasoning capabilities by providing step wise supervision over chain-of-thought (CoT) inference processes. Their success in mathematical reasoning [25, 40] and multimodal tasks [8] has motivated our extensions to medical domains, where diagnosis inherently involves multi step reasoning: identifying relevant evidence regions, extracting morphological or radiological features, integrating observations with clinical context, and synthesizing a final assessment. However, constructing step level reasoning annotations for PRM training in the medical domain demands prohibitive expert effort, and direct annotation by vision models [1, 27] often introduces inaccuracies due to the lack of fine-grained clinical knowledge.

We argue that the following obstacles hinder generalization of vision models over long frame medical imaging sequences. **Evidence Overload and Annotation Dependency**: only a small fraction of medical images are truly decisive for clinical reasoning; aggregating all slices dilutes salient diagnostic cues and encourages shortcut predictions [4]. Current medical VLMs commonly rely on expert annotated lesion regions or domain-specific priors to identify keyframes, and must be supervisedly fine-tuned on such annotations to adapt to each imaging task, precluding scalable deployment when expert labels are unavailable or new modalities emerge. **Reasoning Supervision Scarcity**: Constructing step level CoT annotations for guided training is prohibitively expensive

MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835873

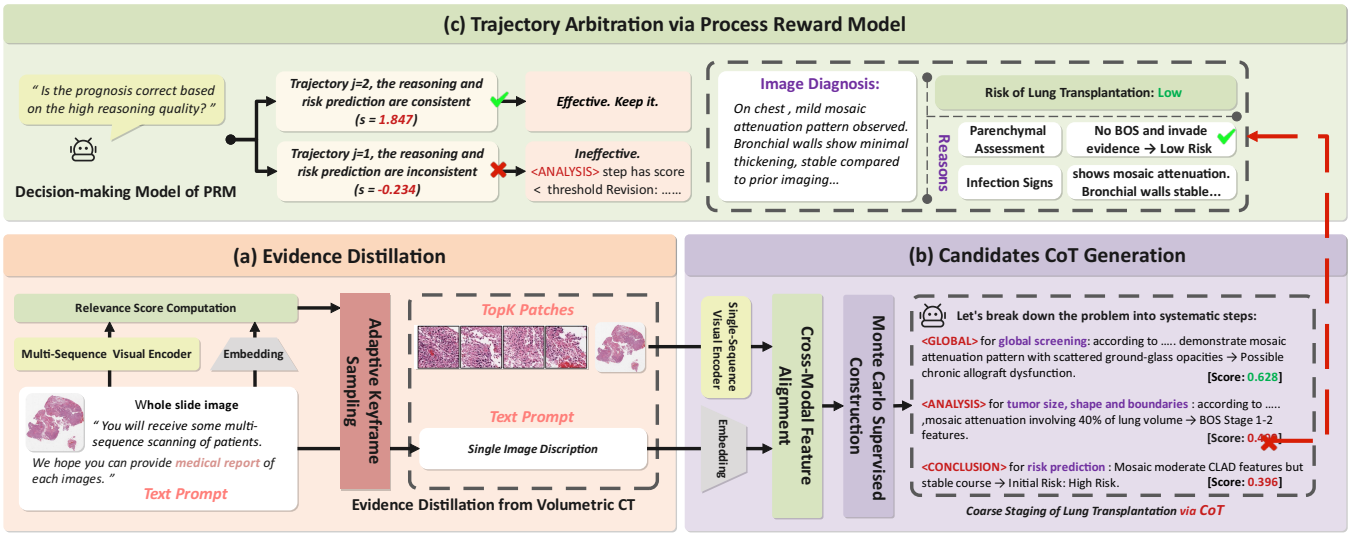

Figure 1: Overview of the proposed three stages reasoning generation(WSI): (a) Evidence Distillation aligns frame-level visual features with the text instruction space to compute relevance scores, enabling adaptive keyframe sampling of diagnostically informative slices. (b) CoT Construction: The selected keyframes and text prompt are fed into a frozen VLM, which produces multiple candidate chain-of-thought (CoT) trajectories. (c) Trajectory Arbitration via PRM: A pre-trained process reward model assesses the step-level quality of candidate trajectories and selects the best-of-N using log-odds aggregation.

in medical domains, and single source automatic signals such as Monte Carlo estimation alone cannot verify whether a reasoning step is grounded in relevant visual evidence. Unlike Supervised Fine-Tuning that relies on final answer supervision leading to overfitting and weaker out of distribution generalization [4] PRMs [22, 42] reward intermediate reasoning steps, enabling models to discover logical trajectories rather than memorize shortcuts. Their success in mathematical reasoning and multimodal tasks [8, 40] motivates extensions to domains requiring systematic analysis of complex medical data. **Modality-Specific Training Dependency**: Existing medical reasoning frameworks require per-task fine-tuning of the VLM or domain-specific reward model adaptation, making deployment on new imaging modalities computationally prohibitive when labeled data are scarce.

In this paper, we propose *MedTRACE*[1] that decouples lightweight PRM training from a training free and zero shot inference pipeline without Supervised Fine-Tuning(SFT), enabling a frozen image level VLM to generalize to unseen medical tasks and modalities. Our contributions are as follows:

*(1)We design a **fused step-level supervision signal** that combines text-vector space matching and Monte Carlo estimation to automatically annotate CoT reasoning steps for adaptive PRM training, replacing costly expert labeling while jointly capturing visual grounding quality and trajectory-level correctness.*

*(2)We propose a **PRM based zero-shot inference pipeline** that enables a frozen image-level VLM to generalize to unseen medical imaging tasks and modalities through three synergistic steps: OT-based evidence distillation, candidate CoT construction, and PRM-guided trajectory arbitration, all without any gradient updates at inference.*

*(3)We identify that **training free cross-modal evidence distillation** aligns frame features effectively with the text instruction space via optimal transport, enabling automatic annotation selection of diagnostically informative keyframes on unseen modalities without domain-specific training or biomarker annotations.*

2 Related Work

Multi Instance Learning and Vision-Language Models. Multi instance learning provides the foundational paradigm for clinical imaging such as WSI-based computational pathology, where gigapixel slides are decomposed into patch-level instances under slide-level supervision. ABMIL [14] introduced attention based pooling to MIL, enabling learnable instance weighting. CLAM [23] improved data efficiency through instance level clustering and gated attention, while TransMIL [32] modeled inter patch dependencies via self-attention to capture morphological and spatial relationships. DSMIL [19] employed pyramid fusion for multi scale feature integration, and H2MIL [13] incorporated multi resolution information through heterogeneous graph learning. More recently, vision-language frameworks have advanced beyond traditional MIL by integrating pretrained models such as CLIP [30] and large language models. TOP-MIL [28] leverages CLIP visual features together with GPT-4 linguistic priors in a two stage few-shot framework for WSI classification. ViLa-MIL [34] fuses vision-language information via prototype guided decoding and dual scale text prompts. FOCUS [12]

[1]Code and appendix available at: https://github.com/Mahiruoshi/MedTRACE

combines pathological features with language priors through progressive three-stage compression, while Concept-MIL [35] predicts pathological concepts to represent key instances through concept-activated patches. In multimodal survival prediction, MCAT [3] employs co-attention between image patches and gene embeddings, MOTCat [50] introduces optimal transport based alignment, and SURVPATH [15] encodes transcriptomic data as pathway informed tokens within a multimodal Transformer. Despite these advances, existing approaches either lack step wise interpretability or require expensive alignment procedures for reasoning annotation, motivating our CoT construction.

