A GENERATIVE PATHOLOGY VISION-LANGUAGE MODEL WITH NATIVE-PRESERVING EXPERT INJECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pathology vision–language reasoning requires fine-grained morphological perception together with flexible multimodal reasoning and generation. General-purpose vision–language models (VLMs) provide broad reasoning capabilities, but their visual representations may miss pathology-specific evidence. Pathology foundation models (PFMs) capture detailed histopathological features but typically lack open-ended reasoning and generation. We introduce PathTIDE (Pathology Token Injection of Domain Experts), an architecture that retains the native VLM pathway while adding a pathology foundation model as a Pathology Expert. PathTIDE aligns Pathology Expert representations with a language-aware feature space and injects spatially structured expert tokens through gated residual fusion. We evaluate PathTIDE on closed-ended pathology VQA, including multiple-choice and Yes/No questions. Across four closed-ended VQA evaluation sets—PathMMU, SlideBench-VQA-TCGA, SlideBench-VQA-CPTAC and the closed-ended portion of WSI-Bench—PathTIDE (PFM-Full) achieves accuracies of 0.869, 0.793, 0.812 and 0.869, respectively. Architecture ablations show that preserving the native General Expert and using gated expert injection outperforms both General Expert replacement and ungated residual addition. These results support adding pathology-specific evidence to the native visual branch of a general-purpose VLM.

1 INTRODUCTION

Histopathological interpretation depends on subtle morphological patterns, from local cellular detail to tissue architecture, and must ultimately be expressed in diagnostic language. A generative pathology vision–language model therefore needs pathology-sensitive visual representations together with language-conditioned reasoning and generation (Srinidhi et al., 2021; Lu et al., 2024b).

General-purpose vision–language models (VLMs) offer broad reasoning and generation capabilities for medical applications. However, pathology remains challenging for general-purpose visual encoders because diagnostic cues are subtle, spatially distributed and morphologically complex (Srinidhi et al., 2021; Chen et al., 2024; Lu et al., 2024a). In parallel, pathology foundation models (PFMs) trained on large-scale pathology datasets have become effective at extracting domain-specific representations of cellular detail, tissue architecture and spatial organization (Chen et al., 2024; Lu et al., 2024a; Xiang et al., 2025). Yet PFMs generally do not provide open-ended, language-conditioned reasoning and generation. The two model families are therefore complementary: VLMs provide flexible reasoning and generation, whereas PFMs provide detailed pathology evidence. Despite this complementarity, few existing approaches retain a general-purpose VLM's native visual pathway while adding a separately encoded pathology specialist.

Existing generative pathology systems can be grouped into four broad architectural routes, although individual models may combine elements from multiple routes. CoCa-style systems train a new pathology-specific visual–language model from paired data (Yu et al., 2022; Lu et al., 2024a; Ding et al., 2025). Native-VLM adaptation retains the original visual–language interface but relies on the native vision encoder alone (Seyfioglu et al., 2024; Zhang et al., 2025; Xu et al., 2025). CLIP-based construction connects a pathology-specific CLIP-style visual encoder to a text-only language model (Sun et al., 2024b; Dai et al., 2024), whereas PFM-based construction adopts a pretrained pathology foundation model as the visual encoder and learns a new interface to a text-only language model (Lu et al., 2024b; Shaikovski et al., 2024; Chen et al., 2025b; Liang et al., 2025). Across these routes,

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.

methods either rebuild multimodal coupling, rely only on the native visual pathway, or replace it with a pathology encoder. This leaves open a central architectural question: how can a pathology foundation model serve as a complementary expert for a native VLM, enhancing pathology-specific perception while preserving the VLM's native visual–language pathway and general reasoning and generation capabilities?

To address this problem, we introduce PathTIDE (Pathology Token Injection of Domain Experts), an architecture that adds a pathology specialist to a pretrained VLM while retaining its native visual pathway. The VLM's native vision pathway is retained as the General Expert Branch, while a separately encoded pathology model forms the Pathology Expert Branch. PathTIDE aligns and projects Pathology Expert representations into expert tokens, retrieves pathology evidence through cross-attention using General Expert tokens as queries, and injects the resulting update through learned token-wise gated residual fusion. Because the gated residual update preserves the native visual-token dimensionality and sequence layout, the original visual–language interface can be reused without reconstruction or replacement.

Our contributions are threefold:

1. We introduce PathTIDE, a PFM-agnostic architecture that augments a pretrained VLM with a complementary pathology expert, preserving the native visual pathway while adding pathology-specific evidence to multimodal reasoning.
2. We develop a gated residual injection mechanism that aligns and projects Pathology Expert representations, retrieves them with General Expert queries through cross-attention, and adaptively injects the resulting update without altering the native visual-token sequence shape.
3. We conduct a cross-benchmark evaluation of PathTIDE on four closed-ended pathology VQA settings—PathMMU, SlideBench-VQA-TCGA, SlideBench-VQA-CPTAC and the closed-ended portion of WSI-Bench—against general-purpose, medical, pathology-specialized and WSI-aware VLMs, with matched ablations that isolate the effects of native-pathway preservation and gated fusion.

2 Related Work

2.1 Pathology foundation models

Pathology foundation models (PFMs) learn pathology-specific visual representations from large pathology image collections and image–text pairs. Representative models include UNI and CONCH (Chen et al., 2024; Lu et al., 2024a), Virchow and Prov-GigaPath (Vorontsov et al., 2024; Xu et al., 2024), and MUSK (Xiang et al., 2025). They provide strong features for classification, retrieval and tissue-level analysis. Yet most PFMs function primarily as visual encoders; even language-aligned PFMs are not natively coupled to the interface of a generative VLM. They therefore supply specialist pathology evidence without the open-ended language-conditioned reasoning of such models.

2.2 Generative vision–language models for pathology

Generative pathology VLMs can be grouped into four broad routes: CLIP-based construction, PFM-based construction, CoCa-style vision–language pretraining and native-VLM adaptation. Figure 7 summarizes these routes, while Figure 1 focuses on the two retrofit routes that connect pathology visual components to text-only LLMs. CLIP- and PFM-based systems use learned interfaces to retrofit text-only LLMs with pathology visual components, whereas CoCa-style systems rebuild visual–language coupling and native-VLM adaptation retains the native pathway without an independent pathology encoder.

CLIP-based multimodal construction. CLIP-based construction connects a pathology-specific CLIP-style visual encoder to a pretrained text-only LLM through a learned interface. PathAsst combines PathCLIP with Vicuna-13B, whereas PA-LLaVA connects PLIP to Llama 3. CPath-Omni constructs CPath-CLIP before coupling it with Qwen2.5-14B (Sun et al., 2024b; Dai et al., 2024; Sun et al., 2025a). These systems retain pathology-specific visual representation learning, but must learn new visual–language coupling.

PFM-based multimodal construction. PFM-based construction uses a pretrained PFM as the visual encoder for a text-only LLM and learns an adaptor, projector or other connector between them. SlideChat and WSI-LLaVA use CONCH or Prov-GigaPath with slide-level connectors; ALPaCA and PathChat connect CONCH or UNI through learned alignment modules, while PRISM and PRISM2 use Virchow-family encoders with Perceiver-based interfaces (Chen et al., 2025b; Liang

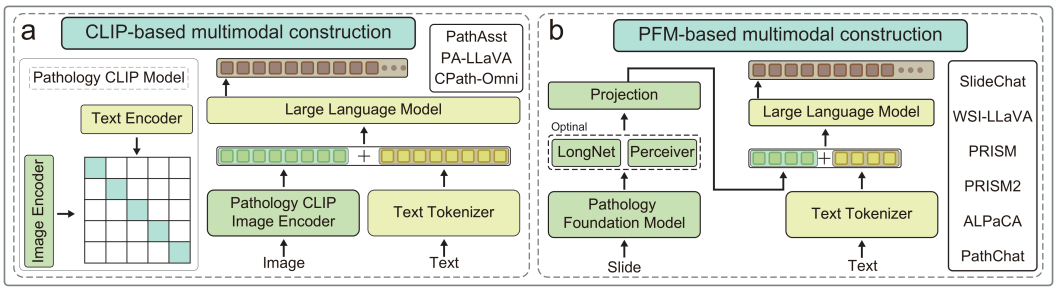


Figure 1: **Existing CLIP- and PFM-based retrofit designs for pathology-specific visual expertise.** Both routes retrofit a pretrained text-only language model by connecting it to pathology visual representations through a learned interface: **(a)** CLIP-based construction trains a pathology CLIP encoder, whereas **(b)** PFM-based construction adapts a pretrained pathology foundation model.

et al., 2025; Gao et al., 2026; Lu et al., 2024b; Shaikovski et al., 2024; Vorontsov et al., 2026). This route supplies strong pathology features, but its interface to the language model is constructed externally rather than inherited from a pretrained VLM.

CoCa-style vision–language pretraining. CoCa-style pretraining learns a multimodal model with contrastive alignment and autoregressive generation objectives (Yu et al., 2022). Pathology models such as CONCH and TITAN connect pretrained unimodal components through cross-attention and then optimize their visual–language coupling on pathology image–text pairs (Lu et al., 2024a; Ding et al., 2025). This route builds a new pathology-specific multimodal system, but requires substantial paired data and reconstructs visual–language coupling.

Native-VLM adaptation. Native-VLM adaptation retains the pretrained topology and visual–language interface while applying domain training. LLaVA-Med and Quilt-LLaVA adapt the LLaVA family through biomedical or pathology alignment and instruction tuning, whereas Patho-R1 and SmartPath-R1 instead adapt Qwen2.5-VL using continued pretraining, supervised fine-tuning, reinforcement learning or LoRA (Li et al., 2023; Seyfioglu et al., 2024; Zhang et al., 2025; Xu et al., 2025). This route preserves native multimodal coupling, but the native vision encoder remains its only source of visual evidence; it does not introduce an independent PFM.

Together, these routes expose an architectural trade-off: existing systems either rebuild visual–language coupling or retain it without an independent pathology encoder. PathTIDE instead retains the General Expert Branch and native visual–language interface while adding a separate, interchangeable Pathology Expert Branch.

3 Method

3.1 Model architecture

The complete PathTIDE architecture and training flow are summarized in Figure 2. PathTIDE couples a pretrained VLM with a pathology foundation model through two expert branches: the retained native pathway forms the General Expert, while the pathology model forms the Pathology Expert. Gated residual fusion injects pathology evidence before language decoding. In the experiments, the pathology branch uses MUSK.

For a source image x, let $x_{\mathrm{N}} = \mathcal{R}_{\mathrm{N}}(x)$ denote the model-compatible native view and let $\mathcal{P}(x)$ denote its retained tissue-patch set. The model computes

$$\begin{aligned} T_{\mathrm{N}} &= E_{\mathrm{N}}(x_{\mathrm{N}}), \qquad T_{\mathrm{P}} = \Pi(F_{\mathrm{M}}(\mathcal{P}(x))), \\ T_{\mathrm{F}} &= \mathcal{I}(T_{\mathrm{N}}, T_{\mathrm{P}}), \qquad y \sim p_\theta(y \mid q, T_{\mathrm{F}}). \end{aligned} \tag{1}$$

Here, E_{N}, F_{M}, Π and $\mathcal{I}$ denote the General Expert, Pathology Expert, shared projector and fusion module, respectively; the experiments use Qwen3.5-9B and MUSK for the two branches.

General Expert Branch. The General Expert Branch encodes the native view as

$$T_{\mathrm{N}} = E_{\mathrm{N}}(x_{\mathrm{N}}) \in \mathbb{R}^{N \times d}, \qquad d = 4096, \tag{2}$$

where N is the number of native visual tokens. This $N \times d$ sequence retains the visual interface expected by the language backbone.

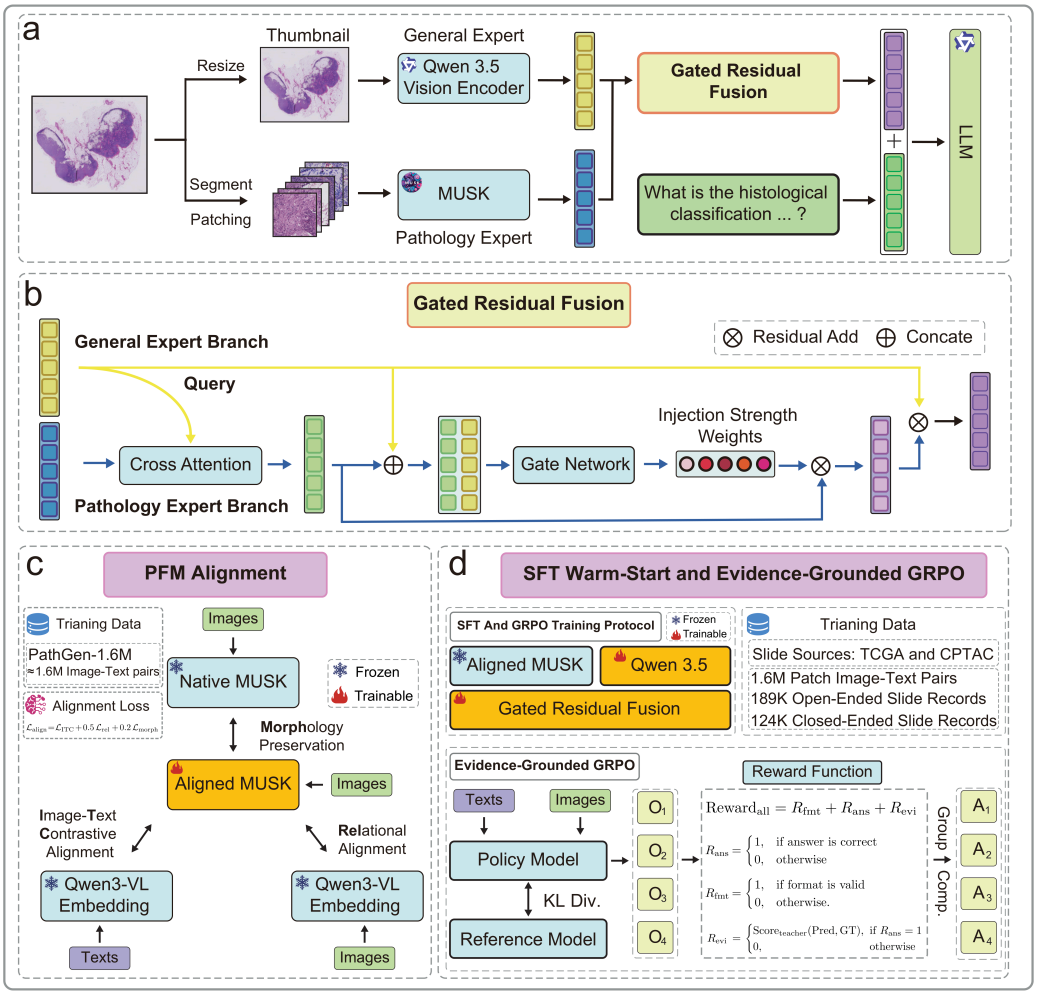


Figure 2: **Overview of PathTIDE. (a)** Native General Expert and pathology Pathology Expert branches with gated residual fusion. **(b)** Cross-attention and token-wise gating. **(c)** MUSK alignment to the frozen Qwen embedding space. **(d)** Stage 2 SFT and evidence-grounded GRPO.

Pathology Expert Branch. After Stage 1 alignment, the frozen Pathology Expert encodes $\mathcal{P}(x) = \{p_i\}_{i=1}^{M}$ as

$$C_{\mathrm{M}} = \left[F_{\mathrm{M}}^{\mathrm{CLS}}(p_i)\right]_{i=1}^{M} \in \mathbb{R}^{M\times 1024}. \tag{3}$$

A shared LayerNorm–dropout–linear projector maps these CLS features to the native hidden dimension:

$$t_{\mathrm{P},i} = \mathrm{Linear}_{\mathrm{P}}(\mathrm{Dropout}(\mathrm{LN}(c_{\mathrm{M},i}))) + e_{\mathrm{pos},i}. \tag{4}$$

The resulting sequence is denoted by $T_{\mathrm{P}} \in \mathbb{R}^{M\times d}$. Here, $e_{\mathrm{pos},i}$ is a fixed two-dimensional sine–cosine encoding of normalized patch coordinates, or zero when coordinates are unavailable. The expert-token contract is designed to permit substitution of other PFMs, but the present experiments evaluate only MUSK.

Gated residual fusion. A gated cross-attention block injects Pathology Expert evidence into the General Expert sequence through a $d_b = 1024$ attention bottleneck (Alayrac et al., 2022):

$$\begin{aligned} \Delta T &= \mathrm{CrossAttn}(T_{\mathrm{N}}W_Q, T_{\mathrm{P}}W_K, T_{\mathrm{P}}W_V)\, W_O, \\ \alpha &= \sigma(G(T_{\mathrm{N}} \,\|\, \Delta T)), \qquad \alpha \in \mathbb{R}^{N\times 1}, \\ T_{\mathrm{F}} &= \mathrm{LN}(T_{\mathrm{N}} + \alpha \odot \Delta T). \end{aligned} \tag{5}$$

The bottleneck uses $d_b = 1024$, with W_O restoring $d = 4096$; $\|$ denotes feature-wise concatenation and α scales each General Expert token. Thus, T_{F} preserves the native $N \times d$ token layout while adding Pathology Expert evidence residually. Coordinate processing and caching are detailed in Appendix B.2.

3.2 Training pipeline

Training first aligns the Pathology Expert to a language-aware embedding space, then optimizes PathTIDE through supervised fine-tuning (SFT) and group-relative policy optimization (GRPO).

Stage 1: MUSK alignment. Stage 1 uses paired PathGen-1.6M patches and descriptions to align LoRA-adapted MUSK with a frozen Qwen3-VL-Embedding-2B teacher (Sun et al., 2025b). For mini-batch $\{(x_i, t_i)\}_{i=1}^{B}$, a two-layer head maps each 1,024-dimensional MUSK CLS feature to the teacher dimension:

$$c_i = F_{\mathrm{M}}^{\mathrm{CLS,LoRA}}(x_i) \in \mathbb{R}^{1024}, \qquad z_i = \mathrm{norm}(A_{\mathrm{M}}(c_i)) \in \mathbb{R}^{2048}. \tag{6}$$

The teacher independently produces normalized image and text embeddings q_i^{img} and q_i^{txt}.

Image–text contrastive alignment. With cosine similarity $s(a, b) = a^\top b$ and temperature $\tau = 0.07$, the symmetric in-batch contrastive loss is

$$\mathcal{L}_{\mathrm{ITC}} := -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(s(z_i, q_i^{\mathrm{txt}})/\tau)}{\sum_{j=1}^{B} \exp(s(z_i, q_j^{\mathrm{txt}})/\tau)} + \log \frac{\exp(s(q_i^{\mathrm{txt}}, z_i)/\tau)}{\sum_{j=1}^{B} \exp(s(q_i^{\mathrm{txt}}, z_j)/\tau)} \right]. \tag{7}$$

The two terms use other batch elements as negatives for image-to-text and text-to-image retrieval.

Relational alignment. To align batch-level geometry, define the row-stacked matrices

$$Z = [z_1^\top; \ldots; z_B^\top], \qquad Q^{\mathrm{img}} = [(q_1^{\mathrm{img}})^\top; \ldots; (q_B^{\mathrm{img}})^\top], \qquad C = [c_1^\top; \ldots; c_B^\top]. \tag{8}$$

For row-wise normalized $\overline{V}$, define $\mathcal{R}(V) = \overline{V}\,\overline{V}^\top$. We match the geometry of aligned MUSK features to that of Qwen image embeddings:

$$\mathcal{L}_{\mathrm{rel}} := \frac{1}{B^2} \left\| \mathcal{R}(Z) - \mathcal{R}(Q^{\mathrm{img}}) \right\|_F^2. \tag{9}$$

Morphology preservation. We disable the MUSK LoRA adapters to obtain frozen reference features

$$C^{\mathrm{ref}} = [(c_1^{\mathrm{ref}})^\top; \ldots; (c_B^{\mathrm{ref}})^\top], \qquad c_i^{\mathrm{ref}} = F_{\mathrm{M}}^{\mathrm{CLS}}(x_i). \tag{10}$$

The morphology-preservation loss constrains changes to the original MUSK geometry:

$$\mathcal{L}_{\mathrm{morph}} := \frac{1}{B^2} \left\| \mathcal{R}(C) - \mathcal{R}(C^{\mathrm{ref}}) \right\|_F^2. \tag{11}$$

The complete Stage 1 objective is

$$\mathcal{L}_{\mathrm{align}} := \mathcal{L}_{\mathrm{ITC}} + 0.5\mathcal{L}_{\mathrm{rel}} + 0.2\mathcal{L}_{\mathrm{morph}}. \tag{12}$$

Only the MUSK LoRA parameters and temporary alignment head are optimized. We then discard the head and freeze aligned MUSK to construct the expert feature bank. Appendix B.1 reports the implementation settings and training dynamics.

Stage 2: SFT warm-start and evidence-grounded GRPO. Aligned MUSK remains frozen within the PathTIDE injection interface. SFT on instruction–response data precedes GRPO (Shao et al., 2024) on answer-checkable multiple-choice, true/false, classification and normalized short-answer tasks. Open-ended descriptions and reports are excluded. For each input x, the policy samples $G = 4$ rollouts in the training format `<RATIONALE>...</RATIONALE>` followed by `<FINAL_ANSWER>...</FINAL_ANSWER>`; benchmark evaluation uses the separate answer-only contract in Appendix B.6.

Deterministic answer and format rewards. The parser requires exactly one such block pair and extracts the answer from `<FINAL_ANSWER>...</FINAL_ANSWER>`; short answers receive task-specific normalization. The format reward is

$$R_{\mathrm{fmt},g} = \begin{cases} 1, & \text{if the response passes the format check,} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The answer reward is

$$R_{\mathrm{ans},g} = \begin{cases} 1, & \text{if the response is well formed and matches the gold answer,} \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Both rewards are deterministic and independent of the evidence teacher.

Signed evidence-grounding reward. GPT-5.6-Luna evaluates evidence only for well-formed rollouts with correct answers. Given the visual input, question, available choices, verified reference answer and response, it scores whether visible pathology supports the response and penalizes unsupported findings. The reference answer constrains answer correctness but is not treated as visual evidence. It returns

$$R_{\mathrm{evi},g} \in [-1, 1], \tag{15}$$

If either deterministic check fails, $R_{\mathrm{evi},g} = 0$ and teacher evaluation is skipped. Teacher outputs determine rewards only and never serve as supervised targets. The exact evidence-judge instructions and output contract are reproduced in Appendix B.6.

Composite reward and GRPO update. The three components have equal unit weight:

$$R_g := R_{\mathrm{ans},g} + R_{\mathrm{fmt},g} + R_{\mathrm{evi},g}. \tag{16}$$

Unsupported evidence can therefore reduce the reward of a correct answer. Within-group normalization produces $\widehat{A}_g$; identical group rewards set all advantages to zero and skip the update. A frozen post-SFT checkpoint serves as the KL reference policy π_{ref}. With

$$\rho_g(\theta) = \exp(\log \pi_\theta(y_g \mid x) - \log \pi_{\mathrm{old}}(y_g \mid x)), \qquad \epsilon = 0.2, \tag{17}$$

Writing $\mathrm{clip}_\epsilon(\rho_g)$ for clipping to $[1-\epsilon, 1+\epsilon]$, the GRPO objective is

$$\mathcal{L}_{\mathrm{GRPO}} := -\tfrac{1}{G} \sum_{g=1}^{G} \min\Big\{\rho_g \widehat{A}_g, \mathrm{clip}_\epsilon(\rho_g)\widehat{A}_g\Big\} + \beta D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}}), \qquad \epsilon = 0.2,\ \beta = 0.02. \tag{18}$$

Here Eq. 18 is the core policy–KL objective. The implementation adds the injection-change and gate regularizers reported in Appendix B.5; Appendix Eq. 28 is therefore the total loss optimized in the full training route.

Alternating PFM-Full and PFM-Free training. Stage 2 follows a deterministic 10-batch cycle with nine PFM-Full batches and one PFM-Free batch. PFM-Full activates the Pathology Expert Branch and updates the injection modules together with LoRA adapters in the General Expert Branch and language backbone. PFM-Free bypasses the complete gated residual fusion module, including the pathology projector, cross-attention, gate and injection LayerNorm, and forwards the retained native tokens directly to the language backbone:

$$T_{\mathrm{free}} = T_{\mathrm{N}}. \tag{19}$$

The same bypassed route is used for PFM-Free training batches and PFM-Free inference. It is therefore not equivalent to setting $\alpha = 0$ in Eq. 5, which would still apply the injection LayerNorm to T_{N}. At inference, PathTIDE (PFM-Free) bypasses the complete gated residual fusion module, whereas PathTIDE (PFM-Full) activates it. Both modes use the same final checkpoint, so their comparison measures the benefit of the complete expert-enabled fusion route rather than the pathology residual update alone. The complete schedule is given in Appendix B.4.

4 Experiments and Results

4.1 Experimental setup

Training data. Stage 1 aligns MUSK on the PathGen-1.6M image–text corpus (Sun et al., 2025b). Stage 2 combines supervised fine-tuning (SFT) and GRPO using PathGen-1.6M, SlideInstruct, and the WSI-Bench training release (Chen et al., 2025b; Liang et al., 2025); slide-level pools are decontaminated against the evaluation slides. Exact corpus composition and decontamination statistics are reported in Appendix A.

Evaluation benchmarks. We evaluate frozen representation quality on seven patch- and ROI-level datasets and pathology reasoning on PathMMU, SlideBench-VQA-TCGA, SlideBench-VQA-CPTAC and the closed-ended portion of WSI-Bench. Full split, category and question-count summaries are provided in Supplementary Table 1.

Baselines and comparison protocol. Baselines cover open-weight and proprietary general-purpose VLMs, medical and pathology-specialized models, and WSI-aware systems; the complete roster is shown in Table 1. Conventional image-interface models receive the benchmark image or an aspect-ratio-preserving WSI thumbnail used by the General Expert Branch, whereas WSI-specific systems

Table 1: **Overall accuracy across pathology benchmarks.** Expert performance† is available only for PathMMU test-tiny ($N = 1{,}138$) and is not directly comparable to the combined PathMMU scores. Dashes indicate unavailable results. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	PathMMU (N=9,583)	SlideBench-VQA-TCGA (N=3,176)	SlideBench-VQA-CPTAC (N=240)	WSI-Bench (N=1,336)
LLaVA-Med-v1.5-7B	0.283 (0.274, 0.292)	0.325 (0.309, 0.341)	0.188 (0.142, 0.242)	0.279 (0.255, 0.305)
MedGemma-1.5-4B-it	0.495 (0.485, 0.505)	0.388 (0.371, 0.405)	0.246 (0.192, 0.300)	0.580 (0.555, 0.609)
Patho-R1-7B	0.625 (0.616, 0.635)	–	–	–
GLM-4.6V-Flash	0.568 (0.559, 0.578)	0.492 (0.475, 0.509)	0.312 (0.258, 0.375)	0.674 (0.651, 0.702)
GLM-4.6V	0.617 (0.607, 0.627)	0.534 (0.517, 0.552)	0.396 (0.338, 0.458)	0.682 (0.658, 0.710)
Qwen3.5-2B	0.585 (0.576, 0.595)	0.364 (0.347, 0.380)	0.179 (0.133, 0.229)	0.585 (0.561, 0.613)
Qwen3.5-4B	0.675 (0.666, 0.685)	0.483 (0.465, 0.500)	0.267 (0.217, 0.325)	0.657 (0.632, 0.685)
Qwen3.5-9B	0.702 (0.693, 0.711)	0.504 (0.486, 0.521)	0.308 (0.250, 0.367)	0.660 (0.635, 0.688)
Qwen3.7-Flash	*0.730 (0.721, 0.738)*	0.558 (0.541, 0.576)	0.421 (0.358, 0.483)	0.707 (0.683, 0.732)
Gemini-3.5-Flash-Lite	0.689 (0.679, 0.699)	0.521 (0.504, 0.540)	0.417 (0.354, 0.479)	0.652 (0.627, 0.680)
GPT-5.6-Luna	0.653 (0.643, 0.662)	0.558 (0.540, 0.576)	0.417 (0.354, 0.479)	0.730 (0.707, 0.755)
GPT-5.6-Luna-Pro	0.663 (0.653, 0.674)	0.575 (0.558, 0.593)	0.383 (0.321, 0.446)	0.722 (0.698, 0.747)
PathChat+	0.728 (0.719, 0.736)	–	–	–
Expert† (test-tiny)	0.718	–	–	–
PRISM2	–	0.556 (0.539, 0.574)	0.554 (0.492, 0.617)	0.688 (0.664, 0.712)
WSI-LLaVA	–	0.629 (0.614, 0.646)	0.571 (0.512, 0.629)	**0.882 (0.865, 0.899)**
SlideChat	–	*0.685 (0.668, 0.701)*	*0.600 (0.537, 0.662)*	0.733 (0.710, 0.758)
PathTIDE (PFM-Free)	0.800 (0.792, 0.808)	0.695 (0.678, 0.710)	0.671 (0.613, 0.729)	*0.824 (0.803, 0.843)*
PathTIDE (PFM-Full)	**0.869 (0.862, 0.875)**	**0.793 (0.778, 0.807)**	**0.812 (0.758, 0.863)**	0.869 (0.851, 0.887)

Table 2: **Ablation of pathology adaptation and expert integration.** *+PT* denotes matched pathology post-training. *Replace* bypasses the General Expert Branch and conditions the language backbone on projected Pathology Expert tokens. *Add* retains both branches but uses ungated residual addition after cross-attention alignment. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Setting	PathMMU	SlideBench-VQA-TCGA	SlideBench-VQA-CPTAC	WSI-Bench
Qwen3.5-9B	0.702 (0.693, 0.711)	0.504 (0.486, 0.521)	0.308 (0.250, 0.367)	0.660 (0.635, 0.688)
Qwen3.5-9B+PT	*0.755 (0.745, 0.763)*	*0.609 (0.591, 0.625)*	*0.508 (0.446, 0.571)*	*0.754 (0.730, 0.777)*
Qwen3.5-9B(Replace)+PT	0.718 (0.708, 0.727)	0.548 (0.530, 0.565)	0.408 (0.346, 0.471)	0.700 (0.674, 0.721)
Qwen3.5-9B(Add)+PT	0.781 (0.772, 0.789)	0.661 (0.643, 0.676)	0.604 (0.542, 0.667)	0.800 (0.778, 0.820)
PathTIDE (PFM-Full)	**0.869 (0.862, 0.875)**	**0.793 (0.778, 0.807)**	**0.812 (0.758, 0.863)**	**0.869 (0.851, 0.887)**

retain their native slide-processing pipelines while using the same questions and answer-scoring procedure.

Metrics and evaluation scope. We report accuracy for closed-ended pathology reasoning and frozen-representation classification, with 95% confidence intervals estimated from 1,000 nonparametric bootstrap resamples unless otherwise stated. Generative comparisons are restricted to multiple-choice and Yes/No questions for deterministic answer parsing and consistent scoring; open-ended description and report-generation tasks are outside the present scope. Prompt templates and model-specific inference settings are provided in Appendix B.6.

4.2 PathTIDE-Full achieves the highest accuracy on PathMMU

On the 9,583-question combined PathMMU evaluation, PathTIDE (PFM-Free) achieved 0.800 accuracy (95% CI, 0.792–0.808), whereas PathTIDE (PFM-Full) achieved 0.869 (0.862–0.875; Table 1). The strongest external result was 0.730 from Qwen3.7-Flash, followed by 0.728 from PathChat+ (Chen et al., 2025a). Relative to Qwen3.7-Flash, PathTIDE (PFM-Free) and PathTIDE (PFM-Full)

Table 3: **Stage-wise training and evidence-reward ablation.** Nested progression from the Qwen3.5-9B base through SFT, GRPO without the evidence reward, and Full GRPO; all variants use the same backbone and matched evaluation protocols.

Setting	PathMMU	SlideBench-VQA-TCGA	SlideBench-VQA-CPTAC	WSI-Bench
Qwen3.5-9B	0.702 (0.693, 0.711)	0.504 (0.486, 0.521)	0.308 (0.250, 0.367)	0.660 (0.635, 0.688)
PathTIDE (SFT)	0.778 (0.770, 0.787)	0.650 (0.633, 0.666)	0.600 (0.538, 0.662)	0.790 (0.768, 0.812)
PathTIDE (SFT + GRPO; no evidence reward)	0.834 (0.827, 0.842)	0.738 (0.723, 0.753)	0.750 (0.695, 0.805)	0.842 (0.823, 0.862)
PathTIDE (SFT + GRPO; full reward)	0.869 (0.862, 0.875)	0.793 (0.778, 0.807)	0.812 (0.758, 0.863)	0.869 (0.851, 0.887)

gained 7.0 and 13.9 percentage points, respectively. PathTIDE (PFM-Full) produced the highest point estimate across all five PathMMU source subsets, ranging from 0.846 on PubMed to 0.905 on PathCLS. On test-tiny, PathTIDE (PFM-Free) and PathTIDE (PFM-Full) reached 0.830 and 0.895, respectively. The reported expert accuracy was measured under a different Atlas protocol and is included only as a contextual reference. Split- and source-specific results are reported in Supplementary Tables 2–7.

4.3 PathTIDE leads on SlideBench and ranks second on WSI-Bench

On SlideBench-VQA-TCGA, SlideChat was the strongest external comparator at 0.685 overall accuracy. PathTIDE (PFM-Free) and PathTIDE (PFM-Full) reached 0.695 (95% CI, 0.678–0.710) and 0.793 (0.778–0.807), respectively, corresponding to gains of 1.0 and 10.8 percentage points. PathTIDE (PFM-Full) also led Microscopy, Diagnosis, and Clinical questions, with accuracy ranging from 0.760 to 0.856. Detailed results appear in Supplementary Tables 8– 12.

On SlideBench-VQA-CPTAC, PathTIDE (PFM-Free) and PathTIDE (PFM-Full) reached 0.671 and 0.812, versus 0.600 for SlideChat; gains were 7.1 and 21.2 percentage points, respectively. PathTIDE (PFM-Full) led the LSCC, LUAD and UCEC cohorts and ranked second on CM (Supplementary Table 13).

On the 1,336 closed-ended WSI-Bench questions, PathTIDE (PFM-Free) achieved 0.824 overall accuracy (95% CI, 0.803–0.843), whereas PathTIDE (PFM-Full) achieved 0.869 (0.851–0.887). PathTIDE (PFM-Full) ranked second overall, 1.3 percentage points behind WSI-LLaVA and above all other comparators, while achieving the highest Diagnosis accuracy at 0.830. Across the three slide-level benchmarks, activating the Pathology Expert Branch improved the shared-checkpoint PFM-Free mode by 9.8, 14.1 and 4.5 percentage points, respectively. Category-level results are reported in Supplementary Tables 14 and 15.

4.4 Ablations of native-preserving pathology expert integration

Under matched pathology post-training, Table 2 compares pathology adaptation and expert integration. Replacing the General Expert Branch with projected Pathology Expert tokens reduced accuracy by 6.3–19.6 percentage points relative to retaining both branches with ungated addition. Learned token-wise gating in PathTIDE further improved accuracy over Add+PT by 6.9–20.8 percentage points. Variant definitions are provided in Appendix B.3, with category-level results in Supplementary Tables 16–19.

4.5 Stage-wise training and evidence-reward ablation

Table 3 isolates cumulative training gains rather than independent comparisons. Accuracy rises from the base model to SFT, improves further with GRPO without the evidence reward, and increases again under Full GRPO on all four benchmarks. All variants use the same backbone, rollout, optimizer and evaluation protocols.

Figure 3 summarizes the reward dynamics: panels (a) and (b) show mean binary format and answer rewards, while panel (c) reports normalized evidence reward over eligible rollouts, $\mathcal{E}_t = \{i : \text{format valid}_i = 1 \wedge \text{answer correct}_i = 1\}$. The evidence reward rises from approximately 0.62 at the SFT checkpoint to 0.77 under Full GRPO. Full GRPO provides only a small additional format-validity gain because format compliance is already near ceiling without the evidence reward. In

contrast, the answer reward improves more visibly, suggesting that the evidence signal mainly complements answer correctness once basic output formatting is reliable.

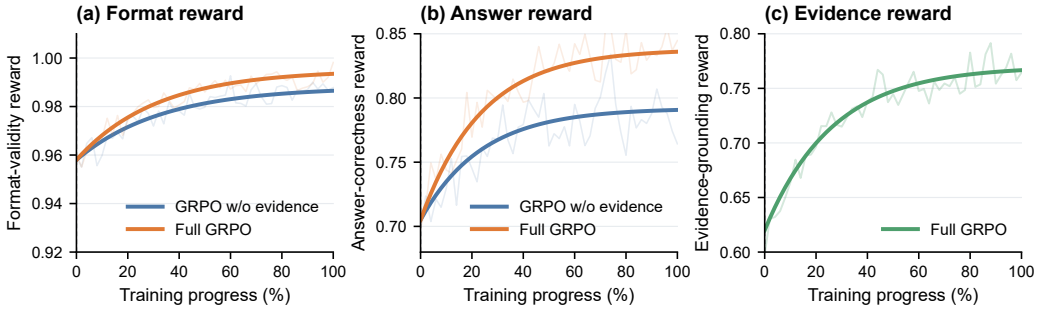


Figure 3: **GRPO reward analysis.** Format-validity and answer-correctness rewards are compared between Full GRPO and GRPO without the evidence reward. Evidence-grounding reward is shown for Full GRPO because it is disabled in the ablated run; panel (c) averages it over eligible rollouts only.

4.6 Expert injection improves frozen pathology representations

For MUSK+Qwen3.5-9B, we extracted and mean-pooled the fused visual-token sequence T_{F} before the language backbone, then applied ℓ_2 normalization to match the native Qwen representation protocol. MUSK+Qwen3.5-9B achieved the highest point estimate on seven of eight patch- and ROI-level benchmarks and tied MUSK on LC25000-Colon (Figure 4); no benchmark showed a lower point estimate than the original MUSK representation. The largest gains over MUSK occurred on SICAPv2, from 0.805 to 0.883, and MHIST, from 0.846 to 0.916. The other five non-tied gains ranged from 0.001 to 0.021. Detailed accuracies and 95% confidence intervals are reported in Supplementary Tables 20 and 21.

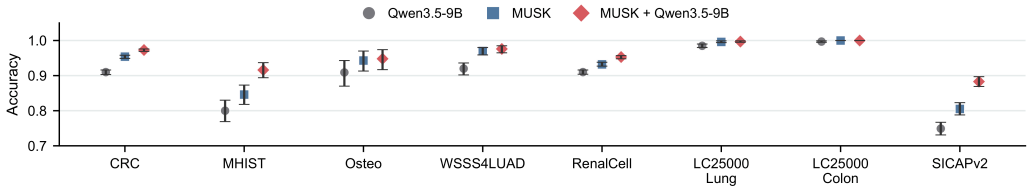


Figure 4: **Frozen pathology representation linear-probe accuracy.** Points show accuracy and error bars show 95% confidence intervals for the native Qwen3.5-9B representation, frozen MUSK representation and their fused representation across eight benchmarks.

4.7 Qualitative case studies

We include two challenging SlideBench cases to illustrate complementary advantages in reasoning coherence and sensitivity to fine-grained pathology cues; all three displayed general-purpose baselines failed both cases, whereas PathTIDE correctly staged a TCGA bladder tumor by recognizing macroscopic perivesical invasion and identified an amelanotic CPTAC melanoma by linking atypical melanocytic cells with the near absence of visible melanin. