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

Zhang Jiahua
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
Zhanjiang, China

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

Abstract

Leveraging vision large language models (VLMs) for long frame medical imaging analysis 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 while requiring no target domain adaptation.

CCS Concepts

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

Keywords

Process Reward Model, Vision-language model, Medical Imaging

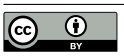
ACM Reference Format:
Zhang Jiahua, Yidong Tian, Jinghao Liang, Dianhan Lin, Yiwen Cai, Yuanqing Liu, Zishan Huang, Jingchun Ni, and He Jianxing. 2026. Process Reward Model as Cross-Task Learner. An Extension of Vision-Language Model to Zero-Shot Medical Reasoning. In *Proceedings of the 34th ACM International 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].

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



MM '26, Rio de Janeiro, Brazil

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

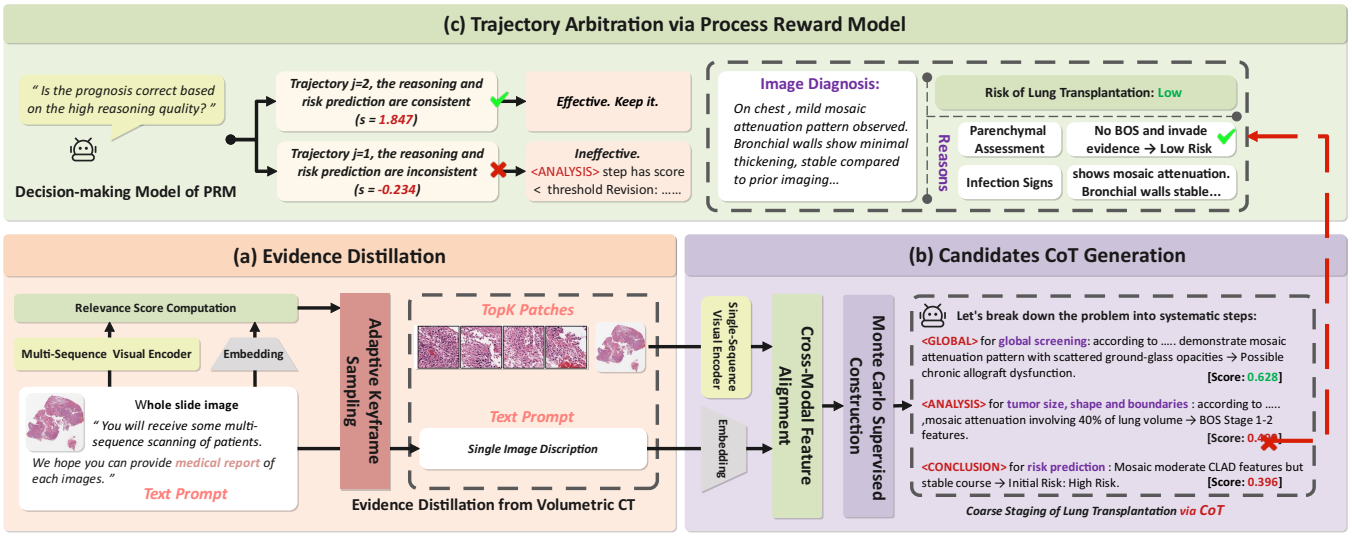


Figure 1: Overview of the proposed three stages reasoning pipeline(WSI): (a) Evidence Distillation aligns frame-level visual features with the text instruction space to compute relevance scores, enabling adaptive keyframe sampling of diagnostically informative slices. (b) CoT Construction: The selected 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.