Process Reward Models and Reinforcement Learning for Reasoning. Process Reward Models provide finer-grained verification than Outcome Reward Models (ORMs) [5, 33] by scoring each intermediate step of a reasoning trajectory rather than only the final answer [21, 22]. A central challenge is obtaining step level supervision signals without costly human annotation. Math-Shepherd [42] addresses this through Monte Carlo estimation that produces both hard and soft labels for automatic process supervision, while OmegaPRM [24] employs Monte Carlo Tree Search for finer-grained exploration, and MiPS [46] further investigates aggregation strategies for step-level PRM signals during inference. In the multimodal domain, extending CoT reasoning to vision-language models introduces additional difficulties: hallucinated reasoning outputs are more prevalent than in text only settings [44, 53], and distribution shifts across visual-linguistic modalities exacerbate generalization failures [8]. Post-training methods such as InternVL-MPO [44] and LLaVA-CoT [49] improve multimodal reasoning through preference optimization and structured thinking datasets, while inference-time approaches including RLAIF-V [51] and AR-MCTS [8] scale reasoning via self-feedback and retrieval-augmented tree search. Orthogonal to these efforts, domain reweighting techniques have been developed to modulate the influence of heterogeneous data sources during training. DoReMi [48] assigns domain weights via distributionally robust optimization, and DOGE [10] proposes first-order bi-level optimization using gradient alignment between source and target domains. SlidePRM extends these ideas to the medical PRM setting, where multi-center quality imbalance and distribution shift are particularly acute, by jointly optimizing domain weights and PRM parameters through a bi-level framework with an aggregation function loss.

3 Preliminaries

3.1 Problem Formulation

Let $\mathcal{S}, \mathcal{T}, \mathcal{Z}$ and $\mathcal{Y}$ denote the spaces of imaging sequences, textual instructions, individual reasoning steps and final answers. A sequence $\mathbf{V} = (I_1, \ldots, I_T) \in \mathcal{S}$ consists of T frames (the slices of a CT volume or the tiles of a histopathology WSI), each $I_t \in \mathbb{R}^{W\times H\times C}$. Given an instruction $\mathbf{Q} \in \mathcal{T}$, the goal is a correct answer $y \in \mathcal{Y}$. A pre-trained vision encoder $f_\eta^{(\mathrm{V})}$ maps each frame to a feature

$$\mathbf{F}_t = f_\eta^{(\mathrm{V})}(I_t) \in \mathbb{R}^{d_v}, \quad t = 1, \ldots, T, \tag{1}$$

and we write $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_T)$. Since visual features and language embeddings live in different spaces ($d_v \neq d_h$), every cross-modal comparison is mediated by a *lifting operator*

$$\Pi : \mathbb{R}^{d_v} \to \mathbb{R}^{d_h}, \tag{2}$$

instantiated in two ways: Π_{vlm}, the VLM's frozen visual projector (Sec. 4.1.3), and Π_{rnd}, an untrained random linear map (Sec. 4.2.1). We write $f_{\mathrm{LM}} : \mathcal{T} \cup \mathcal{Z}^* \to \mathbb{R}^{d_h}$ for the VLM's frozen language embedder, defined on any text span by mean-pooling its token embeddings. The frozen image-level VLM G accepts only a limited number of frames, so feeding all T slices is both prohibitive and noisy. The pipeline therefore uses a keyframe selector

$$\mathrm{KS}_M(\mathbf{Q}, \mathbf{V}) = \mathcal{I} \subseteq \{1, \ldots, T\}, \quad |\mathcal{I}| = M, \tag{3}$$

with M bounded by the VLM's context capacity. Given

$$x = \big(\mathbf{Q}, \{\mathbf{F}_t\}_{t\in\mathcal{I}}\big) \in \mathcal{X} := \mathcal{T} \times \bigcup_{m\geq 1} (\mathbb{R}^{d_v})^m \tag{4}$$

for the multimodal input (the union over m so that the space is defined for any retained budget, though training and inference use the same M by default), the VLM samples a chain-of-thought (CoT) trajectory $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n) \in \mathcal{Z}^n$,

$$\hat{y} \sim G(\cdot \mid x). \tag{5}$$

Throughout, $\hat{y}_i \in \mathcal{Z}$ is the i-th *step* and $\hat{y}_{\leq i}$ the *prefix* of the first i steps; $\mathcal{Z}^* = \bigcup_{m\geq 1} \mathcal{Z}^m$ is the set of complete and partial trajectories and $\mathrm{ans} : \mathcal{Z}^* \to \mathcal{Y}$ extracts a final answer.

3.2 Process Reward Model

A process reward model (PRM) $\mathcal{V}_\phi : \mathcal{X} \times \mathcal{Z}^* \to (0, 1)$ scores each partial reasoning state [2, 43]; unrolling x recovers the signature $\mathcal{T} \times (\mathbb{R}^{d_v})^{|\mathcal{I}|} \times \mathcal{Z}^*$ of [2]. For step i it reads the prefix $\hat{y}_{\leq i}$ and predicts a correctness score from two designated tokens (+/−) at the step delimiter:

$$\hat{p}_i = \mathcal{V}_\phi\big(x, \hat{y}_{\leq i}\big) = \frac{\exp\big(z_i^{(+)}\big)}{\exp\big(z_i^{(+)}\big) + \exp\big(z_i^{(-)}\big)}, \tag{6}$$

where $z_i^{(\pm)}$ are the logits of the two tokens. The codomain is the *open* interval, which the softmax guarantees and Eq. (7) requires. Unlike outcome reward models, the PRM localizes *where* a chain deviates. Step scores are aggregated by log-odds summation [2, 46]:

$$\mathcal{A}(\hat{\mathbf{p}}) = \sum\nolimits_{i=1}^{n} \log \frac{\hat{p}_i}{1-\hat{p}_i}, \quad \hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_n), \tag{7}$$

clipping $\hat{p}_i$ to $[\kappa, 1-\kappa]$, $\kappa = 10^{-6}$, for stability. Among N candidates the highest $\mathcal{A}$ wins (*best-of-N*).

3.3 Optimal Transport

Optimal transport (OT) aligns distributions across representation spaces without paired supervision [6, 38]. Let $\mathcal{D}_{\mathrm{src}} = \{\mathbf{u}_i\}_{i=1}^{n_{\mathrm{src}}}$ and $\mathcal{D}_{\mathrm{tar}} = \{\mathbf{v}_j\}_{j=1}^{n_{\mathrm{tar}}} \subset \mathbb{R}^d$ be sample sets with empirical distributions μ, ν. The Kantorovich problem seeks

$$\gamma^* = \arg\min_{\gamma\in\Gamma(\mu,\nu)} \int c(\mathbf{u}, \mathbf{v})\, \mathrm{d}\gamma(\mathbf{u}, \mathbf{v}), \quad c(\mathbf{u}, \mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|_2^2, \tag{8}$$

over couplings $\Gamma(\mu, \nu)$. The squared Euclidean cost puts the solution in closed form: for Gaussian marginals the Monge map of Eq. (8) is affine, and for diagonal covariances it is exactly per-dimension

moment matching. We therefore match the first two moments of every dimension,

$$\text{scale}_j = \frac{\sigma_{\text{tar},j}}{\sigma_{\text{src},j} + \varepsilon}, \quad \text{shift}_j = \bar{v}_j - \text{scale}_j\, \bar{u}_j, \tag{9}$$

and apply the map row-wise,

$$\mathsf{T}_{\text{OT}}(\mathbf{u}) = \mathbf{scale} \odot \mathbf{u} + \mathbf{shift}, \tag{10}$$

where $\bar{u}_j, \bar{v}_j, \sigma_{\text{src},j}, \sigma_{\text{tar},j}$ are the empirical means and standard deviations of dimension j, $\odot$ is the Hadamard product and $\varepsilon > 0$ guards degenerate coordinates. The transported set then has mean exactly $\bar{v}_j$ and standard deviation $\sigma_{\text{tar},j}$ up to the ε guard.[2] Eq. (10) presumes a shared ambient dimension; otherwise a lifting operator is composed in front of it.

