PaRa-OPD: Pathology Reward-Aware On-Policy Distillation via Token-Level Supervision Reweighting

Anonymous authors
Paper under double-blind review

Abstract

Pathology vision-language reasoning requires models to connect fine-grained histologic evidence with specialized diagnostic concepts, while the task-critical information in a generated rationale is often concentrated in only a small fraction of tokens. On-policy distillation (OPD) provides an effective way to transfer reasoning behavior from a teacher on student-generated trajectories, but its standard token-level objective supervises visually grounded diagnostic tokens in much the same way as generic reasoning and connective text. This uniform treatment limits the effectiveness of OPD on pathology tasks. Existing online visual-gain weighting methods attempt to identify visually dependent tokens during OPD, but their signal can weaken when the sampled student token is already visually incorrect. To make OPD more effective for pathology reasoning, we introduce **PaRa-OPD**, a pathology reward-aware OPD framework for token-level supervision reweighting that directs distillation toward tokens that either depend on visual evidence or encode pathology-specific diagnostic information. PaRa-OPD trains a Pathology Reward Model (PRM) to identify these complementary token patterns in generated trajectories. The resulting PRM scores are converted into token-level weights that dynamically reweight the OPD objective, strengthening supervision where visual grounding and diagnostic knowledge are most consequential. On PathMMU, PaRa-OPD improves Qwen3.5-2B from 63.88 to 71.44 and Qwen3.5-4B from 66.96 to 74.96 over standard OPD, demonstrating strong and consistent effectiveness across model scales. Code and processed data will be released in the future.

1 Introduction

Pathology vision-language reasoning requires models to connect fine-grained histologic evidence with specialized diagnostic concepts. A correct answer often depends on recognizing subtle morphology, tissue architecture, staining patterns, and spatial relationships, and then mapping these observations to medical terminology. In generated rationales, however, such task-critical information is rarely distributed uniformly across the whole response. A small number of tokens may carry the decisive visual observation or diagnostic concept, while many others serve as generic reasoning scaffolds, option comparisons, or connective text.

On-policy distillation (OPD)(Agarwal et al., 2024; Zhao et al., 2026; Jin et al., 2026; Song & Zheng, 2026) has recently emerged as an effective training paradigm between supervised fine-tuning (SFT)(Ouyang et al., 2022; Wei et al., 2021; Wang et al., 2023; Touvron et al., 2023) and reinforcement learning (RL)(Schulman et al., 2017; Rafailov et al., 2023; Shao et al., 2024; Yu et al., 2026). By supervising student-generated trajectories with teacher token distributions, OPD mitigates the exposure mismatch of SFT while avoiding the sparse rewards, high-variance optimization, and stability challenges commonly associated with RL. Despite these advantages, directly applying OPD-based training paradigms to pathology vision-language tasks still yields limited and sometimes unstable improvements. This exposes a practical mismatch between the standard OPD objective and pathology reasoning: the objective supplies dense token-level supervision, but it does not distinguish whether a token encodes image-grounded evidence, pathology-specific diagnostic information, or generic language structure.

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

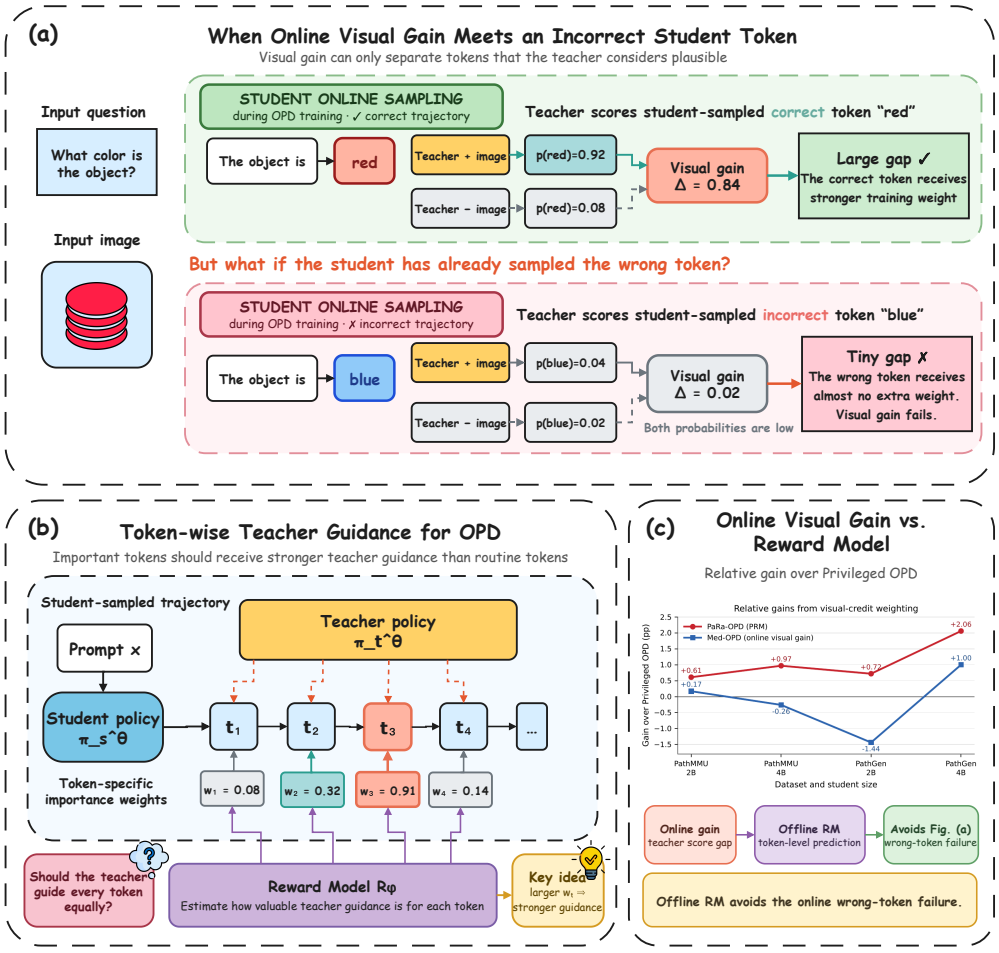


Figure 1: Motivation for PaRa-OPD. (a) Online visual-gain weighting can underweight incorrect sampled tokens. (b) OPD should allocate supervision weights non-uniformly across tokens. (c) PaRa-OPD yields more stable component gains than online visual-gain weighting.

Motivated by this mismatch, recent OPD variants have begun to depart from uniform token weighting, treating the allocation of teacher guidance itself as part of the training design (Xu et al., 2026b; Wang et al., 2026; Zheng et al., 2026). Figure 1(b) illustrates this broader motivation: different training settings and optimization goals may call for additional supervision on different subsets of generated tokens, rather than applying the same distillation strength uniformly across the entire trajectory. In the multimodal setting, online visual-gain weighting is a representative attempt to realize such non-uniform supervision: for each sampled student token, it evaluates the teacher-assigned probability under an image-conditioned prompt and under a counterfactual prompt without the image, or with a weakened visual signal such as a perturbed image, and uses the probability gap as the visual-advantage score(Liu et al., 2026; Qian et al., 2026). This strategy is especially relevant to pathology, where reliable reasoning depends on whether the model actually uses histologic evidence rather than relying on medical language priors alone. Yet the online gain signal remains tied to the token that the student has already sampled. As illustrated in Figure 1(a), when the sampled token is visually correct, the image-conditioned teacher can assign it a much higher probability than the no-image teacher, producing the desired large gain. When the sampled token is visually wrong, both teachers may assign low probability to that token, so the probability gap becomes small precisely where stronger correction is needed.

Motivated by this observation, we propose PaRa-OPD, a pathology reward-aware OPD framework for token-level supervision reweighting. PaRa-OPD first constructs PRM training data from correct

reasoning trajectories, using image/no-image counterfactuals to identify visually dependent tokens and a pathology expert vocabulary signal to identify tokens strongly associated with diagnostic reasoning. A lightweight Pathology Reward Model (PRM) is then trained to predict these token patterns on new student trajectories. During OPD, PRM scores are converted into normalized token-level weights that reweight the reverse-KL distillation loss, strengthening supervision on tokens that are most consequential for visual grounding and pathology reasoning.

We evaluate PaRa-OPD by systematically studying the adaptation of OPD to pathology vision-language reasoning. Across PathMMU(Sun et al., 2024) and PathGen-Instruct(Sun et al., 2025), we compare SFT and rejection-SFT with standard OPD, TRD, VA-OPD, privileged OPD, Med-OPD, GRPO, and PaRa-OPD using both 2B and 4B students. The results show that privileged OPD is already a strong baseline for pathology VLMs, while PRM-based token weighting brings further gains and more stable behavior. On PathMMU's official Tiny test split, PaRa-OPD improves 2B and 4B students over privileged OPD by 0.62 and 0.97 percentage points, respectively, and the cross-dataset PathGen-Instruct Test1800 evaluation shows consistent advantages over the main OPD variants.

Our key contributions are as follows:

- To the best of our knowledge, we are the first to integrate an explicit reward model into OPD training for token-level supervision reweighting, turning dense teacher supervision from a uniform objective into a reward-aware distillation signal.
- We propose PaRa-OPD, which trains a Pathology Reward Model to identify visually dependent and pathology-informative tokens from generated reasoning trajectories, and converts its scores into token-level weights for reweighting the OPD objective.
- We systematically investigate the adaptation of OPD to pathology vision-language reasoning, comparing standard OPD, TRD, VA-OPD, privileged OPD, Med-OPD, GRPO, and PaRa-OPD across PathMMU and PathGen-Instruct with 2B and 4B students, together with analyses of PRM behavior and qualitative cases.

2 Related Work

2.1 Vision-language models for pathology

Pathology vision-language models have developed rapidly, driven by stronger multimodal pretraining recipes and the growing availability of pathology image-text data. Representative systems include PathChat, a multimodal generative AI copilot designed for interactive pathology assistance (Lu et al., 2024b); CONCH, a contrastive vision-language foundation model for computational pathology that supports cross-modal retrieval and downstream recognition tasks (Lu et al., 2024a); and TITAN, a whole-slide multimodal foundation model that extends pathology vision-language modeling from patch-level images to gigapixel slide representations (Ding et al., 2025). These models indicate a broader shift from task-specific pathology classifiers toward general-purpose pathology VLMs that can connect histologic morphology with natural-language concepts.

Meanwhile, a series of pathology-oriented VQA and instruction-following benchmarks have been proposed to evaluate visual understanding and reasoning over histology images, including PathVQA, PathMMU, and PathGen-Instruct (He et al., 2020; Sun et al., 2024; 2025).

2.2 Knowledge distillation and on-policy distillation

Knowledge distillation transfers behavior from a stronger teacher to a smaller student through hard labels, generated rationales, or soft token distributions (Hinton et al., 2015). In sequence-generation settings, supervised fine-tuning on teacher outputs can suffer from exposure mismatch: the student is trained on teacher trajectories but evaluated on its own generations. On-policy distillation addresses this by sampling student trajectories and asking the teacher to supervise the states that the student actually visits. This idea appears in generalized knowledge distillation for autoregressive models (Agarwal et al., 2024), reverse-KL distillation for LLM compression (Gu et al., 2024), and efficiency-oriented variants such as DistiLLM (Ko et al., 2024).

Compared with SFT and RL, OPD can be viewed as an intermediate training paradigm. SFT provides dense token supervision but only on fixed reference trajectories, while RL optimizes on-policy samples but often relies on sparse trajectory-level rewards. OPD keeps the on-policy states visited by the student and replaces sparse rewards with dense teacher distributions. Given an input x, a student trajectory $y \sim p_\phi(\cdot \mid x)$, a student p_ϕ, and a teacher p_θ, the standard reverse-KL OPD objective is

$$\mathcal{L}_{\mathrm{OPD}}(\phi) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim p_\phi(\cdot|x)} \sum_t \mathcal{D}_{\mathrm{RKL}}(p_\phi(\cdot \mid x, y_{<t}) \,\|\, p_\theta(\cdot \mid x, y_{<t}))\,.$$

Recent OPD research has further expanded this training paradigm along several directions. Self-distillation variants use the model itself under privileged answer information as an augmented teacher for its own on-policy trajectories (Zhao et al., 2026; Shenfeld et al., 2026). Trajectory-repair variants revise flawed student rollouts before distillation, so teacher supervision is applied to more informative corrected states (Jiang et al., 2026; Xu et al., 2026a). In vision-language settings, some work strengthens the teacher signal by exposing fine-grained image evidence (Yuan et al., 2026), while another line assigns token-wise distillation weights according to online visual advantage, measured by the teacher's probability gap on the same sampled student token between an image-conditioned prompt and a counterfactual prompt with no image or weakened visual evidence (Liu et al., 2026). In medical VLMs, related approaches combine privileged information with online visual-gain weighting for pathology reasoning (Qian et al., 2026).

These extensions show that OPD can be improved by enriching the teacher context or by making token-level supervision less uniform. In pathology VQA, this direction is especially relevant because reliable reasoning often depends on whether the model actually uses histologic evidence rather than relying on medical language priors alone. Existing online visual-gain weighting methods attempt to capture this dependence during OPD, but as discussed in the Introduction, their signal can become weak when the sampled student token is already visually incorrect. We therefore move visual-credit estimation outside the OPD loop: PaRa-OPD learns an offline Pathology Reward Model from counterfactual visual-credit labels and pathology vocabulary signals, and uses it to reweight OPD supervision at the token level while keeping the student input unchanged.

3 METHOD

3.1 PROBLEM SETUP

Given an image-question pair $x = (I, q)$, a student model generates a reasoning trajectory $y = (y_1, \ldots, y_T)$. A stronger teacher provides token-level distillation supervision on the same trajectory. The goal is to improve the student by reweighting teacher supervision toward tokens whose correctness depends on visual evidence.

Figure 2 summarizes the overall PaRa-OPD pipeline. We first use image/no-image teacher-forced scoring to assign counterfactual visual credit to sampled reasoning tokens. These token-level signals, together with pathology vocabulary matches, supervise an offline PRM. During OPD, the student still learns from teacher token distributions on its own generated trajectory, but the distillation loss is reweighted by PRM scores so that visually dependent pathology evidence receives stronger guidance.

3.2 PRIVILEGED OPD

We first define the privileged OPD baseline that PaRa-OPD builds on. In standard OPD, the student samples a trajectory from the original image-question input $x = (I, q)$, and the teacher supervises the same on-policy prefixes. Privileged OPD keeps the student input unchanged but augments the teacher-side context with privileged pathology information. Let $x^+ = (I, q, a^\star, e^\star)$ denote a teacher prompt that additionally contains the reference answer $a^\star$ and reference explanation or diagnostic rationale $e^\star$. For a student trajectory $y \sim p_\phi(\cdot \mid x)$, the privileged teacher supplies token distributions conditioned on x^+ and the student prefix $y_{<t}$:

$$\ell_{\mathrm{PrivOPD}}(t) = \mathcal{D}_{\mathrm{RKL}}\big(p_\phi(\cdot \mid x, y_{<t}) \,\big\|\, p_\theta(\cdot \mid x^+, y_{<t})\big)\,.$$

The student therefore learns from teacher feedback on its own visited states, while privileged information remains available only to the teacher during training.

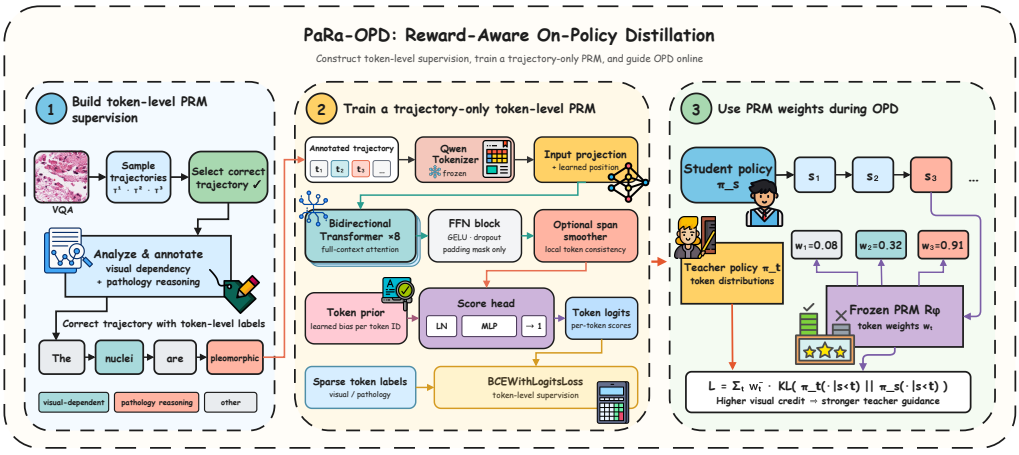


Figure 2: Overview of PaRa-OPD. We first sample reasoning trajectories and annotate tokens through visual-dependence analysis and pathology-reasoning vocabulary matching. These labels train an offline PRM, whose scores are converted into token-level weights for subsequent OPD. The final distillation objective keeps teacher token supervision but reweights it toward visually dependent and pathology-informative tokens.

This construction is inspired by on-policy self-distillation (OPSD)(Zhao et al., 2026), where privileged information can make the model itself a stronger teacher for its own on-policy trajectories. We extend this idea beyond self-distillation: privileged information is also useful in teacher–student OPD, where a stronger external teacher distills into a smaller student. In our setting, we keep the student input unchanged and provide richer privileged pathology information directly to the external teacher. This gives a strong OPD baseline for pathology VQA without changing the student input or inference-time interface.

3.3 Pathology Reward Model

The Pathology Reward Model (PRM) is designed to identify which tokens in a reasoning trajectory should receive stronger distillation supervision. It addresses the failure mode illustrated in Figure 1(a): online visual-gain weighting evaluates the currently sampled student token, so a visually wrong token can receive a small image/no-image probability gap when both teacher variants assign it low probability. We avoid constructing PRM supervision from such incorrect online student tokens. Instead, we build offline token labels from answer-correct teacher trajectories, so visual credit is estimated on trajectories that already lead to the correct option.

Sampling. We sample Qwen3.5-9B reasoning trajectories on PathMMU and PathGen-Instruct under image-conditioned prompts and no-image counterfactual prompts. We retain answer-correct image-conditioned trajectories from cases where no-image samples are majority wrong over eight attempts. This selection focuses the PRM training data on examples where the image is likely necessary for the correct reasoning trajectory. The resulting PRM training pool contains 30,662 selected cases, 125,076 labeled trajectories, and 9.96M trajectory tokens.

Token annotation. For each retained trajectory, we compute counterfactual visual credit by teacher-forcing the same token sequence with and without the pathology image:

$$g_t = p_\theta(y_t \mid I, q, y_{<t}) - p_\theta(y_t \mid q, y_{<t}).$$

High g_t indicates that the image increases the teacher's confidence in token y_t under the same textual prefix. We convert this continuous score into binary visual-dependence supervision by marking tokens in the global top 10% of visual credit and tokens among the top 5 within each trajectory.

We further add pathology expert-vocabulary supervision. We first form word candidates from vocabulary entries whose underlying tokens appear in the global top 30% of visual-credit tokens. An LLM judge then filters this candidate set to retain words and phrases strongly related to pathology reasoning or medical professional concepts, such as morphology, stain or marker patterns, tissue

<reasoning>The inset reveals a high nuclear-to-cytoplasmic ratio with nuclei that exhibit a "salt and pepper" appearance, where the chromatin is finely granular and evenly dispersed without the coarse clumping or prominent nucleoli typically seen in high-grade carcinomas or lymphomas. This specific fine, stippled texture is characteristic of neuroendocrine differentiation.</reasoning>
<answer>A</answer>

Figure 3: Trajectory-level PRM illustration. Tokens highlighted in the trajectory receive higher PRM-derived supervision weights.

architecture, anatomical locations, and diagnostic entities. For every trajectory, matched word spans are projected back to tokenizer pieces, and all covered tokens are marked as pathology-vocabulary positives. The final PRM target is

$$z_t = \mathbf{1}\{z_t^{\text{vis}} = 1 \text{ or } z_t^{\text{path}} = 1\},$$

where z_t^{vis} is the visual-credit label and z_t^{path} is the pathology-vocabulary label.

Training. We define the PRM as a non-generative token-level scorer $r_\psi(y)_t$ over pathology reasoning trajectories. Given only the trajectory token IDs $y = (y_1, \ldots, y_T)$, the PRM predicts one scalar logit for each token. It does not receive the image, question, answer options, final answer, or privileged reference information at inference time, preventing it from acting as a second VQA model. The network uses a frozen Qwen token embedding, a bidirectional Transformer encoder, a lightweight token-wise score head, a learned token prior, and a local span smoother. We train it with binary cross-entropy against z_t:

$$\mathcal{L}_{\text{PRM}} = -\frac{1}{\sum_t m_t} \sum_{t=1}^{T} m_t \left[z_t \log \sigma(r_\psi(y)_t) + (1 - z_t) \log(1 - \sigma(r_\psi(y)_t)) \right],$$

where m_t is the response-token mask. The trained PRM is then used as a soft token-weighting model rather than as a hard verifier; Figure 3 illustrates the resulting token-level scores on a generated pathology reasoning trajectory.

3.4 PRM-BASED SUPERVISION REWEIGHTING

During final OPD training, the student first generates a trajectory from the original image-question input x. The privileged teacher supplies the reverse-KL token loss $\ell_{\text{PrivOPD}}(t)$, while the PRM supplies a token score $r_\psi(y)_t$. We convert the PRM score into a positive token weight and normalize it over response tokens:

$$w_t = 1 + \lambda \sigma(r_\psi(y)_t), \qquad \bar{w}_t = \frac{w_t}{\frac{1}{\sum_j m_j} \sum_j m_j w_j}. \tag{1}$$

The final training objective is

$$\mathcal{L}_{\text{PaRa-OPD}} = \frac{1}{\sum_t m_t} \sum_{t=1}^{T} m_t \, \bar{w}_t \, \ell_{\text{PrivOPD}}(t), \tag{2}$$

where the normalization preserves the average loss scale. In our main experiments, privileged OPD provides the teacher prompt and PaRa-OPD adds only PRM-derived token weighting. We do not combine PaRa-OPD with wrong-tail masking, corrective teacher prompts, or additional SFT losses in the reported main results.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We use PathMMU(Sun et al., 2024) as the main training and evaluation benchmark. PathMMU is a multiple-choice pathology VQA benchmark covering five data sources: PubMed, Atlas, EduContent, SocialPath, and PathCLS. We use the prepared full PathMMU training pool with 8,445 usable image-question pairs.

Evaluation uses PathMMU's official Tiny test split, which contains 1,138 usable examples after filtering cases with missing images. Table 1 reports the subset composition after this filtering.

Table 1: PathMMU data used in our experiments. Counts are usable image-question pairs after filtering missing images.

Subset	Training pool	Tiny test evaluation
PubMed	2,787	281
Atlas	799	208
EduContent	1,683	255
SocialPath	1,544	217
PathCLS	1,632	177
Total	8,445	1,138

We also use small-patch PathGen-Instruct(Sun et al., 2025) VQA data as a cross-dataset evaluation setting and as an additional source for reward-model construction. For evaluation, we report results on a 1,800-example PathGen-Instruct diagnostic test set. For PRM training, we combine visually dependent cases from PathMMU and PathGen-Instruct, yielding 30,662 selected cases, 125,076 labeled trajectories, and 9.96M trajectory tokens.

Models and baselines. Qwen3.5 (Qwen Team, 2026) is a family of multimodal large vision-language models that accept interleaved image-text inputs and generate text responses. In our pathology VQA setting, each model receives the pathology image and a multiple-choice question, then generates a short reasoning trajectory and a final answer. We use Qwen3.5-2B and Qwen3.5-4B as student models, and Qwen3.5-9B as the teacher model. The teacher provides token-level supervision for OPD, and is also used to construct visual-credit labels by comparing image-conditioned and no-image teacher-forced token probabilities.

We compare against base Qwen3.5 students, SFT, rejection-sampled SFT(Liu et al., 2024), GRPO(Shao et al., 2024), standard OPD, TRD(Jiang et al., 2026), VA-OPD(Liu et al., 2026), privileged OPD, Med-OPD(Qian et al., 2026), and PaRa-OPD. SFT uses reasoning traces synthesized from PathMMU explanations, while Rejection-SFT retains answer-correct trajectories sampled from Qwen3.5-9B. Among OPD variants, privileged OPD gives the teacher reference answer and explanation information, TRD distills from teacher-repaired trajectories, VA-OPD weights token losses by the teacher probability gap between image-conditioned and no-image or weak-image prompts, and Med-OPD combines this online visual-advantage weighting with privileged teacher information. GRPO uses binary answer-level rewards over multiple sampled responses. PaRa-OPD instead replaces online visual-advantage scoring with an offline PRM for token-level OPD weighting.

4.2 Results

4.2.1 Main PathMMU results

Table 2 shows that PaRa-OPD improves both model sizes over strong privileged OPD baselines. On 2B, PaRa-OPD reaches 71.44, improving over privileged OPD by 0.62 percentage points. On 4B, PaRa-OPD reaches 74.96, improving over privileged OPD by 0.97 percentage points and giving the strongest current result. Med-OPD is competitive but less stable than PaRa-OPD.

The comparison also highlights the strength of privileged teacher supervision. Basic OPD improves over the base students but remains below SFT on both model sizes. GRPO with sparse answer-level rewards is not competitive in this low-data pathology setting, especially for the 2B student. Privileged OPD closes much of the SFT gap and becomes the strongest non-reward-aware baseline. PaRa-OPD should therefore be interpreted as an incremental gain on top of a strong privileged OPD recipe rather than as a replacement for privileged supervision.

4.2.2 PathGen-Instruct results

We further evaluate cross-dataset transfer on the PathGen-Instruct close-patch VQA diagnostic test set. Since this evaluation set does not include subset labels, Table 3 reports the number of correct predictions and overall accuracy.

PaRa-OPD also transfers to PathGen-Instruct despite the different data source and lower absolute accuracy. The Qwen3.5-9B teacher reference obtains 955/1800 correct predictions, corresponding to 53.06% accuracy. For 2B, PaRa-OPD improves over privileged OPD by 0.72 percentage points and over Med-OPD by 2.16 points. For 4B, PaRa-OPD improves over privileged OPD by 2.06 points, over Med-OPD by 1.06 points, and slightly exceeds the 9B teacher reference. GRPO remains below rejection-SFT and OPD for 2B, and reaches only 48.22 on 4B, indicating that sparse answer-

Table 2: PathMMU Tiny test accuracy. Bold and underline indicate the best and second-best results within each student size; shaded rows show PaRa-OPD and gains over standard OPD.

Size	Method	PubMed	Atlas	EduContent	SocialPath	PathCLS	Overall
9B	Teacher	70.46	72.12	69.02	67.74	61.58	68.54
2B	Base	55.16	62.50	59.22	57.60	55.37	57.91
	SFT	68.68	**73.56**	69.41	70.05	56.50	68.10
	Rejection-SFT	62.63	60.58	65.10	61.29	66.67	63.18
	GRPO(Shao et al., 2024)	63.70	63.46	60.00	62.21	48.02	60.11
	OPD(Agarwal et al., 2024)	63.35	67.31	65.49	59.91	63.28	63.88
	TRD(Jiang et al., 2026)	66.19	68.27	67.45	63.13	65.54	66.17
	VA-OPD(Liu et al., 2026)	69.04	69.23	71.76	71.89	64.41	69.51
	Privileged OPD	70.11	70.19	72.55	74.65	65.54	70.83
	Med-OPD(Qian et al., 2026)	**71.17**	70.19	70.98	**75.12**	66.67	71.00
	PaRa-OPD (Ours)	70.82	69.71	**73.33**	73.27	**69.49**	**71.44**
	△ vs standard OPD	+7.47	+2.40	+7.84	+13.36	+6.21	+7.56
4B	Base	66.19	69.23	70.59	63.59	55.93	65.64
	SFT	74.73	74.52	74.90	73.73	58.19	71.97
	Rejection-SFT	67.62	67.79	73.73	67.28	66.67	68.80
	GRPO(Shao et al., 2024)	73.67	68.27	72.16	67.28	57.63	68.63
	OPD(Agarwal et al., 2024)	69.04	67.31	70.59	67.28	57.63	66.96
	TRD(Jiang et al., 2026)	70.82	68.27	72.55	68.20	64.41	69.24
	VA-OPD(Liu et al., 2026)	72.24	69.71	73.33	69.59	63.84	70.21
	Privileged OPD	75.09	**75.96**	76.47	72.35	68.36	73.99
	Med-OPD(Qian et al., 2026)	76.51	72.60	**77.65**	73.27	65.54	73.73
	PaRa-OPD (Ours)	**78.65**	74.52	74.90	**75.12**	**69.49**	**74.96**
	△ vs standard OPD	+9.61	+7.21	+4.31	+7.84	+11.86	+8.00

Table 3: PathGen-Instruct Test1800 results. Bold and underline indicate the best and second-best results within each student size; shaded rows show PaRa-OPD and gains over standard OPD.

Qwen3.5-2B

Method	Correct / Total	Acc
Base	758 / 1800	42.11
Rejection-SFT	886 / 1800	49.22
GRPO	832 / 1800	46.22
OPD	885 / 1800	49.17
TRD	891 / 1800	49.50
VA-OPD	901 / 1800	50.06
Privileged OPD	938 / 1800	52.11
Med-OPD	912 / 1800	50.67
PaRa-OPD (Ours)	**951 / 1800**	**52.83**
△ vs standard OPD	+66	+3.66

Qwen3.5-4B

Method	Correct / Total	Acc
Base	843 / 1800	46.83
Rejection-SFT	910 / 1800	50.56
GRPO	868 / 1800	48.22
OPD	872 / 1800	48.44
TRD	894 / 1800	49.67
VA-OPD	912 / 1800	50.67
Privileged OPD	924 / 1800	51.33
Med-OPD	942 / 1800	52.33
PaRa-OPD (Ours)	**961 / 1800**	**53.39**
△ vs standard OPD	+89	+4.95

level rewards are weaker than token-level distillation signals in this setting. This result is important because the PathGen-Instruct diagnostic set is not the same benchmark used for the main PathMMU table.

4.2.3 Component ablation

Table 4 isolates the two main ingredients of PaRa-OPD: privileged teacher context and PRM-based token weighting. PRM weighting improves over standard OPD even without privileged teacher context, increasing PathMMU accuracy by 1.15 and 1.41 points for 2B and 4B students, respectively. Privileged teacher supervision provides a larger gain, and combining it with PRM weighting gives the best result across all four settings. This indicates that PRM-based supervision reweighting is complementary to teacher-side privileged information rather than merely duplicating its effect.

Table 4: Component ablation for PaRa-OPD. Entries are accuracies in percentage points.

Teacher context	Token weighting	PathMMU 2B	PathMMU 4B	PathGen 2B	PathGen 4B
Standard teacher	Uniform (OPD)	63.88	66.96	49.17	48.44
Privileged teacher	Uniform (Privileged OPD)	70.83	73.99	52.11	51.33
Standard teacher	PRM weighting	65.03	68.37	50.11	50.94
Privileged teacher	PRM weighting (PaRa-OPD)	**71.44**	**74.96**	**52.83**	**53.39**

4.3 Analysis

PaRa-OPD improves most consistently when token weighting is added on top of privileged teacher supervision. Both Med-OPD and PaRa-OPD start from this strong privileged OPD setting and then reallocate token-level distillation strength. The difference is how the token weights are obtained: Med-OPD computes an online visual-advantage signal on the currently sampled student token, whereas PaRa-OPD uses an offline PRM trained to predict visually dependent and pathology-informative tokens. This distinction matters empirically. Relative to privileged OPD, Med-OPD changes accuracy by $+0.17$ points on PathMMU-2B, -0.26 on PathMMU-4B, -1.44 on PathGen-2B, and $+1.00$ on PathGen-4B. Thus, despite using the same privileged teacher context, online visual-advantage weighting underperforms the unweighted privileged baseline in two of the four settings.

This instability is consistent with the mechanism illustrated in Figure 1(a): when the sampled student token is already visually incorrect, the image-conditioned and no-image teacher distributions may both assign it low probability, producing a weak online advantage signal at precisely the token where stronger correction is needed. In contrast, PaRa-OPD yields positive gains over privileged OPD in all four settings: $+0.62$ and $+0.97$ points on PathMMU-2B/4B, and $+0.72$ and $+2.06$ points on PathGen-2B/4B. As summarized in Figure 1(c), the offline PRM-based weighting therefore provides a more stable component gain than online visual-advantage weighting. Table 4 further shows that PRM weighting improves standard OPD even without privileged teacher context, and gives the best results when combined with privileged supervision. Appendix B.1 shows that a continuous sigmoid PRM weight with moderate scale is preferable to either raw scores or hard token masks.

4.4 PRM analysis

The trained PRM provides a measurable token-level signal rather than only an auxiliary heuristic. Importantly, this signal does not replace the teacher distribution or directly judge final-answer correctness. It only determines how much reverse-KL supervision each generated token receives, so the semantic target of distillation remains the privileged teacher while the PRM controls the allocation of supervision strength. On the held-out token-level validation split, it achieves an AUROC of 0.924, with precision 0.613 and recall 0.742 at threshold 0.5; the full validation table and ROC curve are reported in Appendix C.1. The target is also sparse: visual-credit labeling marks 10.31% of trajectory tokens, and the combined visual-or-expert target covers 12.74% after adding pathology expert-vocabulary matches (Appendix C.2). Additional diagnostics in Appendix C show the visual-credit score distribution, vocabulary-frequency patterns, labeled training examples, and post-training PRM scoring examples, supporting the interpretation that the PRM primarily upweights compact visual and pathology-specific spans rather than entire responses.

5 Conclusion

We introduced counterfactual visual credit as a token-level supervision signal for pathology VLM distillation. PaRa-OPD uses an offline reward model to predict visually dependent tokens and reweight OPD loss on student-generated trajectories. Across PathMMU and PathGen-Instruct, this approach improves over strong privileged OPD baselines for both 2B and 4B students. The results support token-level visual credit as a practical direction for pathology VLM distillation, while also showing that separating visual evidence from medical language priors remains a central challenge.

AI USE STATEMENT

We used AI tools for writing polish, code and experiment assistance, and data processing. All experimental claims, analyses, and manuscript contents were reviewed and verified by the authors.

ETHICS STATEMENT

This work uses open-source pathology datasets for research on pathology vision-language models. The models studied in this paper are intended for research evaluation and should not be used for autonomous clinical diagnosis.

REPRODUCIBILITY STATEMENT

We will release the code and processed data to support reproducibility. The appendix provides additional implementation details, prompts, training configurations, and ablation settings.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, 31(11):3749–3761, 2025.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Li Jiang, Haoran Xu, Yichuan Ding, and Amy Zhang. Trajectory-refined distillation. *arXiv preprint arXiv:2606.08432*, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *International Conference on Machine Learning*, 2024.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. In *International conference on learning representations*, volume 2024, pp. 54605–54624, 2024.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024a.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024b.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Yunhang Qian, Jiaquan Yu, Jiawei Liu, Meng Wang, Hongwei Bran Li, and Xiaobin Hu. Med-opd: Improving medical vision-language models via evidence-aware on-policy distillation. *arXiv preprint arXiv:2607.16303*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026. Official model release blog.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, et al. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, pp. 56–73. Springer, 2024.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Jingxiong Li, Xuan Gong, Xinheng Lyu, Tao Lin, et al. Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In *International Conference on Learning Representations*, volume 2025, pp. 94611–94653, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Haolei Xu, Xiaowen Xu, Haiwen Hong, Zixuan Ni, Hongxing Li, Yiwen Qiu, Weiming Lu, and Yongliang Shen. Pass the baton: Trajectory-relayed on-policy distillation. *arXiv preprint arXiv:2607.26057*, 2026a.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. Scope: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *arXiv preprint arXiv:2604.10688*, 2026.

A EXPERIMENTAL SETUP DETAILS

A.1 EXPERIMENTAL PARAMETERS

Tables 5, 6, 7, and 8 summarize the training data scale and main optimization, sampling, and inference settings used in our experiments. Data counts refer to unique training examples and prompts used by the algorithm. For generation-based methods, training uses visible reasoning prompts with a maximum generation length of 512 tokens.

Table 5: PathMMU training data used by each algorithm.

Algorithm	Training data
Base	No task-specific training.
SFT	8,445 reference reasoning examples.
Rejection-SFT	8,445 selected correct 9B trajectories from 7,163 unique cases.
OPD-family methods	8,445 on-policy prompts. This group includes OPD, TRD, VA-OPD, Privileged OPD, Med-OPD, and PaRa-OPD; the variants differ in trajectory repair, teacher context, and token weighting.
GRPO	8,445 prompts with 8 sampled responses per prompt.

Table 6: PathGen-Instruct training data.

Algorithm	Training data
Base	No task-specific training.
Rejection-SFT	30,000 selected correct 9B trajectories, one per case/question.
OPD-family methods	30,000 on-policy prompts. This group includes OPD, TRD, VA-OPD, Privileged OPD, Med-OPD, and PaRa-OPD; the variants differ in trajectory repair, teacher context, and token weighting.
GRPO	5,000 prompts with 8 sampled responses per prompt.

Table 7: Default optimization configuration.

Training stage	Batch size	Micro batch	Rollout batch	Learning rate
SFT	64	2	–	2×10^{-6}
Rejection-SFT	64	2	–	2×10^{-6}
OPD-family methods	64	2	64	2×10^{-6}
GRPO	64	2	8	1×10^{-6}

Training uses 2–8 NVIDIA H20 GPUs depending on model size and method. GRPO uses reward 1.0 for correct answers and 0.0 otherwise, with clip range 0.2. Batch-size-only adjustments required by memory limits were not treated as algorithmic hyperparameters.

A.2 PROMPT TEMPLATES

The student always receives the standard image-question prompt below; the same prompt format is used for test-time inference. Privileged information is exposed only to the teacher during OPD training. The placeholder `<image>` denotes the image slot in KDFlow records and corresponds to the image content block in API-based inference. For all generation and evaluation runs, we disable the model's native thinking mode and use a prompt-controlled visible CoT format instead: the model is instructed to place its reasoning in `<reasoning>...</reasoning>` and its final option in `<answer>...</answer>`. This keeps the reasoning trace visible, bounded, and consistently

Table 8: Sampling and inference configuration.

Stage	Samples per prompt	Temperature	Top-p	Maximum tokens
OPD-family training	1	0.7	0.95	512
GRPO training	8	0.7	0.95	512
Test-time inference	1	0.0	1.0	512

parseable across training and inference. Figures 4 and 5 show concrete examples of the standard student prompt and the privileged teacher prompt, respectively.

Prompt 1 gives the standard student-side template used for rollout generation and test-time inference. Prompt 2 gives the privileged teacher-side template used only for OPD supervision.

Prompt 1: Student and inference prompt template

```
SYSTEM:
You are a pathology vision-language assistant. Use only the histology image
and the provided multiple-choice options. Give concise, evidence-based
pathology reasoning, then one final option letter.

USER:
<image>
Please answer this pathology multiple-choice question.

Question:
[QUESTION]

Options:
[OPTIONS]

Reasoning guidance:
- Ground the reasoning in visible histology, morphology, staining, or tissue
  architecture.
- Compare plausible options only when it helps decide the answer.
- Keep the reasoning to at most two sentences and do not invent clinical
  context.

Output exactly this format:
<reasoning>one or two concise pathology reasoning sentences</reasoning>
<answer>X</answer>

Put exactly one valid option letter inside <answer>; do not include the option
text there.
```

Prompt 2: Privileged teacher prompt template

```
SYSTEM:
You are a pathology teacher model for on-policy distillation. You may use
teacher-only privileged reference information to calibrate your reasoning,
but your final response should read as if it was derived from the histology
image and options.

USER:
<image>
```

```
Please answer this pathology multiple-choice question.

Question:
[QUESTION]

Options:
[OPTIONS]

Teacher-only privileged reference:
- Correct answer: [CORRECT ANSWER]
- Reference reasoning: [REFERENCE REASONING]

Use the privileged reference only to know the intended pathology conclusion.
Then reconstruct a concise visual reasoning path as if you did not already
know the answer. The reasoning should be suitable for guiding a smaller
student model. Do not mention the privileged reference or reveal that the
answer was provided to you. Ground the reasoning in visible histology,
morphology, staining, or tissue architecture.

Output exactly this format:
<reasoning>one or two concise pathology reasoning sentences</reasoning>
<answer>[CORRECT LETTER]</answer>

Put exactly the correct option letter inside <answer>; do not include the
option text there.
```

The student-side prompt is unchanged in all privileged OPD variants. Figure 4 illustrates the prompt seen by the student at both training-time rollout generation and test-time inference, while Figure 5 illustrates the teacher-only fields used to provide privileged OPD supervision.

B Ablation studies

This section reports additional ablations for PaRa-OPD, focusing on the PRM weighting function, the PRM scorer architecture, and the granularity of teacher-side privileged information. The main component ablation is reported in Table 4.

B.1 PRM weighting variants

This subsection ablates how PRM scores are injected into the OPD objective on the PathMMU 4B setting; Table 9 reports the resulting accuracy. All rows use the same privileged teacher setup; the baseline is therefore Privileged OPD with uniform token weights. Following Eq. 1 and Eq. 2, we write a general token-weighting form as

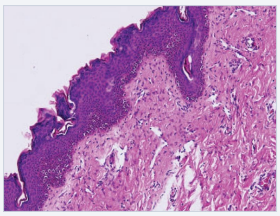
$$w_t = 1 + \lambda a_t, \qquad \bar{w}_t = \frac{w_t}{\frac{1}{\sum_j m_j} \sum_j m_j w_j},$$

where a_t is the token-level signal derived from the PRM output. The main PaRa-OPD configuration uses $a_t = \sigma(r_\psi(y)_t)$ with $\lambda = 1.0$. The ablation therefore studies two components relative to Privileged OPD: the weight scale λ under the same sigmoid mapping, and the score-to-weight mapping itself under a fixed $\lambda = 1.0$. The mapping variants compare the sigmoid probability, direct use of the predicted score, and a hard 0/1 mask obtained by thresholding the sigmoid probability at 0.5.

The best setting uses the sigmoid PRM probability with $\lambda = 1.0$, improving PathMMU 4B accuracy from 73.99 to 74.96 over Privileged OPD. Reducing the scale to 0.5 still improves over the baseline, while increasing it to 2.0 degrades performance, indicating that overly sharp token reweighting can destabilize distillation. Directly using the raw PRM score is weaker than the sigmoid probability, and the hard 0/1 mask performs worst, supporting the use of continuous calibrated token weights.

Student inference prompt examples

CASE 01



Pathology image

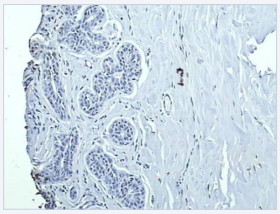
SYSTEM MESSAGE

```
You are a pathology vision-language assistant. Use only the histology image and the provided multiple-choice options. Give
concise, evidence-based pathology reasoning, then one final option letter.
```

USER MESSAGE

```
<image>
Please answer this pathology multiple-choice question.

Question:
What is the inferred histological diagnosis based on the increased melanocytic activity and pigmentation?

Options:
A) Basal cell carcinoma
B) Acute inflammation
C) Dermal nevus
D) Melanocytic junctional activity

Reasoning guidance:
- Ground the reasoning in visible histology, morphology, staining, or tissue architecture.
- Compare plausible options only when it helps decide the answer.
- Keep the reasoning to at most two sentences and do not invent clinical context.

Output exactly this format:
<reasoning>one or two concise pathology reasoning sentences</reasoning>
<answer>X</answer>

Put exactly one valid option letter inside <answer>; do not include the option text there.
Valid option letters: A, B, C, D.
```

CASE 02



Pathology image

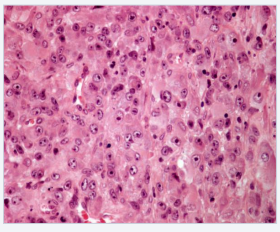
SYSTEM MESSAGE

```
You are a pathology vision-language assistant. Use only the histology image and the provided multiple-choice options. Give
concise, evidence-based pathology reasoning, then one final option letter.
```

USER MESSAGE

```
<image>
Please answer this pathology multiple-choice question.

Question:
What can be inferred about the stromal compartment of this breast tissue sample?

Options:
A) The stromal compartment is poorly demarcated from the ductal structures.
B) There is a significant infiltration of inflammatory cells.
C) The extracellular matrix shows extensive fibrosis and collagen deposition.
D) The connective tissue is well demarcated from the ductal structures with no significant inflammatory infiltrate.

Reasoning guidance:
- Ground the reasoning in visible histology, morphology, staining, or tissue architecture.
- Compare plausible options only when it helps decide the answer.
- Keep the reasoning to at most two sentences and do not invent clinical context.

Output exactly this format:
<reasoning>one or two concise pathology reasoning sentences</reasoning>
<answer>X</answer>

Put exactly one valid option letter inside <answer>; do not include the option text there.
Valid option letters: A, B, C, D.
```

CASE 03



Pathology image

SYSTEM MESSAGE

```
You are a pathology vision-language assistant. Use only the histology image and the provided multiple-choice options. Give
concise, evidence-based pathology reasoning, then one final option letter.
```

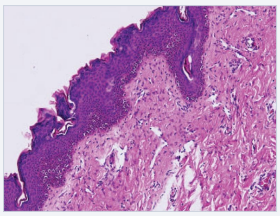
USER MESSAGE

```
<image>
Please answer this pathology multiple-choice question.

Question:
What is the predominant cytoplasmic staining characteristic observed in these cells?

Options:
A) Clear cytoplasm
B) Eosinophilic cytoplasm
C) Basophilic cytoplasm
D) Amphophilic cytoplasm

Reasoning guidance:
- Ground the reasoning in visible histology, morphology, staining, or tissue architecture.
- Compare plausible options only when it helps decide the answer.
- Keep the reasoning to at most two sentences and do not invent clinical context.

Output exactly this format:
<reasoning>one or two concise pathology reasoning sentences</reasoning>
<answer>X</answer>

Put exactly one valid option letter inside <answer>; do not include the option text there.
Valid option letters: A, B, C, D.
```

Figure 4: Concrete example of the standard student prompt used for both OPD student generation and test-time inference. The prompt exposes only the pathology image, question, and answer options, and requests a controlled visible reasoning trace before the final option.

Privileged teacher prompt examples

CASE 01



Pathology image

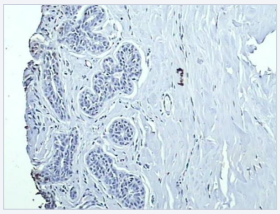
SYSTEM MESSAGE

```
You are a pathology teacher model for on-policy distillation. You may use teacher-only privileged reference information to calibrate your
reasoning, but your final response should read as if it was derived from the histology image and options.
```

USER MESSAGE

```
<image>
Please answer this pathology multiple-choice question.

Question:
What is the inferred histological diagnosis based on the increased melanocytic activity and pigmentation?

Options:
A) Basal cell carcinoma
B) Acute inflammation
C) Dermal nevus
D) Melanocytic junctional activity

Teacher-only privileged reference:
- Correct answer: D) Melanocytic junctional activity
- Reference reasoning: The presence of prominent pigmentation and increased melanocytic activity at the junction of the epidermis and dermis, as
  shown by the darker cells at the bottom layer of the epidermis, suggests melanocytic junctional activity. This is the hallmark for various
  melanocytic lesions and not typical for the other conditions listed.

Use the privileged reference only to know the intended pathology conclusion.
Then reconstruct a concise visual reasoning path as if you did not already know
the answer. The reasoning should be suitable for guiding a smaller student
model. Do not mention the privileged reference or reveal that the answer was
provided to you. Ground the reasoning in visible histology, morphology,
staining, or tissue architecture.

Output exactly this format:
<reasoning>one or two concise pathology reasoning sentences</reasoning>
<answer>D</answer>

Put exactly the correct option letter inside <answer>; do not include the option text there.
```

CASE 02



Pathology image

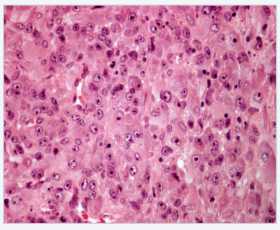
SYSTEM MESSAGE

```
You are a pathology teacher model for on-policy distillation. You may use teacher-only privileged reference information to calibrate your
reasoning, but your final response should read as if it was derived from the histology image and options.
```

USER MESSAGE

```
<image>
Please answer this pathology multiple-choice question.

Question:
What can be inferred about the stromal compartment of this breast tissue sample?

Options:
A) The stromal compartment is poorly demarcated from the ductal structures.
B) There is a significant infiltration of inflammatory cells.
C) The extracellular matrix shows extensive fibrosis and collagen deposition.
D) The connective tissue is well demarcated from the ductal structures with no significant inflammatory infiltrate.

Teacher-only privileged reference:
- Correct answer: D) The connective tissue is well demarcated from the ductal structures with no significant inflammatory infiltrate.
- Reference reasoning: The stromal compartment is well demarcated from the ductal structures, showing a clear boundary between the two components.
  The extracellular matrix, which consists of fibrous connective tissue, provides structural support, embedding the ducts without significant
  inflammatory infiltrate.

Use the privileged reference only to know the intended pathology conclusion.
Then reconstruct a concise visual reasoning path as if you did not already know
the answer. The reasoning should be suitable for guiding a smaller student
model. Do not mention the privileged reference or reveal that the answer was
provided to you. Ground the reasoning in visible histology, morphology,
staining, or tissue architecture.

Output exactly this format:
<reasoning>one or two concise pathology reasoning sentences</reasoning>
<answer>D</answer>

Put exactly the correct option letter inside <answer>; do not include the option text there.
```

CASE 03



Pathology image

SYSTEM MESSAGE

```
You are a pathology teacher model for on-policy distillation. You may use teacher-only privileged reference information to calibrate your
reasoning, but your final response should read as if it was derived from the histology image and options.
```

USER MESSAGE

```
<image>
Please answer this pathology multiple-choice question.

Question:
What is the predominant cytoplasmic staining characteristic observed in these cells?

Options:
A) Clear cytoplasm
B) Eosinophilic cytoplasm
C) Basophilic cytoplasm
D) Amphophilic cytoplasm

Teacher-only privileged reference:
- Correct answer: B) Eosinophilic cytoplasm
- Reference reasoning: The cells in the image show a predominance of eosinophilic cytoplasm which appears pink to red, consistent with the uptake
  of the eosin dye in a hematoxylin and eosin stain. Basophilic cytoplasm would appear blue as it would take up the hematoxylin, amphophilic would
  have both blue and pink characteristics, and clear cytoplasm would have a pale, almost transparent appearance.

Use the privileged reference only to know the intended pathology conclusion.
Then reconstruct a concise visual reasoning path as if you did not already know
the answer. The reasoning should be suitable for guiding a smaller student
model. Do not mention the privileged reference or reveal that the answer was
provided to you. Ground the reasoning in visible histology, morphology,
staining, or tissue architecture.

Output exactly this format:
<reasoning>one or two concise pathology reasoning sentences</reasoning>
<answer>B</answer>

Put exactly the correct option letter inside <answer>; do not include the option text there.
```

Figure 5: Concrete example of the privileged teacher prompt used during OPD training. The teacher receives the same image-question context together with teacher-only reference answer and reasoning fields, while the student prompt remains unchanged.

Table 9: PRM weight-scale and score-to-weight ablation on PathMMU 4B.

Component	Variant	Token signal a_t	λ	PathMMU 4B
Baseline	Privileged OPD	0	0.0	73.99
Weight scale	Sigmoid PRM	$\sigma(r_\psi(y)_t)$	0.5	<u>74.42</u>
Weight scale	Sigmoid PRM (main)	$\sigma(r_\psi(y)_t)$	1.0	**74.96**
Weight scale	Sigmoid PRM	$\sigma(r_\psi(y)_t)$	2.0	73.64
Score mapping	Direct PRM score	$r_\psi(y)_t$	1.0	74.25
Score mapping	Hard PRM mask	$\mathbf{1}\{\sigma(r_\psi(y)_t) > 0.5\}$	1.0	73.02

B.2 PRM ARCHITECTURE ABLATION

This subsection isolates the capacity of the PRM token scorer. All variants use the same frozen Qwen3.5-9B token embedding, PRM supervision data, train/validation split, binary-cross-entropy objective, token prior, span smoother, and threshold 0.5 for precision and recall. We vary only the number of bidirectional Transformer layers and the hidden size of the PRM encoder, and Table 10 reports token-level ROC performance together with the operating-point precision and recall.

Table 10: PRM architecture ablation over encoder depth and hidden size. Precision@0.5 and Recall@0.5 use a fixed threshold of 0.5.

Layers	Hidden size	AUROC	Precision@0.5	Recall@0.5
4	1024	0.873	0.552	0.668
4	2048	0.892	0.579	0.701
8	1024	0.906	0.596	0.721
8	2048	**0.924**	**0.613**	**0.742**

The architecture ablation shows a consistent benefit from increasing PRM capacity. Moving from 4 to 8 Transformer layers improves token ranking performance, and the 8-layer, 2048-hidden model obtains the best AUROC and operating-point precision and recall. We therefore use this configuration as the default PRM in the main PaRa-OPD experiments.

B.3 PRIVILEGED-INFORMATION GRANULARITY

This subsection studies how much privileged information should be exposed to the teacher during OPD on PathMMU; Table 11 reports the comparison. The student input remains the original image-question prompt in all rows; only the teacher-side context changes. In our implementation, full privilege corresponds to providing both the reference answer and the reference reasoning to the teacher. PathMMU's official Tiny test split contains 1,138 usable examples after missing-image filtering, so we report both the number of correct predictions and the corresponding accuracy.

Table 11: Ablation of teacher-side privileged information on PathMMU. Entries are correct predictions out of 1,138 examples, with accuracy in parentheses.

Teacher context	Answer	Reasoning	PathMMU 2B	PathMMU 4B
None (OPD)	No	No	727/1138 (63.88)	762/1138 (66.96)
Answer only	Yes	No	771/1138 (67.75)	806/1138 (70.83)
Answer + reasoning (Privileged OPD)	Yes	Yes	**806/1138 (70.83)**	**842/1138 (73.99)**

Providing the answer alone already substantially improves over standard OPD for both model sizes, indicating that teacher-side access to the target option helps supervise the student's on-policy states.

Adding reference reasoning gives a further gain, suggesting that pathology rationales provide useful intermediate structure beyond the final answer.

C PATHOLOGY REWARD MODEL ANALYSIS

This section provides additional diagnostics for the PRM, including the continuous visual-credit scores used to construct labels, validation ROC behavior, vocabulary-frequency statistics, and examples of both training labels and post-training trajectory scores.

C.1 PRM VALIDATION PERFORMANCE

Table 12 reports the held-out token-level validation result of the trained PRM, and Figure 6 shows the corresponding ROC curve. The model achieves an AUROC of 0.924 for identifying visually dependent or pathology-informative trajectory tokens. At the 0.5 operating threshold, it obtains precision 0.613 and recall 0.742, indicating that the learned scorer has strong ranking performance while retaining sufficient coverage of positive token spans.

C.2 PRM TARGET SPARSITY

The PRM target remains sparse at the token level. In the checked training-label pool, visual-credit labeling marks 1,026,030 of 9,955,712 trajectory tokens as positive, corresponding to 10.31% of all tokens. After adding pathology expert-vocabulary supervision and accounting for overlap with the visual-credit labels, the combined visual-or-expert target covers 1,268,851 tokens, or 12.74% of the pool. Thus, the PRM is trained to emphasize a compact subset of trajectory tokens rather than uniformly increasing the loss on entire responses.

C.3 PRM VISUAL-CREDIT STATISTICS

We further characterize the continuous counterfactual scores used before constructing binary PRM labels. For each sampled teacher trajectory, we teacher-force the same token sequence with and without the image and record $g_t = p_\theta(y_t \mid I, q, y_{<t}) - p_\theta(y_t \mid q, y_{<t})$ for every response token. The analyzed set contains 125,076 trajectories and 9,955,712 labeled tokens from the visually dependent PathMMU and PathGen-Instruct candidate pool.

Figure 7 shows the empirical distribution of these probability differences over the full feasible interval $[-1, 1]$. The scores are concentrated near zero, while the positive tail contains tokens whose likelihood increases substantially when the teacher observes the image. Table 13 reports the corresponding percentiles. The 90th percentile is 0.2083, matching the global top-10% threshold used in the binary visual-dependence labels; the additional per-trajectory top-5 rule recovers locally informative tokens even when an entire trajectory has smaller absolute gains. Figure 8 and Figure 9 further summarize the most frequent visual-credit and expert-vocabulary terms used in the PRM targets.

C.4 PRM TRAINING LABELS AND TRAJECTORY SCORING EXAMPLES

Figure 10 shows examples of the token-level labels used to train the PRM. These labels are constructed from counterfactual visual-credit signals and pathology vocabulary matches, so the supervision marks sparse visual or pathology-specific spans rather than the whole response. Figure 11 shows the trained PRM applied to generated reasoning trajectories from PathMMU and PathGen-Instruct. In this post-training visualization, high-scoring spans tend to correspond to morphology, staining, tissue architecture, disease entities, and anatomic descriptors, while generic connective words and answer-format tokens usually receive low scores. The examples illustrate the intended role of the PRM: it does not replace the VQA model or directly judge image correctness, but supplies a token-level weighting signal over the trajectory for OPD.

Table 12: Token-level validation performance of the trained PRM. Precision and recall are computed at threshold 0.5.

Metric	Value
AUROC	0.924
Precision@0.5	0.613
Recall@0.5	0.742
FPR@0.5	0.054

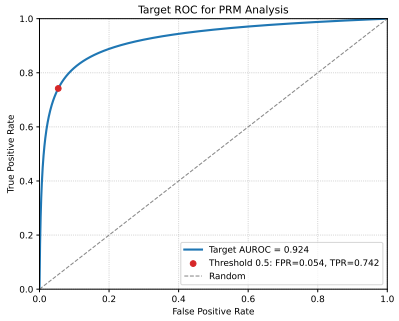


Figure 6: ROC curve of the trained PRM on the held-out token-level validation split. The red marker indicates the threshold 0.5 operating point.

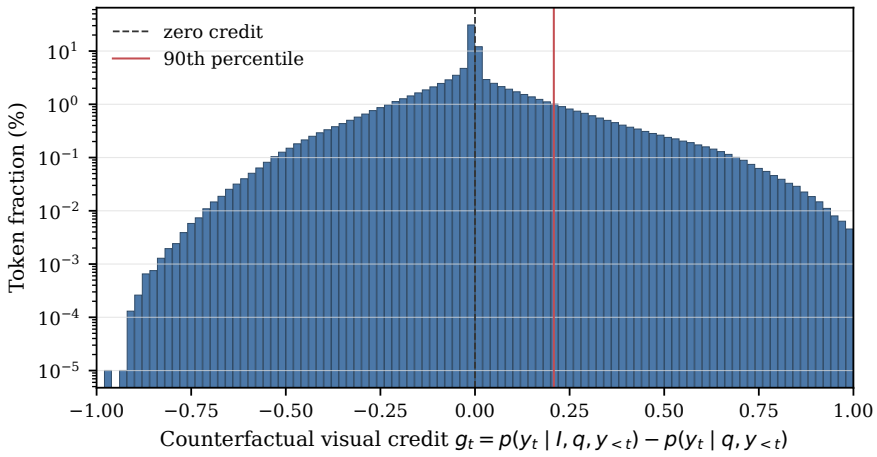


Figure 7: Distribution of token-level counterfactual visual-credit scores in the sampled PRM training trajectories. The x-axis covers the full range $[-1, 1]$, and the y-axis shows the token fraction on a log scale. The red line marks the 90th percentile, corresponding to the global top-10% visual-credit threshold.

Table 13: Percentiles of token-level counterfactual visual-credit scores g_t in the PRM label set.

Percentile	g_t value
0	-0.9728
1	-0.4525
5	-0.2681
10	-0.1736
25	-0.0424
50	-0.0001
75	0.0322
90	0.2083
95	0.3447
99	0.6203
100	1.0000

D QUALITATIVE CASE STUDIES AND FAILURE ANALYSIS

D.1 PATHGEN-INSTRUCT CASE OVERVIEW

To complement the aggregate accuracy results, we conduct a qualitative analysis of six PathGen-Instruct examples. The goal is to examine not only whether each method predicts the correct option, but also how its reasoning trajectory uses visual evidence, handles negative findings, and maps pathology observations to the answer choices. Each example shows the original pathology image and question together with sampled or saved trajectories from the original Qwen3.5-2B base model and Qwen3.5-2B student checkpoints after SFT, OPD, Privileged OPD, Med-OPD, GRPO, and PaRa-OPD training. For OPD-family methods, Qwen3.5-9B is used only as the training-time teacher; the displayed trajectories are generated by the trained 2B students at inference time.

The six examples are organized to cover both representative improvements and a remaining limitation. Figures 12–16 are cases where PaRa-OPD answers correctly while several baselines fail. They cover tissue-processing artifact recognition, absence of a glomerulus, malignant morphology recognition, anatomic layer localization around colonic crypts, and selection of the most specific absent lung structure. Figure 17 presents a failure case after PaRa-OPD training: the question asks for the feature most indicative of cholangiocarcinoma, where the correct answer depends on recognizing tubular structures lined by cuboidal to columnar epithelium rather than generic nuclear atypia.

Case 1: processing artifact versus biological lumen. Figure 12 asks what clear spaces in the tissue section represent. The base model, SFT, and OPD incorrectly interpret the round spaces as ductal, lobular, or glandular lumina, mainly relying on their shape and surrounding stroma without verifying epithelial or endothelial lining. Privileged OPD, Med-OPD, GRPO, and PaRa-OPD answer correctly. The PaRa-OPD trajectory is more focused: it notes that the spaces lack cellular detail and contrast with surrounding tissue that retains nuclear structure, supporting tissue-processing artifact rather than a true anatomic lumen.

Case 2: absence of a glomerulus. Figure 13 asks whether a glomerulus is clearly visible. The base model and GRPO hallucinate glomerular capillary loops or filtration structures from local tubular profiles. SFT, OPD, Privileged OPD, Med-OPD, and PaRa-OPD answer correctly, although SFT reaches the correct answer with an incorrect tissue description. The PaRa-OPD trajectory identifies renal tubules and uses negative visual evidence, namely the absence of clustered capillary loops, mesangial cells, and a glomerular tuft, to reject the glomerulus option.

Case 3: malignant tissue morphology. Figure 14 requires selecting the feature set that best characterizes the section. The base model, OPD, and GRPO overemphasize eosinophilic amorphous regions and disorganized tissue, choosing extensive necrosis and debris. GRPO even mentions enlarged, irregular, hyperchromatic nuclei but maps these observations to the wrong final option. SFT, Privileged OPD, Med-OPD, and PaRa-OPD correctly identify marked nuclear atypia and abortive gland formation; PaRa-OPD additionally rejects necrosis and inflammation as the dominant findings.

Case 4: anatomic layer around colonic crypts. Figure 15 asks which layer surrounds the colonic crypts. The base model describes the lamina propria in its reasoning but outputs the submucosa, showing a reasoning-to-answer mapping error. Privileged OPD, Med-OPD, and GRPO confuse the cellular connective tissue adjacent to crypts with deeper smooth muscle or submucosal layers. SFT, OPD, and PaRa-OPD answer correctly; PaRa-OPD grounds the choice in the immediate spatial relationship between crypts and loose cellular connective tissue and excludes the deeper submucosa.

Case 5: option granularity in lung pathology. Figure 16 asks which structure is absent in the observed lung section. Several baselines correctly describe dense eosinophilic necrotic material and disrupted architecture, and some even mention loss of alveolar spaces, but still choose the broader option "normal lung parenchyma." GRPO and PaRa-OPD select the more specific correct answer: distinct alveolar spaces are absent. The PaRa-OPD trajectory directly connects the image finding to normal lung architecture, where thin alveolar septa and open alveolar spaces should be visible but are not.

Across the five success cases, the base model fails all examples, while SFT, OPD, Privileged OPD, Med-OPD, and GRPO show partial improvements but still suffer from hallucinated structures, over-broad option choices, and inconsistent reasoning-to-answer mapping. PaRa-OPD is correct on all five examples and tends to align the final answer with discriminative visual evidence rather than with the most salient or generic pathology phrase.

Remaining failure case. Figure 17 shows a representative failure after PaRa-OPD training. The question asks which feature is most indicative of cholangiocarcinoma in liver tissue. The correct answer is tubular structures lined by cuboidal to columnar epithelium, which is more diagnosis-specific than generic nuclear atypia. The base model, SFT, OPD, Privileged OPD, GRPO, and PaRa-OPD all select enlarged hyperchromatic nuclei with increased nuclear-to-cytoplasmic ratios. Med-OPD is the only method in this example that selects the correct tubular-epithelial feature. This failure suggests that PRM-based token reweighting can still overemphasize salient malignancy-associated descriptors when the correct option requires a more specific architectural or lineage-defining cue.

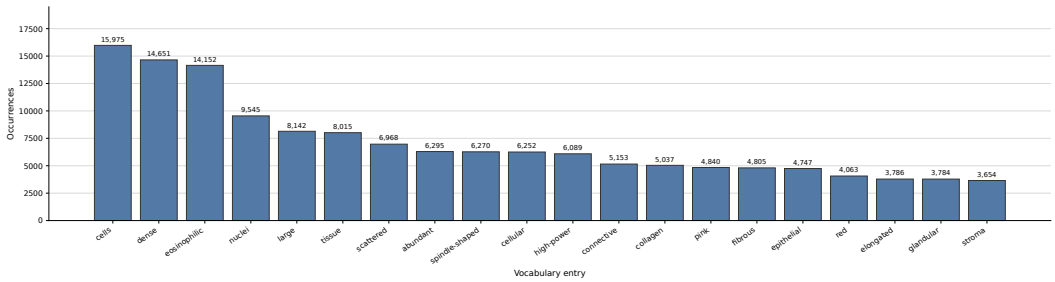


Figure 8: Top-20 word occurrences among tokens marked by the visual-credit labels. Formatting tokens, punctuation, and common function words are excluded for readability.

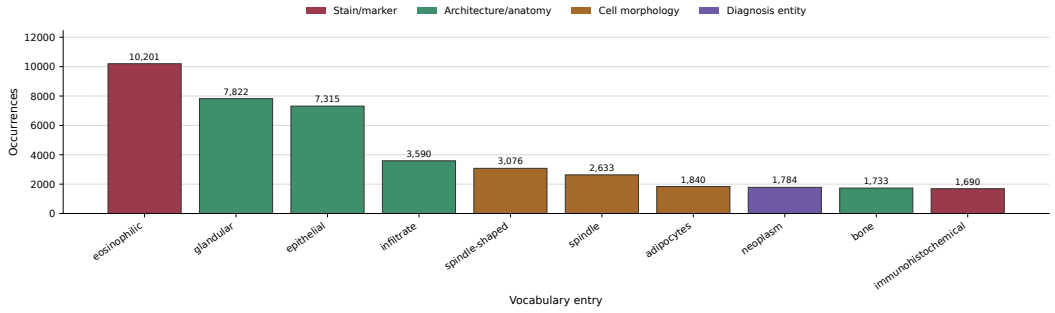


Figure 9: Top-10 pathology expert-vocabulary entries by occurrence after filtering generic descriptors. Colors indicate the expert category assigned during vocabulary construction.

From PubMed

<reasoning>The inset reveals a high nuclear-to-cytoplasmic ratio with nuclei that exhibit a "salt and pepper" appearance, where the chromatin is finely granular and evenly dispersed without the coarse clumping or prominent nucleoli typically seen in high-grade carcinomas or lymphomas. This specific fine, stippled texture is characteristic of neuroendocrine differentiation.</reasoning>
<answer>A</answer>

From SocialPath

<reasoning>The image displays thick, interwoven bundles of eosinophilic collagen fibers with fibroblast nuclei in between, which is the characteristic architecture of dense irregular connective tissue. This structure provides tensile strength in multiple directions and lacks the loose matrix of areolar tissue, lipid vacuoles of adipose tissue, or chondrocytes in lacunae found in cartilage.</reasoning>
<answer>C</answer>

From EduContent

<reasoning>The bottom right corner of the image shows a dense aggregation of small cells with dark-stained nuclei, which is characteristic of a lymphoid infiltrate seen in chronic inflammation. This morphology contrasts with the surrounding loose connective tissue and does not resemble the organized structure of normal dermis, muscle, or malignant cell clusters.</reasoning>
<answer>D</answer>

From PathCLS

<reasoning>The image displays large, empty-appearing circular structures with thin rims, which is the classic histological appearance of adipocytes where lipid droplets are dissolved during processing. This morphology is distinct from the cellular features of lymphocytes, smooth muscle, or epithelial tissues, and lacks the fibroblastic density of cancer-associated stroma.</reasoning>
<answer>A</answer>

From Atlas

<reasoning>The image shows a renal tubule containing a dense aggregate of small, dark-staining lymphocytes within its lumen, which is not characteristic of typical casts or necrotic debris. This cellular infiltration suggests an inflammatory process rather than a degenerative or proteinaceous change.
<answer>A</answer>

From PathGen-Instruct

<reasoning>The image shows a cluster of cells with segmented, dark-staining nuclei characteristic of neutrophils, which are the hallmark cellular component of acute inflammation. The surrounding tissue appears relatively preserved without significant fibrosis or lymphocytic infiltration typical of chronic inflammation.</reasoning>
<answer>A</answer>

From PubMed

<reasoning>The image displays a large area of pale, amorphous, eosinophilic material consistent with necrotic debris, labeled as 'd'. This is surrounded by a dense infiltrate of inflammatory cells, which is characteristic of the exudative and cellular response seen in acute inflammation. The presence of significant necrosis alongside an acute cellular infiltrate rules out purely chronic, fibrotic, or serous processes.</reasoning>
<answer>C</answer>

From SocialPath

<reasoning>
The image displays chorionic villi separated by wide, open spaces that are filled with maternal blood (indicated by the red blood cells), which is characteristic of the intervillous space. The histiocytes are located within these spaces between the villous structures, rather than being confined to the subchorionic space or clustered around the periphery of the villi.
</reasoning>
<answer>D</answer>

From EduContent

<reasoning>The image displays a cross-section of a squamous epithelial structure with distinct, large, eosinophilic keratin-filled cysts (horn cysts) embedded within the epithelium. This specific morphology, characterized by islands of proliferating keratinocytes surrounded by a rim of basal cells, is the hallmark histologic definition of a benign cystic formation such as an epidermoid or trichilemmal cyst. The presence of these well-defined cystic spaces rules out solid neoplastic growth, dysplasia, or granulomatous inflammation.</reasoning>
<answer>A</answer>

From PathCLS

<reasoning>The image displays dense, eosinophilic fibrous tissue with scattered inflammatory cells, which is the characteristic appearance of desmoplastic stroma associated with invasive carcinoma. This reactive fibrotic tissue contrasts with the glandular structures of normal mucosa or the cellular density of the tumor epithelium itself.</reasoning>
<answer>H</answer>

Figure 10: Examples of labeled data used to train the PRM. Highlighted tokens indicate positive supervision targets constructed from counterfactual visual-credit labels and pathology vocabulary matches.

From PubMed

<reasoning>The inset reveals a high nuclear-to-cytoplasmic ratio with nuclei that exhibit a "salt and pepper" appearance, where the chromatin is finely granular and evenly dispersed without the coarse clumping or prominent nucleoli typically seen in high-grade carcinomas or lymphomas. This specific fine, stippled texture is characteristic of neuroendocrine differentiation.</reasoning>
<answer>A</answer>

From SocialPath

<reasoning>The image displays thick, interwoven bundles of eosinophilic collagen fibers with fibroblast nuclei in between, which is the characteristic architecture of dense irregular connective tissue. This structure provides tensile strength in multiple directions and lacks the loose matrix of areolar tissue, lipid vacuoles of adipose tissue, or chondrocytes in lacunae found in cartilage.</reasoning>
<answer>C</answer>

From EduContent

<reasoning>The bottom right corner of the image shows a dense aggregation of small cells with dark-stained nuclei, which is characteristic of a lymphoid infiltrate seen in chronic inflammation. This morphology contrasts with the surrounding loose connective tissue and does not resemble the organized structure of normal dermis, muscle, or malignant cell clusters.</reasoning>
<answer>D</answer>

From PathCLS

<reasoning>The image displays large, empty-appearing circular structures with thin rims, which is the classic histological appearance of adipocytes where lipid droplets are dissolved during processing. This morphology is distinct from the cellular features of lymphocytes, smooth muscle, or epithelial tissues, and lacks the fibroblastic density of cancer-associated stroma.</reasoning>
<answer>A</answer>

From Atlas

<reasoning>The image shows a renal tubule containing a dense aggregate of small, dark-staining lymphocytes within its lumen, which is not characteristic of typical casts or necrotic debris. This cellular infiltration suggests an inflammatory process rather than a degenerative or proteinaceous change.
<answer>A</answer>

From PathGen-Instruct

<reasoning>The image shows a cluster of cells with segmented, dark-staining nuclei characteristic of neutrophils, which are the hallmark cellular component of acute inflammation. The surrounding tissue appears relatively preserved without significant fibrosis or lymphocytic infiltration typical of chronic inflammation.</reasoning>
<answer>A</answer>

From PubMed

<reasoning>The image displays a large area of pale, amorphous, eosinophilic material consistent with necrotic debris, labeled as 'd'. This is surrounded by a dense infiltrate of inflammatory cells, which is characteristic of the exudative and cellular response seen in acute inflammation. The presence of significant necrosis alongside an acute cellular infiltrate rules out purely chronic, fibrotic, or serous processes.</reasoning>
<answer>C</answer>

From SocialPath

<reasoning>
The image displays chorionic villi separated by wide, open spaces that are filled with maternal blood (indicated by the red blood cells), which is characteristic of the intervillous space. The histiocytes are located within these spaces between the villous structures, rather than being confined to the subchorionic space or clustered around the periphery of the villi.
</reasoning>
<answer>D</answer>

From EduContent

<reasoning>The image displays a cross-section of a squamous epithelial structure with distinct, large, eosinophilic keratin-filled cysts (horn cysts) embedded within the epithelium. This specific morphology, characterized by islands of proliferating keratinocytes surrounded by a rim of basal cells, is the hallmark histologic definition of a benign cystic formation such as an epidermoid or trichilemmal cyst. The presence of these well-defined cystic spaces rules out solid neoplastic growth, dysplasia, or granulomatous inflammation.</reasoning>
<answer>A</answer>

From PathCLS

<reasoning>The image displays dense, eosinophilic fibrous tissue with scattered inflammatory cells, which is the characteristic appearance of desmoplastic stroma associated with invasive carcinoma. This reactive fibrotic tissue contrasts with the glandular structures of normal mucosa or the cellular density of the tumor epithelium itself.</reasoning>
<answer>H</answer>

Figure 11: Examples of PRM scoring on generated reasoning trajectories from PathMMU and PathGen-Instruct. Highlight intensity indicates the predicted token-level score used for trajectory weighting.

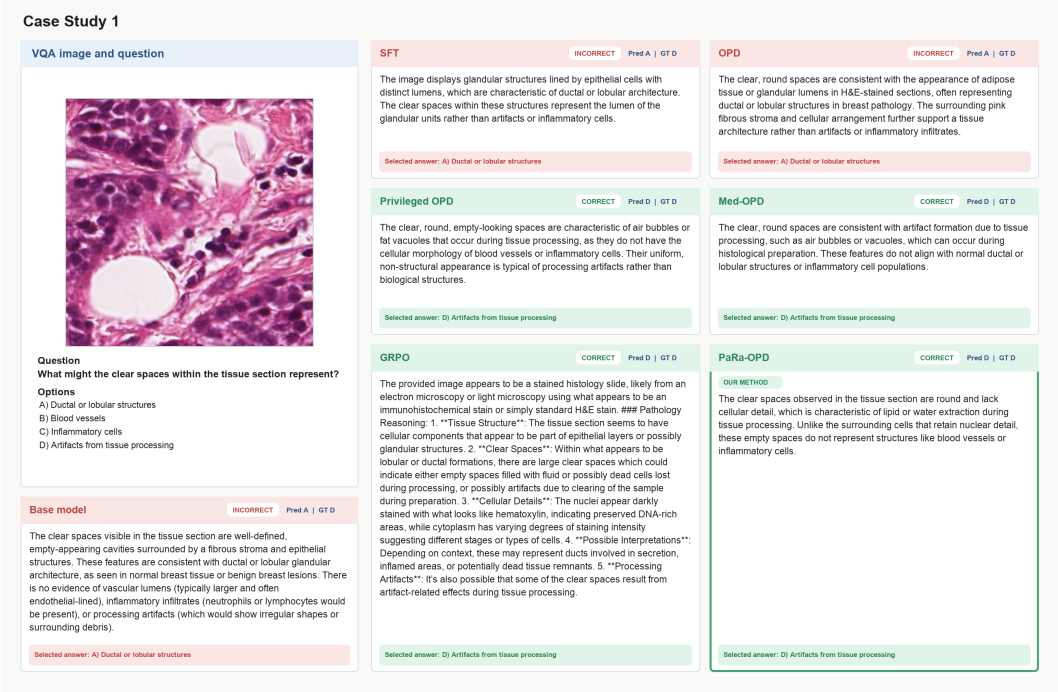


Figure 12: Qualitative case study 1: tissue-processing artifact recognition. The question asks whether clear spaces represent biological structures or preparation artifacts; PaRa-OPD focuses on the lack of cellular lining/detail and selects the artifact option.

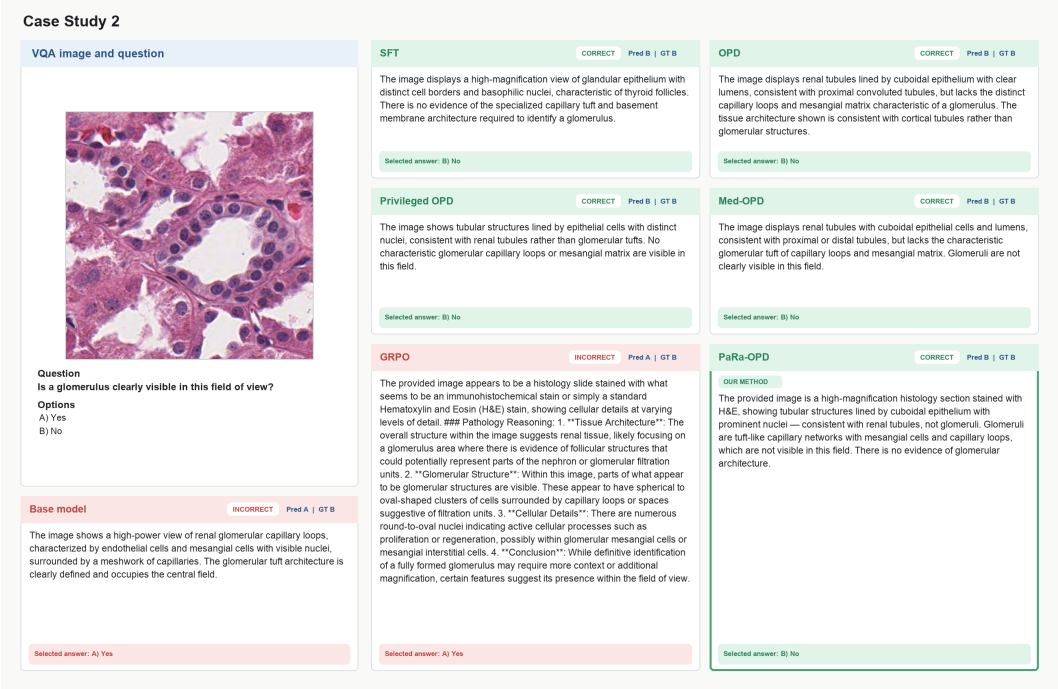


Figure 13: Qualitative case study 2: ruling out an absent glomerulus. The key evidence is negative visual evidence: renal tubules are visible, but glomerular capillary loops and mesangial architecture are not.

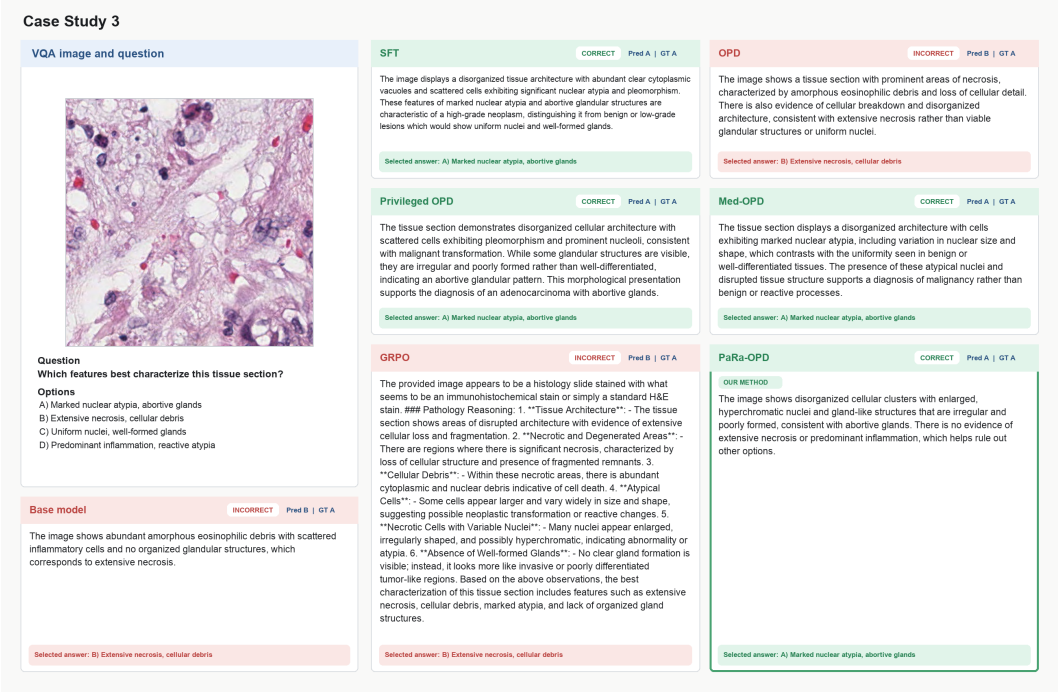


Figure 14: Qualitative case study 3: malignant morphology recognition. The case tests whether the model prioritizes nuclear atypia and abortive glands over more generic necrosis or debris descriptions.

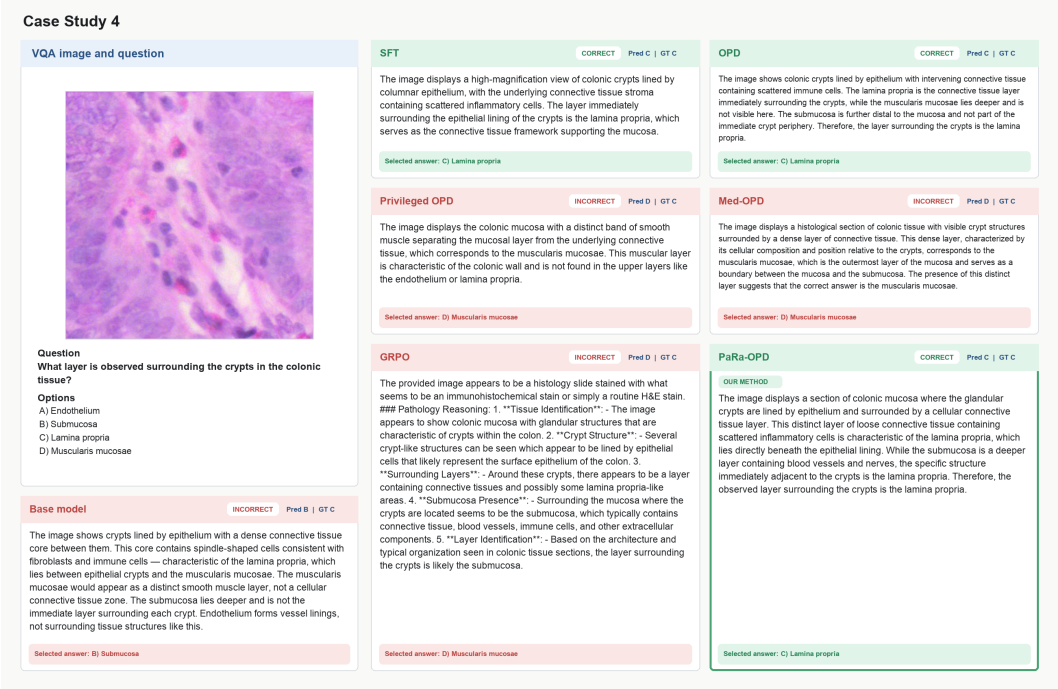


Figure 15: Qualitative case study 4: anatomic layer localization around colonic crypts. The correct answer depends on distinguishing lamina propria immediately adjacent to crypts from deeper submucosa or muscularis mucosae.

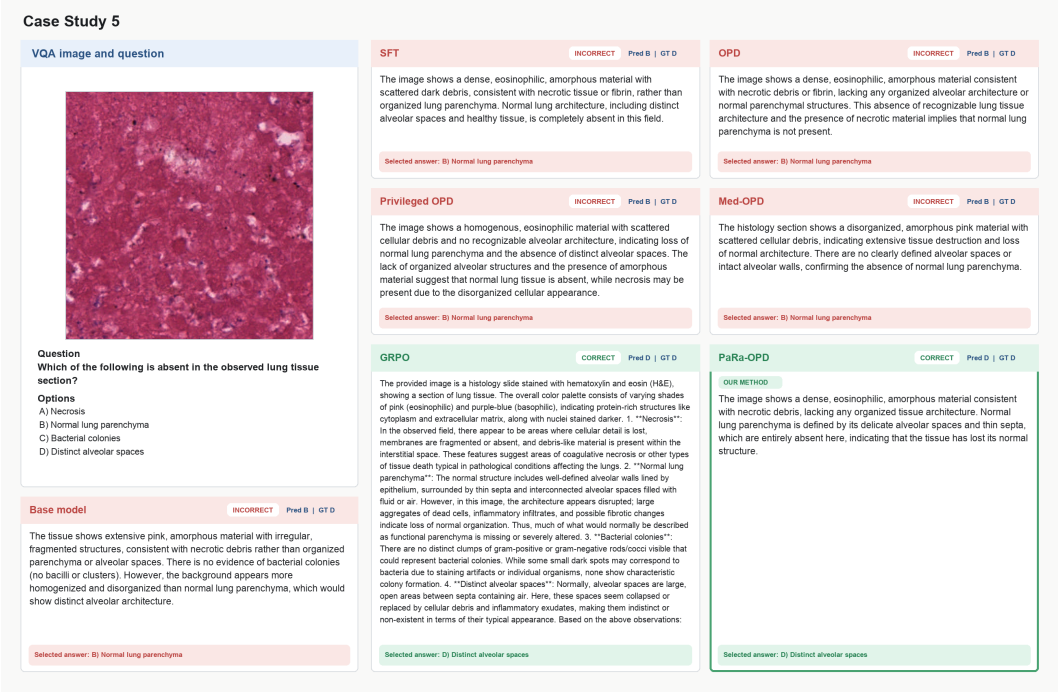


Figure 16: Qualitative case study 5: selecting the most specific absent lung structure. Several baselines describe loss of normal architecture but choose an overbroad option, whereas PaRa-OPD selects the absent alveolar spaces.

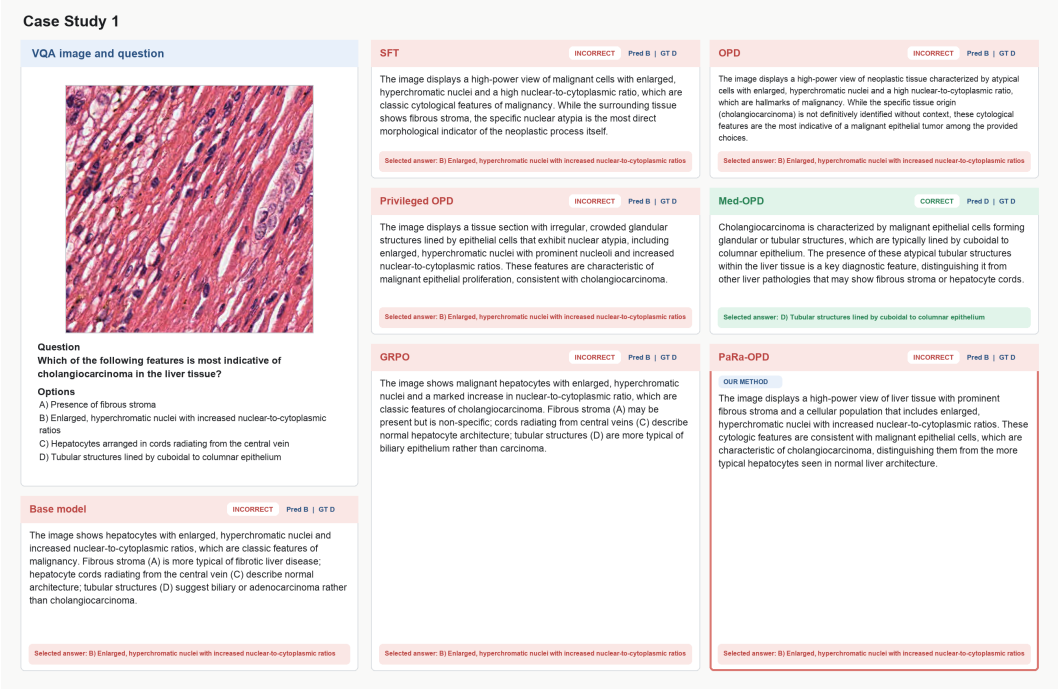


Figure 17: PaRa-OPD failure case: cholangiocarcinoma feature selection. Most methods, including PaRa-OPD, choose a generic malignant nuclear feature, while the correct answer depends on recognizing tubular structures lined by cuboidal to columnar epithelium.

E LIMITATIONS

Our experiments are currently conducted within the Qwen3.5 model family, using Qwen3.5-2B and Qwen3.5-4B students with a Qwen3.5-9B teacher. This controlled setting enables consistent comparisons across OPD variants, but it does not fully characterize how PaRa-OPD transfers across different VLM backbones. Future work will extend the distillation study to additional model families and teacher–student pairings.

The current PRM is also intentionally lightweight. While this design keeps token-level reward estimation efficient during OPD training, larger reward models or alternative scorer architectures may provide stronger token discrimination and better generalization across datasets. Exploring the scaling behavior and architectural choices of PRMs remains an important direction for future work.