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 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 zero-shot inference pipeline, 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 annotation-free 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

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

information via prototype-guided decoding and dual-scale text prompts. FOCUS [12] 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, 52], 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

Our framework addresses the task of enabling an image-level vision large language model (VLM) to perform zero-shot cross-task generalization across diverse long-frame medical imaging modalities. In this work, we focus on volumetric CT and histopathology whole-slide images through a training-free inference pipeline equipped with a lightweight pre-trained process reward model (PRM). During PRM training on source-domain tasks, adaptive keyframe sampling [36] isolates diagnostically critical frames from long imaging sequences, a frozen VLM generates diverse CoT trajectories from the selected evidence, and each reasoning step is assigned a supervision score via a fused signal combining text-vector space matching with Monte Carlo estimation, entirely avoiding expert annotation. At inference on unseen modalities, the pipeline is training-free: optimal transport aligns frame features with the text instruction space for annotation-free keyframe selection, the frozen VLM constructs candidate reasoning trajectories, and the pre-trained PRM selects the best trajectory via step-wise scoring—all without gradient updates. This section formalizes the problem setting, core notation, and background components used throughout the framework.

3.1 Problem Formulation

Let $\mathbf{V} = \{I_1, I_2, \ldots, I_T\}$ denote a medical imaging sequence of T frames (*e.g.*, multi-slice CT volume or tiled patches from a histopathology WSI), where each frame $I_t \in \mathbb{R}^{W\times H\times C}$. Given a text instruction $\mathbf{Q} \in \mathcal{T}$ describing a diagnostic task, the goal is to produce a correct textual answer $y \in \mathcal{Y}$.

We employ a frozen, image-level VLM G that accepts a limited number of frames as visual context. Since T can be large (*e.g.*, hundreds of slices), feeding all frames is computationally prohibitive and introduces noise from uninformative slices. The pipeline therefore requires a keyframe selection function $\mathrm{KS}_M(\mathbf{Q}, \mathbf{F})$ that selects an index set $\mathcal{I} \subseteq \{1, \ldots, T\}$ with $|\mathcal{I}| = M$, where M is constrained by the VLM's context capacity. Each frame is encoded by a pre-trained vision encoder $f_\eta^{(\mathrm{V})}$ to obtain frame-level features $\mathbf{F}_t = f_\eta^{(\mathrm{V})}(I_t) \in \mathbb{R}^{d_v}$. The VLM then generates a chain-of-thought (CoT) reasoning trajectory $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ conditioned on the selected keyframes and the text instruction $\hat{y} \sim G(\cdot \mid \{\mathbf{F}_t \mid t \in \mathcal{I}\}, \mathbf{Q})$.

3.2 Process Reward Model

A process reward model (PRM) $\mathcal{V}_\phi : \mathcal{T} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$, parameterized by ϕ, assigns a scalar reward to each partial reasoning state [2, 43]. For step i of a CoT trajectory $\hat{y}$, the PRM predicts a correctness score based on two designated tokens (+ for correct, − for incorrect):

$$p_i = \frac{\exp(z_i^{(+)})}{\exp(z_i^{(+)}) + \exp(z_i^{(-)})}, \tag{1}$$

where $z_i^{(+)}$ and $z_i^{(-)}$ are the logits for the positive and negative tokens at the scoring position of step i. Unlike outcome reward models that evaluate only final answers, the PRM provides fine-grained, step-level supervision that identifies where reasoning deviates.

Step-level scores are aggregated into a trajectory-level score via log-odds summation [46]:

$$\mathcal{A}(\mathbf{p}) = \sum_{i=1}^{n} \log \frac{p_i}{1 - p_i}. \tag{2}$$

Given N candidate trajectories, the one with the highest $\mathcal{A}(\mathbf{p})$ is selected as the final prediction (*best-of-N* selection).

3.3 Optimal Transport

Optimal transport (OT) provides a mathematical framework for aligning distributions across different representation spaces without requiring paired supervision [6, 38]. Given a source distribution $\mathcal{D}_s \in \mathbb{R}^{N\times d}$ and a target distribution $\mathcal{D}_t \in \mathbb{R}^{N'\times d}$, the OT objective seeks the transport plan γ^* that minimizes the total transportation

cost:

$$\gamma^* = \arg\min_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{D}_s \times \mathcal{D}_t} c(u,v)\, d\gamma(u,v), \tag{3}$$

where $\Pi(\mu,\nu)$ denotes the set of joint distributions with marginals μ and ν over $\mathcal{D}_s$ and $\mathcal{D}_t$, and $c(u,v)$ is the cost of transporting u to v. For practical implementation, we adopt an affine alignment that matches the mean and variance of each dimension of $\mathcal{D}_s$ to those of $\mathcal{D}_t$:

$$\text{shift}_j = \bar{v}_j - \bar{u}_j, \quad \text{scale}_j = \frac{\sigma_{t,j}}{\sigma_{s,j}}, \tag{4}$$

$$\text{OT}(\mathcal{D}_s, \mathcal{D}_t) = \text{scale} \cdot \mathcal{D}_s + \text{shift}, \tag{5}$$

where $\bar{u}_j, \bar{v}_j$ and $\sigma_{s,j}, \sigma_{t,j}$ denote the means and standard deviations of the j-th dimension. In our framework, OT alignment is used for cross-modal keyframe selection during zero-shot inference (Sec. 4.2.1).

3.4 Monte Carlo Estimation for Step-Level Scoring

To obtain step-level supervision for PRM training without costly human annotation, we adopt Monte Carlo (MC) estimation [2, 43]. Given a partial reasoning state $\hat{y}_i$ (the first i steps of a trajectory), the MC score estimates the probability that this partial trajectory eventually leads to the correct final answer $p_i^{(\text{MC})} = \frac{\text{num(correct completions from } \hat{y}_i)}{\text{num(total completions from } \hat{y}_i)}$, where multiple completions of the remaining reasoning steps are sampled from the base model and each completion's final answer is compared against the ground-truth label y. This provides a trajectory-aware quality measure for each intermediate step: high scores indicate that the reasoning is on a productive path, while low scores signal that the partial trajectory is unlikely to recover.

4 Method

Conventional medical VLMs trained through supervised fine-tuning (SFT) rely on outcome-level supervision, which encourages pattern matching rather than principled diagnostic inference. Our framework instead combines lightweight process-supervised PRM training with a training-free zero-shot inference pipeline, enabling a frozen image-level VLM to generalize to unseen medical imaging tasks and modalities without any gradient updates at deployment.

As depicted in Fig. 1, the framework operates through three functional modules: (1) **Evidence Distillation** extracts diagnostically salient frames via adaptive keyframe sampling during training [36] and optimal transport-based cross-modal alignment during inference; (2) **Structured Inference** constructs multi-step diagnostic rationales anchored to the extracted evidence through a frozen VLM with temperature-controlled diverse sampling; (3) **Trajectory Arbitration** employs a process reward model trained with fused step-level supervision to score and select reasoning chains, yielding not only diagnostic predictions but also traceable reasoning sequences with per-step quality estimates (Sec. 4.1, 4.2).

4.1 PRM Training

4.1.1 Adaptive Keyframe Sampling. During training, ground-truth labels and domain-specific biomarker descriptions are available. We adopt Adaptive Relevance-based Keyframe Sampling (ADA) [36] to isolate diagnostically critical frames from each imaging sequence $\mathbf{V} = \{I_1, \ldots, I_T\}$. For a diagnostic query $\mathbf{Q}$, frame-level relevance is quantified through cross-modal similarity:

$$s(\mathbf{Q}, I_t) = \frac{f_\eta^{(\text{V})}(I_t)^\top f_{\text{LM}}(\mathbf{Q})}{\|f_\eta^{(\text{V})}(I_t)\| \cdot \|f_{\text{LM}}(\mathbf{Q})\|}, \tag{6}$$

where $f_\eta^{(\text{V})}$ and f_{LM} are the vision and text encoders, respectively. The ADA procedure alternates between two modes depending on the score distribution: (1) **Focal extraction**: when salient peaks exist ($s_t > s_{\text{thr}}$), top-scoring frames are directly selected; (2) **Spatial bisection**: when scores are diffuse, the sequence is recursively partitioned to guarantee coverage of spatially distributed diagnostic cues. The output is a compact evidence set $\mathcal{I} \subseteq \{1, \ldots, T\}$ with $|\mathcal{I}| = M$ ($M = 10$ by default).

ADA-based selection at training time serves two purposes: (1) it concentrates the frozen VLM's limited visual context on diagnostically relevant frames, producing higher-quality CoT trajectories as positive training signals for the PRM; (2) combined with temperature-controlled diverse sampling (Sec. 4.1.2), it still yields trajectories of varying quality—those where the VLM's stochastic decoding deviates from the visual evidence provide negative signals—ensuring a balanced training distribution.

4.1.2 Temperature-Controlled Diverse CoT Generation. For each input $x = (\mathbf{Q}, \{\mathbf{F}_t \mid t \in \mathcal{I}\})$, we query the frozen VLM G at an elevated sampling temperature τ_s to produce N candidate CoT trajectories $\{\hat{y}^{(c)}\}_{c=1}^N$, where each trajectory $\hat{y}^{(c)} = (\hat{y}_1^{(c)}, \hat{y}_2^{(c)}, \ldots, \hat{y}_{n_c}^{(c)})$ is a multi-step reasoning chain terminating in a prediction. The temperature parameter scales the logit vector $\mathbf{z}$ before the softmax operation:

$$P(y_t \mid y_{<t}, x) = \frac{\exp(z_t/\tau_s)}{\sum_v \exp(z_v/\tau_s)}, \tag{7}$$

where higher τ_s yields flatter distributions, encouraging exploration of alternative reasoning branches. This stochastic sampling naturally produces trajectories of varying quality, providing a diverse candidate pool for step-level scoring.

4.1.3 Step-Level Fused Scoring. Each reasoning step of the generated CoT trajectories must be assigned a quality score to serve as the supervision signal for PRM training. We design a fused scoring mechanism that combines two complementary signals: text-vector space matching and Monte Carlo estimation.

Text-Vector Space Matching. Each reasoning step r_i in a CoT trajectory produces a textual representation $\mathbf{t}_{r_i} \in \mathbb{R}^{d_h}$ via the language model encoder. To measure whether the reasoning step is grounded in diagnostically relevant visual evidence, we compute the structural grounding score between the step representation and the selected frame embeddings:

$$s_i^{(\text{match})} = \max_{t \in \mathcal{I}} \alpha_t(x) \cdot \frac{\mathbf{t}_{r_i}^\top \mathbf{F}_t}{\|\mathbf{t}_{r_i}\| \|\mathbf{F}_t\|}, \tag{8}$$

where $\mathbf{F}_t$ is the frame-level embedding and $\alpha_t(x)$ is a per-frame relevance weight reflecting the diagnostic importance of frame I_t. The score is high when a reasoning step engages with visually relevant evidence and low when the reasoning is ungrounded or hallucinated. The per-frame relevance weight $\alpha_t(x)$ can be instantiated in two ways:

ITab (Instance Table): a sample-specific lookup with clipping,

$$\alpha_t(x) = \mathrm{clip}(\alpha_{x,t}^{\mathrm{tab}},\ \alpha_{\min},\ \alpha_{\max}), \tag{9}$$

which maximizes per-instance flexibility.
INet (Instance Net): a lightweight parametric predictor,

$$\alpha_t(x) = a + (b-a)\,\sigma\big(f_\psi(\mathbf{F}_t)\big), \tag{10}$$

which keeps parameter size independent of dataset scale.

Monte Carlo Estimation. Following the formulation in Sec. 3.4, given the partial reasoning state $\hat{y}_i$ and the ground-truth label y, we sample R completions of the remaining steps from the base VLM and compute the MC score $p_i^{(\mathrm{MC})}$. This signal estimates the probability that the partial trajectory leads to the correct answer.

Fused Supervision Signal. The final step-level supervision target is a convex combination of the structural grounding score and the MC estimate:

$$p_i = \beta \cdot s_i^{(\mathrm{match})} + (1-\beta)\cdot p_i^{(\mathrm{MC})}, \tag{11}$$

where β balances as bias. The structural grounding term encourages the PRM to reward reasoning steps anchored to relevant visual evidence, while the MC term preserves trajectory-level correctness. Based on the fused score, a binary correctness marker (+ if $p_i > 0.5$, − otherwise) is appended to each reasoning step.

4.1.4 Supervised PRM Training. The PRM $\mathcal{V}_\phi$ is initialized from a pre-trained language model and fine-tuned on the step-scored CoT data. For each training sample, the PRM predicts step-wise correctness scores p_i via Eq. 1. The training loss is the cross-entropy between predicted and annotated correctness markers:

$$\mathcal{L}_{\mathrm{CE}}(x,\hat{y},\phi) = -\sum_{i=1}^{n}\Big[\mathbf{1}_{[c_i=+]}\log p_i + \mathbf{1}_{[c_i=-]}\log(1-p_i)\Big], \tag{12}$$

where $c_i \in \{+,-\}$ is the correctness marker derived from the fused score (Eq. 11). The overall training objective aggregates over all source-domain tasks $\mathcal{D} = \{\mathcal{D}_1,\dots,\mathcal{D}_K\}$:

$$\mathcal{L}_{\mathrm{train}}(\phi) = \sum_{k=1}^{K}\sum_{(x,\hat{y})\in\mathcal{D}_k}\mathcal{L}_{\mathrm{CE}}(x,\hat{y},\phi), \tag{13}$$

with parameter updates $\phi^{(t+1)} = \phi^{(t)} - \eta\nabla_\phi\mathcal{L}_{\mathrm{train}}(\phi^{(t)})$.

4.1.5 Algorithmic View of PRM Training Pipeline. For clarity, Algorithm 1 summarizes the complete PRM training workflow for imaging tasks, including random keyframe sampling, CoT candidate generation, fused step-level scoring, and supervised PRM optimization.

Algorithm 1 PRM Training Pipeline for Imaging Tasks

Require: Imaging training set $\mathcal{D} = \{(\mathbf{V}^{(j)},\mathbf{Q}^{(j)},y^{(j)})\}_{j=1}^{J}$, keyframe count M, candidate count N, MC rollout count R, fusion weight β, PRM learning rate η
Ensure: Trained PRM $\mathcal{V}_\phi$
Initialize scored trajectory set $\mathcal{D}_{\mathrm{CoT}} \leftarrow \emptyset$
for each sample $(\mathbf{V},\mathbf{Q},y) \in \mathcal{D}$ **do**
  Select keyframe index set $\mathcal{I}$ via ADA [36] with query $\mathbf{Q}$ (Eq. 6)
  Encode selected frames: $\mathbf{F}_t = f_\eta^{(\mathrm{V})}(I_t),\ t\in\mathcal{I}$
  Generate N candidate CoT trajectories $\{\hat{y}^{(c)}\}_{c=1}^{N}$ with frozen VLM G at temperature τ_s (Eq. 7)
  for each trajectory $\hat{y}^{(c)}$ and each step i **do**
    Compute structural grounding score $s_i^{(\mathrm{match})}$ using selected frames (ITab/INet instantiation optional) (Eq. 8)
    Estimate MC score $p_i^{(\mathrm{MC})}$ from R rollouts conditioned on partial trajectory $\hat{y}_i$
    Fuse scores: $p_i \leftarrow \beta s_i^{(\mathrm{match})} + (1-\beta)p_i^{(\mathrm{MC})}$ (Eq. 11)
    Assign correctness marker $c_i \in \{+,-\}$ by thresholding p_i
  end for
  Add $(\mathbf{Q},\{\mathbf{F}_t\}_{t\in\mathcal{I}},\hat{y}^{(c)},\{c_i\})$ to $\mathcal{D}_{\mathrm{CoT}}$
end for
Initialize PRM parameters ϕ from pretrained language model
for training iteration $t = 1,\dots,T_{\mathrm{train}}$ **do**
  Sample mini-batch $(x,\hat{y},\{c_i\})$ from $\mathcal{D}_{\mathrm{CoT}}$
  Predict step scores via token logits: $p_i = \frac{\exp(z_i^{(+)})}{\exp(z_i^{(+)})+\exp(z_i^{(-)})}$ (Eq. 1)
  Compute batch loss with step-level cross-entropy (Eq. 12)
  Update parameters by gradient descent on $\mathcal{L}_{\mathrm{train}}$ (Eq. 13)
end for
return $\mathcal{V}_\phi$

4.2 Training-Free Zero-Shot Inference Pipeline

At inference time on unseen medical modalities and tasks, all model parameters (VLM G, vision encoder $f_\eta^{(\mathrm{V})}$, PRM $\mathcal{V}_\phi$) remain frozen. The entire pipeline is training-free and proceeds through three steps: OT-based evidence distillation for keyframe selection, candidate CoT construction, and PRM-guided trajectory arbitration.

4.2.1 Step 1: OT-Based Evidence Distillation. At inference on unseen modalities, neither ground-truth labels nor domain-specific biomarker descriptions are available, rendering the ADA similarity-based frame selection used during training (Sec. 4.1.1) inapplicable—the cross-modal relevance in Eq. 6 requires a clinically grounded text query $\mathbf{Q}$ with biomarker semantics, which cannot be constructed for unseen tasks without domain expertise. Inspired by the optimal matching flow in [54], we instead leverage optimal transport to perform annotation-free evidence distillation, identifying the most diagnostically informative frames by aligning frame features directly with the task instruction embedding space.

For each frame I_t in the target imaging sequence, we compute a prompt-frame relevance score. Let $\mathbf{q} = f_{\mathrm{LM}}(\mathbf{Q}) \in \mathbb{R}^{d_h}$ be the text instruction embedding and $\mathbf{F}_t \in \mathbb{R}^{d_v}$ be the frame feature. Since $\mathbf{q}$ and $\mathbf{F}_t$ reside in different representation spaces, we first apply OT affine alignment (Eq. 5) to map the frame features into the text embedding space:

$$\tilde{\mathbf{F}}_t = \mathrm{OT}(\mathbf{F}_t,\ \mathcal{D}_{\mathrm{text}}), \tag{14}$$

where $\mathcal{D}_{\mathrm{text}}$ is the distribution of text embeddings from the instruction and few-shot demonstrations. The prompt-frame relevance

score is then:

$$s(\mathbf{Q}, \mathbf{F}_t) = \frac{\mathbf{q}^\top \tilde{\mathbf{F}}_t}{\|\mathbf{q}\| \|\tilde{\mathbf{F}}_t\|}. \tag{15}$$

We select the top-M' frames ($M' = 5$ by default) with the highest relevance scores as keyframes for inference:

$$\mathcal{I}^* = \text{argtop}_{M'} \; s(\mathbf{Q}, \mathbf{F}_t), \quad t \in \{1, \ldots, T\}. \tag{16}$$

This OT-based evidence distillation exploits the geometric structure of the cross-modal embedding spaces to identify diagnostically informative frames without any target-domain annotations or biomarker knowledge, replacing the ADA-based selection that requires domain-specific text queries during training.

4.2.2 Step 2: Candidate CoT Construction. Given the OT-selected keyframes $\{\mathbf{F}_t \mid t \in \mathcal{I}^*\}$, we construct the input for the frozen VLM G and generate N candidate CoT trajectories at elevated temperature τ_s (Eq. 7), following the same procedure as in training (Sec. 4.1.2). This step relies entirely on the base VLM's inherent reasoning capability — no task-specific adaptation is required.

4.2.3 Step 3: PRM-Guided Trajectory Arbitration. The pre-trained PRM $\mathcal{V}_\phi$ then scores each step of every candidate trajectory via Eq. 1. Step-level scores are aggregated into trajectory-level scores via log-odds summation (Eq. 2), and the trajectory with the highest aggregated score is selected as the final prediction:

$$c^* = \arg \max_{c \in \{1,\ldots,N\}} \mathcal{A}(\mathbf{p}^{(c)}), \quad \hat{y}^* = \hat{y}^{(c^*)}. \tag{17}$$

This best-of-N selection mechanism identifies the reasoning path that is both logically coherent (high PRM scores across steps) and grounded in informative visual evidence (OT-selected frames). The PRM's learned reward signal, trained on source-domain fused supervision, transfers to unseen modalities because the step-level scoring captures general patterns of grounded medical reasoning rather than modality-specific features.

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 evaluate on two complementary dataset categories spanning public and private sources.

Public: TCGA Pan-Cancer WSI+CT. We construct a multicenter 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**

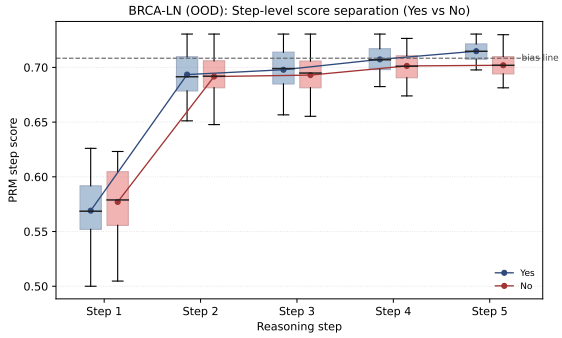


Figure 2: CT-LNE (cross-task OOD with PRM trained in TCGA) with step-level score distribution.

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 rigorously assess 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-2B [41] as the frozen LLM backbone for pan-cancer evaluation and lung CT disease classification tasks as the CoT generator and InternVL3-1B [53] as the PRM backbone. We also evaluate additional PRM backbones (Qwen2.5-3B, GPT-Qwen) and CoT builders (Gemini, template-based) in the appendix to study component interactions. All PRM training uses AdamW with learning rate 5×10^{-7} and cross-entropy loss (Eq. 12); the fusion weight β is cross-validated on a held-out meta-dataset.

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

Table 1 presents the main comparison across five TCGA cancer cohorts under both pathology and CT input modalities. Baselines including domain-specific medical 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

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.

Method	Pathology (WSI)					CT				
	BLCA (N=373)	UCEC (N=480)	BRCA (N=511)	GBMLGG (N=569)	KIRC (N=247)	BLCA (N=373)	UCEC (N=480)	BRCA (N=511)	GBMLGG (N=569)	KIRC (N=247)
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]	0.464±0.196	0.462±0.232	0.454±0.208	0.535±0.026	0.489±0.050	0.602±0.044	0.652±0.033	0.622±0.041	0.675±0.163	0.605±0.054
Proposed Pipeline (Qwen2.5-3B based)										
PRM Baseline	0.567±0.196	0.446±0.210	0.450±0.204	0.670±0.033	0.506±0.058	0.688±0.041	0.693±0.040	0.721±0.052	0.560±0.154	0.726±0.036
Zero-shot (Ours)	**0.859±0.069**	**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**

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. A vanilla PRM baseline trained without structural grounding shows promising CT results yet fails to transfer to pathology, indicating that step-level supervision alone is insufficient without cross-modal alignment.Notably, the gains are most pronounced on the pathology modality, where all prior methods struggle—this 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 case. 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—the model overwhelmingly predicts pathology-positive regardless of image content, producing near-zero specificity. This bias mirrors findings in molecular tasks where

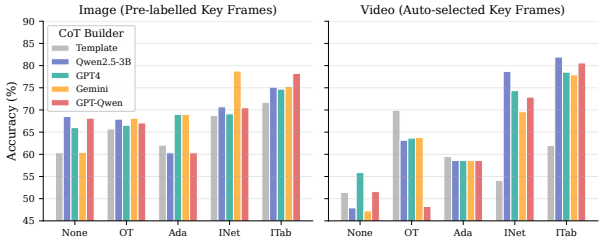


Figure 3: CoT builder comparison across structural grounding methods in CT-LNE.

LLMs default to "yes" 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. Figure 2 reveals two complementary properties of the PRM's step-level scoring when transferring from BRCA survival prediction to CT-LNE binary classification. First, 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.

Table 2: Component ablation (AUC) on two private lung CT binary classification tasks with segmentation annotations: CT-LNE (lymph node enlargement, N=103) and CT-PTX (pneumothorax, N=172). Top8 Candidates eval during inference stage

A. Cross-Modal Projection during inference stage (Evidence Distillation, Sec. 4.2.1)								
Task	**Ran-Noi**	**PCA**	**Zero-Pad**	**OT-Embed**	**Ran-Pro**	**Ada**	**Linear**	**No Linear**
CT-LNE	0.815±0.021	0.780±0.020	0.748±0.018	0.846±0.019	0.766±0.020	**0.912±0.016**	0.890±0.017	0.894±0.017
CT-PTX	0.779±0.023	0.771±0.022	0.589±0.020	0.794±0.021	0.787±0.022	**0.821±0.018**	0.806±0.018	0.785±0.019
B. Structural Grounding / Evidence Matching (Fused Step-Level Scoring, Eq. 11)								
Task	**MC-only** (β=0)	**Match-only** (β=1)	**Top5**	**Top12**	**INet** (Eq. 10)	**ITab** (Eq. 9)	**14b**	**78b**
CT-LNE	0.797±0.020	0.804±0.019	0.839±0.017	0.865±0.018	0.853±0.019	0.890±0.017	0.892±0.015	**0.901±0.015**
CT-PTX	0.696±0.022	0.688±0.021	0.726±0.019	0.727±0.020	0.734±0.021	0.767±0.019	**0.806±0.017**	0.786±0.017

Reasoning toward a negative conclusion, by contrast, requires systematically ruling out pathological features across the entire imaging volume, an inherently more uncertain task that yields less structured argumentation. The PRM's learned reward signal thus captures a clinically meaningful property: affirmative diagnostic reasoning is generally more tractable than definitive exclusion; we provide a detailed analysis of this phenomenon in the appendix.

Although the between-class score gap is compressed under cross-task OOD compared to same-task settings with early steps showing minimal separation that grows only gradually, with the spread of scores across candidate trajectories does not collapse. The box-plot whiskers and interquartile ranges remain substantial throughout the reasoning chain, indicating that 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 2), where lesion sparsity (<10% of slices) makes evidence localization particularly challenging, and complement these with CoT builder analysis (Figure 3).

Cross-Modal Projection And Evidence Quality. Table 2A compares strategies for mapping CT frame features into the text instruction space 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 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 ADA 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. Table 2B 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. 9) outperforms INet (Eq. 10), 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 3 compares CoT builders across structural grounding methods under both pre-labelled keyframe (Image) and auto-selected keyframe

(Video) settings. A key finding is that Qwen2.5-3B, operating without any external large-model guidance, produces competitive CoT trajectories across all grounding configurations—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 a framework that decouples lightweight PRM training from a training-free zero-shot inference pipeline for long-frame medical imaging analysis. Cross-source experiments—training on public WSI survival data and testing on private CT classification tasks, and vice versa—demonstrate that the framework achieves competitive zero-shot generalization across modalities and tasks. A current limitation is that the OT affine alignment assumes unimodal feature distributions, which may not hold for highly heterogeneous imaging sequences; future work will explore more expressive transport plans 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 et al. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

[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] 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

[53] 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).

[54] 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.