3.4 Monte Carlo Step-Level Scoring

To supervise steps without human annotation we use Monte Carlo (MC) estimation [2, 43]. Given a prefix $\hat{y}_{\le i}$, we sample R completions $\xi^{(r)} \sim G(\cdot \mid x, \hat{y}_{\le i})$ and compare their answers with the label y:

$$p_i^{(\text{MC})} = \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}\Big[\text{ans}\big(\hat{y}_{\le i} \oplus \xi^{(r)}\big) = y\Big] \in [0, 1], \tag{11}$$

with $\oplus$ denoting step concatenation. This estimates $\Pr[\text{ans}(\hat{y}) = y \mid \hat{y}_{\le i}]$: high values mark a productive path, low values a prefix unlikely to recover.

4 Method

SFT on outcome-level supervision rewards the final answer alone and therefore encourages pattern matching rather than principled diagnostic inference. We instead transfer a pre-trained PRM into a training free zero shot pipeline, letting a frozen image level VLM generalize to unseen imaging tasks and modalities without any gradient update at deployment. Three modules realize this: **Evidence Distillation** extracts diagnostically salient frames, by adaptive keyframe sampling at training time [36] and by OT based alignment at inference; **Structured Inference** builds multi-step rationales anchored to that evidence; and **Trajectory Arbitration** scores and selects chains with a PRM trained on fused step level supervision, yielding traceable rationales with per-step quality estimates.

4.1 PRM Training

4.1.1 Adaptive Keyframe Sampling. At training time, labels and domain-specific biomarker descriptions are available. We adopt Adaptive Relevance-based Keyframe Sampling (ADA) [36] to isolate critical frames. ADA scores frames by cross-modal similarity, which requires the two modalities to be *directly comparable*; we therefore use a jointly pre-trained dual encoder pair $g^{(\text{V})}, g^{(\text{T})}$ mapping images and text into a *shared* space $\mathbb{R}^{d_c}$ (a biomedical CLIP-style model), not the VLM encoders $f_\eta^{(\text{V})}, f_{\text{LM}}$, whose codomains $\mathbb{R}^{d_v}, \mathbb{R}^{d_h}$ are neither of equal dimension nor mutually calibrated:

$$s_t := s(\mathbf{Q}, I_t) = \frac{g^{(\text{V})}(I_t)^\top g^{(\text{T})}(\mathbf{Q})}{\|g^{(\text{V})}(I_t)\|\, \|g^{(\text{T})}(\mathbf{Q})\|} \in [-1, 1]. \tag{12}$$

[2]The factor scale_j inside shift_j is required for the transported first moment to equal $\bar{v}_j$ *after* scaling; the variant $\text{shift}_j = \bar{v}_j - \bar{u}_j$ of [52] matches the target mean only when $\text{scale}_j \approx 1$.

ADA then alternates two modes according to how concentrated the relevance profile is. Writing $\pi_t \propto e^{s_t/\tau_r}$ for the normalized profile over the T frames and $H(\pi) = -\sum_t \pi_t \log \pi_t$ for its entropy, *focal extraction* takes the top-scoring frames when the profile is peaked, $H(\pi)/\log T < \rho$ with $\rho \in (0, 1)$ a fixed concentration budget, and *axial bisection* recursively partitions the sequence when the profile is diffuse, so that spatially distributed cues are still covered.

4.1.2 Temperature-Controlled CoT Generation. For each input $x = (\mathbf{Q}, \{\mathbf{F}_t\}_{t \in \mathcal{I}})$ we query the frozen VLM G at an elevated temperature τ_s, which rescales the decoding logits $\zeta_{\ell,v}$ of vocabulary item $v \in \mathcal{W}$ at position ℓ before the softmax,

$$P\big(w_\ell = v \mid w_{<\ell}, x\big) = \frac{\exp\big(\zeta_{\ell,v}/\tau_s\big)}{\sum_{v' \in \mathcal{W}} \exp\big(\zeta_{\ell,v'}/\tau_s\big)}, \tag{13}$$

yielding N candidates $\{\hat{y}^{(c)}\}_{c=1}^N$ with $\hat{y}^{(c)} = (\hat{y}_1^{(c)}, \ldots, \hat{y}_{n_c}^{(c)})$. Larger τ_s flattens the distribution and encourages alternative branches, so the pool spans a range of quality.

4.1.3 Step-Level Fused Scoring. Every step needs a quality score to serve as the supervision *target*. We fuse two complementary signals. Targets are written $\tilde{p}_i$ to distinguish them from the PRM's *prediction* $\hat{p}_i$ of Eq. (6).

Text–vector space matching. Embedding step $\hat{y}_i$ as $\mathbf{h}_i = f_{\text{LM}}(\hat{y}_i)$ and lifting the selected frames by the VLM's frozen projector Π_{vlm}, the structural grounding score is

$$s_i^{(\text{match})} = \max_{t \in \mathcal{I}} \; \alpha_t \cdot \text{sim}^+\big(\mathbf{h}_i, \Pi_{\text{vlm}}(\mathbf{F}_t)\big) \in [0, 1], \tag{14}$$

where $\text{sim}^+(\mathbf{a}, \mathbf{b}) = \frac{1}{2}\big(1 + \cos(\mathbf{a}, \mathbf{b})\big) \in [0, 1]$ is a rectified cosine, which puts the score on the same scale as the MC signal so the two can be combined convexly, and $\alpha_t \in [\alpha_{\min}, \alpha_{\max}] \subseteq (0, 1]$ is a per-frame evidence gate. The score is high when a step engages relevant evidence and low when it is ungrounded or hallucinated.

Evidence gates. By default the gate is the ADA relevance, normalized relative to a uniform allocation so that the gates have mean 1 before clipping, independently of $|\mathcal{I}|$:

$$\alpha_t = \text{clip}\big(|\mathcal{I}| \cdot \pi_t, \alpha_{\min}, \alpha_{\max}\big), \;\; \pi_t = \frac{e^{s_t/\tau_r}}{\sum_{t' \in \mathcal{I}} e^{s_{t'}/\tau_r}}, \tag{15}$$

so an average-relevance frame receives $\alpha_t = \alpha_{\max}$ ($\alpha_{\max} = 1$, $\alpha_{\min} = 0.1$) and below-average frames are attenuated. Two learnable instantiations, obtained by applying the bi-level reweighting scheme of [2] at the *instance* rather than the domain level, are also evaluated. For the j-th training instance, **ITab** (Instance Table) is a sample-specific lookup and **INet** (Instance Net) a lightweight predictor $f_\psi : \mathbb{R}^{d_v} \to \mathbb{R}$:

$$\text{ITab:} \;\; \alpha_t = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})\, \text{sigmoid}\big(\alpha_{j,t}^{\text{tab}}\big), \tag{16}$$

$$\text{INet:} \;\; \alpha_t = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})\, \text{sigmoid}\big(f_\psi(\mathbf{F}_t)\big). \tag{17}$$

ITab maximizes per-instance flexibility at a parameter cost that grows with the dataset; INet keeps its size constant. All three map into $[\alpha_{\min}, \alpha_{\max}] \subseteq (0, 1]$, so Eq. (14) stays in $[0, 1]$ under any choice. Since α_t enters a supervision *target*, the learnable variants cannot be fitted by the inner loss of Eq. (21): $\{\alpha_{j,t}^{\text{tab}}\}$ or ψ is optimized in an outer loop on a held-out meta set, with the targets held constant inside.

Monte Carlo estimation. Given the prefix $\hat{y}_{\le i}$ and the label y, we draw R completions from the frozen VLM and compute $p_i^{(\mathrm{MC})}$ of Eq. (11), the probability that the prefix leads to the correct answer. The target is a convex combination,

$$\tilde{p}_i = \beta \cdot s_i^{(\mathrm{match})} + (1-\beta)\cdot p_i^{(\mathrm{MC})} \in [0,1], \quad \beta \in [0,1], \tag{18}$$

well defined precisely because both terms were placed on the common scale $[0,1]$. The grounding term rewards steps anchored to relevant evidence; the MC term preserves trajectory-level correctness. Thresholding at θ gives the binary marker

$$m_i = + \text{ if } \tilde{p}_i > \theta, \qquad m_i = - \text{ otherwise.} \tag{19}$$

Because cross-modal cosines concentrate near zero, sim^+ concentrates near 0.5 and the distribution of $\tilde{p}_i$ is narrow; we therefore calibrate θ on a held-out split rather than fixing it at 0.5.

4.1.4 Supervised PRM Training. $\mathcal{V}_\phi$ is initialized from a pre-trained *vision–language* model with signature $\mathcal{X} \times \mathcal{Z}^* \to (0,1)$: the frame features $\{\mathbf{F}_t\}_{t\in\mathcal{I}}$. For a pair $(x, \hat{y})$ of n steps it predicts $\hat{p}_i$ by Eq. (6), and the loss is the step-level cross-entropy against the markers:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{n}\sum_{i=1}^{n}\Big[\mathbf{1}_{[m_i=+]}\log\hat{p}_i + \mathbf{1}_{[m_i=-]}\log\big(1-\hat{p}_i\big)\Big]. \tag{20}$$

This is the hard-label form of [43]; the soft-label alternative $\mathcal{L}_{\mathrm{MSE}} = \frac{1}{n}\sum_i(\hat{p}_i - \tilde{p}_i)^2$ of [2] discards no information from $\tilde{p}_i$ and is reported as an ablation. Writing $\mathcal{D}_{\mathrm{CoT}}^{(k)}$ for the scored trajectories of source domain k, the objective is

$$\mathcal{L}_{\mathrm{train}}(\phi) = \sum_{k=1}^{K}\omega_k \sum_{(x,\hat{y},\{m_i\})\in\mathcal{D}_{\mathrm{CoT}}^{(k)}} \mathcal{L}_{\mathrm{CE}}\big(x,\hat{y},\{m_i\};\phi\big), \tag{21}$$

updated as $\phi^{(\mathrm{iter}+1)} = \phi^{(\mathrm{iter})} - \lambda\nabla_\phi\mathcal{L}_{\mathrm{train}}$. We set $\omega_k \equiv 1$, the unweighted special case of the domain-reweighted objective of [2]; learning $\{\omega_k\}$ by bi-level optimization is orthogonal to our contribution. Algorithm 1 summarizes the whole procedure: lines 2–16 form an offline data-construction pass, lines 17–22 the fine-tuning loop.

Algorithm 1 PRM Training Pipeline for Imaging Tasks

Require: $\mathcal{D}_{\mathrm{raw}} = \{(\mathbf{V}^{(j)}, \mathbf{Q}^{(j)}, y^{(j)})\}_{j=1}^{J}$ from K domains; keyframe count M; candidate count N; rollout count R; fusion weight β; threshold θ; projector Π_{vlm}; rate λ; budget N_{iter}
Ensure: Trained PRM $\mathcal{V}_\phi$
1: $\mathcal{D}_{\mathrm{CoT}} \leftarrow \emptyset$
2: **for** each $(\mathbf{V}, \mathbf{Q}, y) \in \mathcal{D}_{\mathrm{raw}}$ **do**
3: $\quad\mathcal{I} \leftarrow \mathrm{KS}_M(\mathbf{Q}, \mathbf{V})$
4: $\quad\mathbf{F}_t \leftarrow f_\eta^{(\mathrm{V})}(I_t),\ t \in \mathcal{I};\ x \leftarrow (\mathbf{Q}, \{\mathbf{F}_t\}_{t\in\mathcal{I}})$
5: $\quad\alpha_t \leftarrow \mathrm{clip}(|\mathcal{I}|\pi_t, \alpha_{\min}, \alpha_{\max}),\ t \in \mathcal{I}$
6: $\quad\{\hat{y}^{(c)}\}_{c=1}^{N} \sim G(\cdot \mid x)$ at temperature τ_s
7: $\quad$**for** $c = 1, \ldots, N$ **do**
8: $\quad\quad$**for** $i = 1, \ldots, n_c$ **do**
9: $\quad\quad\quad s_i^{(\mathrm{match})} \leftarrow \max_{t\in\mathcal{I}} \alpha_t\, \mathrm{sim}^+(\mathbf{h}_i, \Pi_{\mathrm{vlm}}(\mathbf{F}_t))$
10: $\quad\quad\quad p_i^{(\mathrm{MC})} \leftarrow$ hit rate of R rollouts from $\hat{y}_{\le i}^{(c)}$
11: $\quad\quad\quad \tilde{p}_i \leftarrow \beta s_i^{(\mathrm{match})} + (1-\beta)p_i^{(\mathrm{MC})}$
12: $\quad\quad\quad m_i \leftarrow +$ if $\tilde{p}_i > \theta$ else $-$
13: $\quad\quad$**end for**
14: $\quad\quad\mathcal{D}_{\mathrm{CoT}} \leftarrow \mathcal{D}_{\mathrm{CoT}} \cup \{(x, \hat{y}^{(c)}, \{m_i\}_{i=1}^{n_c})\}$
15: $\quad$**end for**
16: **end for**
17: Initialize ϕ from a pretrained vision–language model (visual projector included in ϕ)
18: **for** iter $= 1, \ldots, N_{\mathrm{iter}}$ **do**
19: $\quad$Sample a batch $\{(x, \hat{y}, \{m_i\})\}$ from $\mathcal{D}_{\mathrm{CoT}}$
20: $\quad\hat{p}_i \leftarrow \mathcal{V}_\phi(x, \hat{y}_{\le i})$ for every step i
21: $\quad\phi \leftarrow \phi - \lambda\nabla_\phi\mathcal{L}_{\mathrm{train}}$
22: **end for**
23: **return** $\mathcal{V}_\phi$

4.2 Inference Pipeline

At inference on unseen modalities all parameters $(G, f_\eta^{(\mathrm{V})}, \mathcal{V}_\phi)$ stay frozen.

4.2.1 Step 1: OT-Based Evidence Distillation. Unseen tasks supply neither labels nor biomarker descriptions, so Eq. (12) is unavailable: it presumes a clinically grounded query *and* a jointly calibrated encoder pair, neither of which exists on a new modality. Following [52, 55], we recast keyframe selection as annotation free distribution alignment against the VLM's own language embedding space, probed by $\mathbf{q} = f_{\mathrm{LM}}(\mathbf{Q}) \in \mathbb{R}^{d_h}$. Frame features and language embeddings differ in *dimension* ($d_v \neq d_h$ leaves $\mathbf{q}^\top\mathbf{F}_t$ undefined) and in *geometry* (mismatched means and anisotropic scales dominate any similarity), so evidence is embedded by the composition

$$\tilde{\mathbf{F}}_t = \big(\mathsf{T}_{\mathrm{OT}} \circ \Pi_{\mathrm{rnd}}\big)(\mathbf{F}_t), \quad \Pi_{\mathrm{rnd}} : \mathbb{R}^{d_v} \to \mathbb{R}^{d_h}, \ \ \mathsf{T}_{\mathrm{OT}} : \mathbb{R}^{d_h} \to \mathbb{R}^{d_h}. \tag{22}$$

Π_{rnd} is a untrained linear projector and bias-free, with i.i.d. entries from $\mathcal{N}(0, 1/d_h)$, and the variance $\mathbb{E}\|\Pi_{\mathrm{rnd}}(\mathbf{x})\|^2 = \|\mathbf{x}\|^2$. Such a map preserves frame norms and, by a union bound over the $\binom{T}{2}$ frame pairs, their pairwise cosines up to $O(\sqrt{\log(T^2/\delta)/d_h})$ with probability $1-\delta$ [52]; the zero bias matters, since a shared additive term inflates both ends of the cosine and pushes every pair toward similarity. This buys the *relative* geometry among frames, hence a stable ranking, but says nothing about the absolute similarity to the native probe $\mathbf{q}$—which is what the transport stage below is for, at the price of trading exact angle preservation for distribution matching. A nonlinear activation would distort that geometry further and, for activations suppressing negative values, inflate similarity [52], driving the lifted frames toward uniformity and flattening the scores.

Let $\mathcal{D}_{\mathrm{src}} = \{\Pi_{\mathrm{rnd}}(\mathbf{F}_t^{(j)})\}_{j,t}$ be the lifted features pooled over a small unlabeled calibration subset of the target domain and $\mathcal{D}_{\mathrm{tar}} = \mathcal{D}_{\mathrm{text}}$ a text pool; T_{OT} is then the moment-matching map of Eqs. (9)–(10). The text pool admits two annotation-free injections: the token embeddings of the task instruction, and those of the stringified, PCA-reduced frame features, which presume no annotated text about the image. Diagonal statistics are necessary as well as cheap: $|\mathcal{D}_{\mathrm{src}}|, |\mathcal{D}_{\mathrm{text}}| \ll d_h$, so the full empirical covariances are rank-deficient and the general closed-form Gaussian map is undefined. Both parameter vectors are estimated once per domain and reused for every test sample.

Since the transport decides where the visual cloud lands and Π_{rnd} which text directions carry its variance, neither stage is neutral. We therefore replace $\mathsf{T}_{\mathrm{OT}} \circ \Pi_{\mathrm{rnd}}$ wholesale by: *Zero-Pad*, padding $\mathbf{F}_t$ to d_h and rescaling to the *global* mean and variance of the embedding matrix rather than a context-specific target; *Ran-Pro*, keeping Π_{rnd}

but dropping the transport; *Ran-Noi*, a vector drawn from a per-frame hash seed, retaining frame identity while destroying content; and *PCA*, serializing the reduced feature into the prompt as text. The probe $\mathbf{q}$ induces a soft frame-level label and, by top-M' selection, its hard counterpart:

$$s_t^{\mathrm{OT}} = \frac{\mathbf{q}^\top \tilde{\mathbf{F}}_t}{\|\mathbf{q}\|\,\|\tilde{\mathbf{F}}_t\|}, \qquad \mathcal{I}^* = \mathrm{Top}_{M'}\left(\{s_t^{\mathrm{OT}}\}_{t=1}^{T}\right),$$
$$\alpha_t^{\mathrm{OT}} = \mathrm{clip}\Big(\frac{M' e^{s_t^{\mathrm{OT}}/\tau_r}}{\sum_{t'\in\mathcal{I}^*} e^{s_{t'}^{\mathrm{OT}}/\tau_r}}, \alpha_{\min}, \alpha_{\max}\Big),\ t \in \mathcal{I}^*. \tag{23}$$

This applies exactly the normalization rule where softmax times the retained count, so the gates have mean 1 before clipping, with s_t^{OT} for s_t and M' for M. It is thus the annotation-free counterpart of the training-time gate α_t, on the same range and with the same meaning, and plays the same role of deciding which frames dominate the evidence, by fixing which frames are retained and in what order they reach G and $\mathcal{V}_\phi$.

4.2.2 Step 2: Candidate CoT Construction. Given $\{\mathbf{F}_t \mid t \in \mathcal{I}^*\}$, G generates N candidates at temperature τ_s exactly as in training (Sec. 4.1.2), needing no task specific adaptation. The VLM consumes the *raw* features $\mathbf{F}_t$ through its own visual projector; the aligned $\tilde{\mathbf{F}}_t$ of Eq. (22) serve only to compute the scores and gates of Eq. (23), never as VLM input.

4.2.3 Step 3: PRM-Guided Trajectory Arbitration. $\mathcal{V}_\phi$ scores every step of every candidate by Eq. (6), yielding $\hat{\mathbf{p}}^{(c)}$; scores are aggregated by Eq. (7) and the best trajectory returned:

$$c^* = \arg\max_{c\in\{1,\dots,N\}} \mathcal{A}\big(\hat{\mathbf{p}}^{(c)}\big), \qquad \hat{y}^* = \hat{y}^{(c^*)}. \tag{24}$$

Best-of-N thus favors paths that are both internally coherent and grounded in the distilled evidence. Distillation is gradient free and costs $O(Td_vd_h)$, linear in T, so deployment remains training-free.

5 Experiments

We conduct experiments to validate the proposed framework across two medical imaging modality categories: (i) pan-cancer survival prediction using public TCGA WSI+CT data, and (ii) lung CT disease classification on private data derived from a lung CT foundation model [11]. The PRM is trained on source domain tasks from public datasets; zero shot cross-task and cross-site generalization is then evaluated on both public held-out cohorts and private cohorts. We organize our evaluation around three research questions:**RQ1:** How effective is the proposed framework? **RQ2:** What mechanism enables the PRM to improve frozen VLM predictions under OOD settings? **RQ3:** How do the OT projection scheme, self-supervised pretraining strategy, structural grounding variant, and CoT builder interact to affect inference process performance?

5.1 Experimental Setup

We construct a multi-center pan cancer dataset from TCGA [37] covering five cancer types: BLCA (N=373), UCEC (N=480), BRCA (N=511), GBMLGG (N=569), and KIRC (N=247), totaling 2,731 patients. **Private: Lung CT Disease Classification.** We evaluate on two clinically relevant binary classification tasks derived from the lung CT vision foundation model of Gao et al. [11]: **CT-LNE** (lymph node enlargement, N=103) and **CT-PTX** (pneumothorax, N=172). To assess MedTRACE ability of zero-shot generalization, we train PRM separately on each data source and perform cross source evaluation: a PRM trained on public WSI survival data is tested zero-shot on private CT classification tasks, and conversely a PRM trained on private CT data is tested zero-shot on public WSI benchmarks, constituting cross-modality and cross-task transfer.We employ Qwen2-VL [41] as the frozen LLM backbone for pan-cancer evaluation and lung CT disease classification tasks as the CoT generator and InternVL3-1B [54] as the PRM backbone. All PRM training uses AdamW with learning rate 5×10^{-7} and cross-entropy loss (Eq. 20), at XingGuang A800 GPU Clusters.

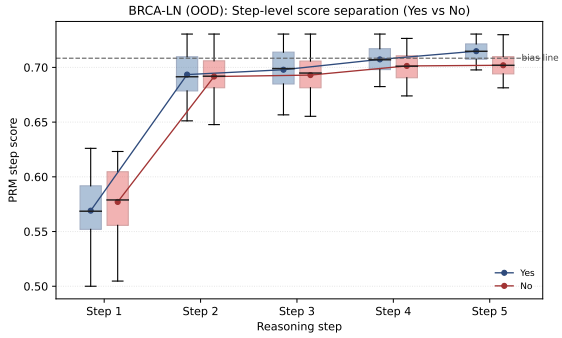

Figure 2: CT-LNE (cross-task OOD with PRM trained in TCGA-BRCA) with step-level score distribution. Steps recover only weak positive separation, and the final aggregated margin (+0.018) is compressed but remains positive.

Baselines. **(1) Zero-shot MLLMs**: domain-specific models including llava-med-7b [20], RadFM [47], and Mantis [16]; general-purpose models including LLaMA3-7B [9] and Qwen2.5-3B [29]. **(2) Test-time scaling methods**: Self-consistency [45], which selects the most consistent answer via majority voting, and ORM [31], which scores final responses to select the most promising one. **(3) PRM-based methods** (Qwen2.5-3B based): Vanilla PRM trained without domain reweighting [22].

5.2 RQ1: Cross-Modal and Cross-Site Transfer Effectiveness

In the TCGA cancer cohorts 1, baselines including domain specific trained VLMs, general purpose LLMs of varying scale and test-time scaling methods, cluster near chance-level C-index on most pathology cohorts, reflecting the inherent difficulty of zero-shot survival ranking from raw imaging frames. Self-consistency voting offers marginal gains on CT but degrades on pathology, with majority voting amplifies the frozen VLM's systematic biases rather than correcting them. ORM-based outcome scoring improves CT performance moderately but remains inconsistent across pathology cohorts, confirming that outcome-level reward signals lack the granularity needed to distinguish reasoning quality under distribution shift. The vanilla PRM baseline shows promising CT results yet fails to transfer to pathology, indicating that vision annotation alone is insufficient without cross-modal alignment. Notably, the

Table 1: Cross-task OOD evaluation (C-Index, mean ± std) on five TCGA pan-cancer cohorts. The PRM is trained on private clinical lung cancer data (WSI and CT) and evaluated zero-shot on public TCGA benchmarks under two input modalities. All methods use the same frozen VLM backbone; no TCGA data is seen during PRM training.

	Pathology (WSI)					CT				
Method	BLCA	UCEC	BRCA	GBMLGG	KIRC	BLCA	UCEC	BRCA	GBMLGG	KIRC
Zero-shot Reasoning of Supervised Fine-Tuning based Methods										
llava-med-7b [20]	0.545±0.088	0.502±0.120	0.520±0.117	0.677±0.032	0.514±0.021	0.558±0.055	0.568±0.072	0.617±0.042	0.590±0.105	0.579±0.068
RadFM [47]	0.578±0.064	0.514±0.147	0.608±0.155	0.593±0.036	0.527±0.077	0.464±0.048	0.484±0.050	0.556±0.066	0.506±0.144	0.508±0.057
Mantis [16]	0.551±0.083	0.532±0.175	0.643±0.109	0.638±0.038	0.567±0.045	0.518±0.040	0.532±0.051	0.603±0.068	0.642±0.144	0.553±0.070
LLaMA3-8B [9]	0.511±0.123	0.464±0.130	0.667±0.145	0.635±0.041	0.531±0.081	0.524±0.033	0.537±0.039	0.568±0.089	0.580±0.077	0.541±0.079
LLaMA3-70B [9]	0.642±0.126	0.474±0.062	0.482±0.187	0.657±0.050	0.542±0.069	0.534±0.041	0.624±0.036	0.614±0.053	0.671±0.205	0.601±0.053
Qwen2.5-3B [29]	0.566±0.093	0.547±0.059	0.562±0.136	0.651±0.030	0.549±0.018	0.506±0.027	0.568±0.066	0.615±0.052	0.620±0.030	0.544±0.073
Test-time Scaling Methods (Qwen2.5-3B based)										
Self-consistency [45]	0.574±0.194	0.359±0.154	0.433±0.184	0.685±0.016	0.451±0.077	0.546±0.069	0.601±0.032	0.593±0.039	0.540±0.119	0.503±0.072
ORM [31]	0.567±0.196	0.462±0.232	0.454±0.208	0.535±0.226	0.489±0.250	0.602±0.144	0.652±0.133	0.622±0.141	0.675±0.163	0.605±0.154
Vanilla PRM[22]	0.567±0.196	0.446±0.210	0.450±0.204	0.670±0.133	0.506±0.258	0.688±0.141	0.693±0.140	0.721±0.052	0.560±0.154	0.726±0.036
Proposed Pipeline (Qwen2.5-3B based)										
MedTRACE	**0.859±0.049**	**0.875±0.025**	**0.814±0.056**	**0.796±0.032**	**0.824±0.050**	**0.804±0.033**	**0.792±0.013**	**0.822±0.047**	**0.747±0.088**	**0.806±0.045**

gains are most pronounced on the pathology modality, where all prior methods struggle, which validates that the OT-based evidence distillation mechanism (Sec. 4.2.1) successfully bridges the visual–textual modality gap that defeats conventional approaches. We further evaluate cross-task generalization on clinical lung CT tasks where lesion regions occupy less than 10% of the full CT volume, which presents harder evidence localization challenge than public datasets where lesion proportions are considerably larger. Under same-task cross-center OOD on CT-LNE, the PRM-guided best-of-N selection consistently outperforms the frozen VLM's Pass@1 baseline, correcting the systematic positive-class bias and recovering specificity without sacrificing recall. Under the harder cross-task OOD protocol (PRM trained on CT-PTX, tested on CT-LNE), the framework maintains positive trajectory-ranking margins, confirming that the learned PRM reward signal captures generalizable reasoning patterns rather than task-specific shortcuts.

5.3 RQ2: PRM Mechanism Analysis Under OOD Settings

We investigate the mechanism by which PRM-guided trajectory selection improves frozen VLM predictions with K=8 candidate trajectories per iter. We identify three complementary mechanisms.
Mechanism 1: Correcting systematic positive class bias. VLMs exhibit a systematic positive class bias when applied zero-shot to medical classification, where the model overwhelmingly predicts pathology positive regardless of image content, producing near-zero specificity. This bias mirrors findings in molecular tasks where LLMs default to one category due to the prevalence of positive activity descriptions in pre-training corpora. PRM guided best-of-N selection corrects this bias by evaluating *reasoning quality* rather than surface-level prediction confidence: trajectories that reach a positive conclusion without adequate radiological evidence receive lower step-level scores, shifting the selection distribution toward better-grounded negative predictions.

Mechanism 2: Asymmetric reasoning quality and persistent inter-candidate spread under OOD. Evacuated in binary classification tasks 2, trajectories concluding with a positive finding consistently receive higher and more tightly clustered step-level scores than those concluding with a negative finding, with the gap widening progressively across reasoning steps. This asymmetry aligns with clinical intuition: when pathology is present, the VLM can anchor its reasoning to concrete radiological signs enlarged lymph nodes, irregular margins, density changes, which produce coherent multi-step rationales that the PRM recognizes as well-grounded, which learned reward signal thus captures a clinically meaningful property: affirmative diagnostic reasoning is generally more tractable than definitive exclusion. Although cross-task OOD compresses the between-class score gap, leaving separation minimal at early steps and widening only gradually, the spread across candidate trajectories does not collapse. Trajectories reasoning more carefully over anatomically relevant CT evidence receive meaningfully different step-level scores from those that do not. This persistent inter-candidate variation, combined with the asymmetric class signal, enables reliable best-of-N ranking: the PRM captures genuine reasoning quality rather than merely mirroring a learned class prior, even under substantial domain shift.

5.4 RQ3: Component Ablation and Interaction Analysis

We conduct ablation studies on private lung CT classification tasks (Table 3), where lesion sparsity (<10% of slices) makes evidence localization particularly challenging, and complement these with CoT builder analysis (Figure 4).

Cross-Modal Projection And Evidence Quality. Figure 3(a) compares strategies for the projection of inference-time annotations during inference-time evidence distillation (Sec. 4.2.1). Training-free unsupervised baselines (Ran-Noi, PCA, Zero-Pad, OT-Embed, Ran-Pro) span a wide performance range, with random noise injection already achieving a surprisingly strong baseline, suggesting

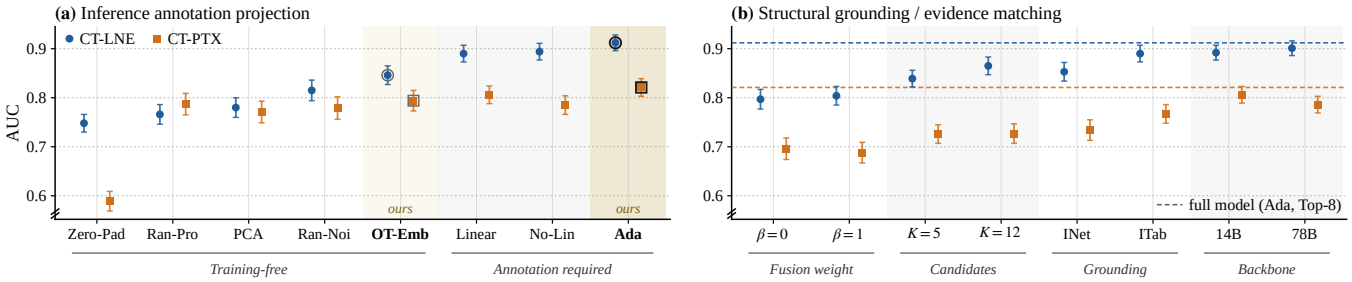

Figure 3: Component ablation (AUC) on two private lung CT binary classification tasks: CT-LNE (lymph node enlargement, N=103) and CT-PTX (pneumothorax, N=172).

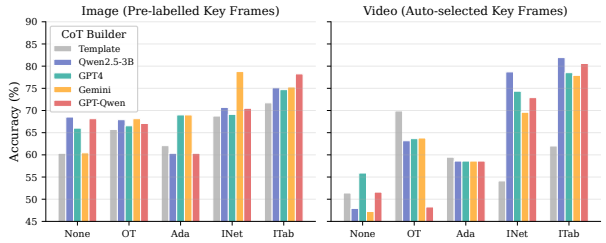

Figure 4: CoT builder comparison across structural grounding methods in CT-LNE.

that even coarse cross-modal perturbation provides sufficient diversity for the PRM to identify informative keyframes. Among these unsupervised methods, OT-Embed stands out by leveraging distributional alignment, approaching the performance of supervised alternatives without any target-domain annotations.

Supervised projection methods (Ada, Linear, No Linear) provide further gains. Ada, which leverages the keyframe sampling with domain specific biomarker queries, achieves the highest overall performance, confirming that access to clinical supervision during evidence distillation yields the strongest frame selection. Linear and non linear contrastive projectors perform comparably to each other, both narrowing the gap with Ada while requiring only generic contrastive objectives. Importantly, our training free OT-based projection achieves performance close to the trained linear projector, validating the central design premise: optimal transport alignment can substitute for gradient-based projection learning, enabling fully annotation-free deployment on unseen modalities with only a modest performance trade off relative to supervised alternatives.

Structural Grounding and Candidate Pool Interactions. Figure 3(b) reveals interaction effects among the fused step-level scoring components, candidate pool size, and backbone scale. Comparing MC-only (β=0) with Match-only (β=1) confirms that neither signal alone is sufficient: the MC estimate captures trajectory level correctness but misses evidence grounding, while structural matching rewards visual anchoring but lacks trajectory awareness. The fused signal consistently outperforms either component in isolation.

Increasing the candidate pool from Top5 to the default Top8 and further to Top12 produces measurable gains, confirming that a larger pool of diverse CoT trajectories provides the PRM with more opportunities to identify well grounded reasoning chains. Among structural grounding instantiations, ITab (Eq. 16) outperforms INet (Eq. 17), indicating that sample specific relevance weighting captures finer diagnostic distinctions than a shared parametric predictor. Scaling the PRM backbone from the default to stronger InternVL variants (14b, 78b with denoising) yields further improvements, with the larger backbone achieving the best CT-LNE performance. However, the computational cost increases substantially of larger base models,with the backbone scaling offers complementary but diminishing benefits relative to the algorithmic contributions of structural grounding.

CoT Builder: Intrinsic VLM Capability Suffices. Figure 4 compares CoT builders across structural grounding methods under both pre-labelled keyframe (Image) and auto-selected keyframe (Video) settings to evaluate the zero shot prediction ability without PRM. A key finding is that Qwen2.5-3B, operating without any external large-model guidance, produces competitive CoT trajectories across all grounding configurations by training free projection backbones, matching or exceeding template based and externally guided alternatives (GPT-4, Gemini, GPT-Qwen) in most settings. This indicates that the frozen VLM's intrinsic reasoning capability is sufficient to generate a diverse and high-quality candidate pool, and that the PRM's trajectory arbitration mechanism is the primary driver of final performance rather than the sophistication of the CoT generation source.

6 Conclusion

This paper proposes MedTRACE, which decouples lightweight process reward model training from a zero shot inference pipeline that requires no further SFT, for the analysis of long medical imaging sequences. However, curating process reward supervision is costly in its own right, since both the time complexity of annotation and the memory required to retain step wise targets grow with the length of the imaging sequence, which restricts the scale at which such data can realistically be collected. Moreover, although our inference stage requires neither manual supervision nor labels prepared in advance, it still relies on a vision model to produce slice level annotations, so the quality of the distilled evidence remains bounded by that of the underlying visual backbone. Future work may explore more expressive transport plans, reduce the dependence on externally generated slice level labels, and extend the framework to additional clinical modalities beyond CT and histopathology.

References

[1] B. F. Boyce. 2015. Whole slide imaging: uses and limitations for surgical pathology and teaching. *Biotechnic & Histochemistry* 90, 5 (2015), 321–330.

[2] Qi Cao et al. 2025. DreamPRM: Domain-Reweighted Process Reward Model for Multimodal Reasoning. arXiv:2505.20241 [cs.LG] https://arxiv.org/abs/2505.20241

[3] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. 2021. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4025.

[4] Tianzhe Chu et al. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161* (2025).

[5] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 [cs.LG] https://arxiv.org/abs/2110.14168

[6] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. 2017. Optimal Transport for Domain Adaptation. *IEEE TPAMI* 39, 9 (2017), 1853–1865.

[7] K. Ding, M. Zhou, D. N. Metaxas, and S. Zhang. 2023. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 622–631.

[8] Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2024. Progressive Multimodal Reasoning via Active Retrieval. arXiv:2412.14835 [cs.CL] https://arxiv.org/abs/2412.14835

[9] Abhimanyu Dubey et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).

[10] Simin Fan, Matteo Pagliardini, and Martin Jaggi. 2024. DoGE: Domain Reweighting with Generalization Estimation. arXiv:2310.15393 [cs.LG] https://arxiv.org/abs/2310.15393

[11] Zhiyuan Gao, Gongning Zhang, Hui Liang, et al. 2026. A lung CT vision foundation model facilitating disease diagnosis and medical imaging. *Nature Communications* 17 (2026), 35. doi:10.1038/s41467-025-66620-z

[12] Zhengrui Guo, Conghao Xiong, Jiabo Ma, Qichen Sun, Lishuang Feng, Jinzhuo Wang, and Hao Chen. 2025. Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 15590–15600.

[13] Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. 2022. Hˆ 2-MIL: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 933–941.

[14] Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*. PMLR, 2127–2136.

[15] Guillaume Jaume, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Paul Pu Liang, and Faisal Mahmood. 2024. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11579–11590.

[16] Dongfu Jiang et al. 2024. MANTIS: Interleaved Multi-Image Instruction Tuning. *arXiv preprint arXiv:2405.01483* (2024).

[17] N. C. Kurian, A. Sethi, A. R. Konduru, A. Mahajan, and S. U. Rane. 2021. A 2021 update on cancer image analytics with deep learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11, 4 (2021), e1410.

[18] M. Lambin et al. 2012. Radiomics: extracting more information from medical images using advanced feature analysis. *European Journal of Cancer* 48, 4 (2012), 441–446.

[19] Bin Li, Yin Li, and Kevin W Eliceiri. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14318–14328.

[20] Chunyuan Li et al. 2024. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *Advances in Neural Information Processing Systems* 36 (2024).

[21] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making Large Language Models Better Reasoners with Step-Aware Verifier. arXiv:2206.02336 [cs.CL] https://arxiv.org/abs/2206.02336

[22] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's Verify Step by Step. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=v8L0pN6EOi

[23] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* 5, 6 (2021), 555–570.

[24] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. 2024. Improve Mathematical Reasoning in Language Models by Automated Process Supervision. arXiv:2406.06592 [cs.CL] https://arxiv.org/abs/2406.06592

[25] Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023. Let's reward step by step: Step-Level reward model as the Navigators for Reasoning. arXiv:2310.10080 [cs.CL] https://arxiv.org/abs/2310.10080

[26] Jiazhen Pan et al. 2025. MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning. arXiv:2502.19634 [cs.CV] https://arxiv.org/abs/2502.19634

[27] L. Pantanowitz et al. 2020. Whole slide imaging in pathology: current perspectives and future directions. *Journal of Digital Imaging* (2020). https://link.springer.com/article/10.1007/s10278-020-00351-z

[28] Linhao Qu, Kexue Fu, Manning Wang, Zhijian Song, et al. 2023. The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. *Advances in Neural Information Processing Systems* 36 (2023), 67551–67564.

[29] Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. *arXiv preprint arXiv:2412.10862* (2024).

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

[31] Zhihong Shao et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

[32] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems* 34 (2021), 2136–2147.

[33] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] https://arxiv.org/abs/2402.03300

[34] Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. 2024. ViLa-MIL: Dual-scale Vision-Language Multiple Instance Learning for Whole Slide Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11248–11258.

[35] Susu Sun, Leslie Tessier, Frédérique Meeuwsen, Clément Grisi, Dominique van Midden, Geert Litjens, and Christian F Baumgartner. 2025. Label-free Concept Based Multiple Instance Learning for Gigapixel Histopathology. *arXiv preprint arXiv:2501.02922* (2025).

[36] Xi Tang et al. 2025. Adaptive Keyframe Sampling for Long Video Understanding. arXiv:2502.21271 [cs.CV] https://arxiv.org/abs/2502.21271

[37] The Cancer Genome Atlas Research Network. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 45, 10 (2013), 1113–1120.

[38] Luis Caicedo Torres, Luiz Manella Pereira, and M. Hadi Amini. 2021. A Survey on Optimal Transport for Machine Learning: Theory and Applications. arXiv:2106.01963 [cs.LG] https://arxiv.org/abs/2106.01963

[39] G. Verghese, M. Li, F. Liu, A. Lohan, N. C. Kurian, S. Meena, P. Gazinska, A. Shah, A. Oozeer, T. Chan, M. Opdam, S. Linn, C. Gillett, et al. 2023. Multiscale deep learning framework captures systemic immune features in lymph nodes predictive of triple negative breast cancer outcome in large-scale studies. *The Journal of Pathology* 260, 4 (2023), 376–389.

[40] Chaojie Wang, Yanchen Deng, Zhiyi Lyu, Liang Zeng, Jujie He, Shuicheng Yan, and Bo An. 2024. Q*: Improving Multi-step Reasoning for LLMs with Deliberative Planning. arXiv:2406.14283 [cs.AI] https://arxiv.org/abs/2406.14283

[41] Peng Wang et al. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191* (2024).

[42] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 9426–9439.

[43] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 9426–9439. doi:10.18653/v1/2024.acl-long.510

[44] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024. Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization. *arXiv preprint arXiv:2411.10442* (2024).

[45] Xuezhi Wang et al. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv:2203.11171* (2023).

[46] Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. 2024. Multi-step Problem Solving Through a Verifier: An Empirical

Analysis on Model-induced Process Supervision. arXiv:2402.02658 [cs.AI] https://arxiv.org/abs/2402.02658

[47] Chaoyi Wu et al. 2023. Towards Generalist Foundation Model for Radiology. *arXiv preprint arXiv:2308.02463* (2023).

[48] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023. DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=lXuByUeHhd

[49] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. arXiv:2411.10440 [cs.CV] https://arxiv.org/abs/2411.10440

[50] Yingxue Xu and Hao Chen. 2023. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 21241–21251.

[51] Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. RLAIF-V: Open-Source AI Feedback Leads to Super GPT-4V Trustworthiness. *arXiv preprint arXiv:2405.17220* (2024).

[52] Tianle Zhang, Wanlong Fang, Jonathan Woo, Paridhi Latawa, Deepak A. Subramanian, and Alvin Chan. 2025. Can LLMs Reason Over Non-Text Modalities in a Training-Free Manner? A Case Study with In-Context Representation Learning. arXiv:2509.17552 [cs.CL] https://arxiv.org/abs/2509.17552

[53] Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. 2024. Thinking Before Looking: Improving Multimodal LLM Reasoning via Mitigating Visual Hallucination. arXiv:2411.12591 [cs.CV] https://arxiv.org/abs/2411.12591

[54] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479* (2025).

[55] Zhenfeng Zhuang, Fangyu Zhou, and Liansheng Wang. 2025. Libra-MIL: Multimodal Prototypes Stereoscopic Infused with Task-specific Language Priors for Few-shot Whole Slide Image Classification.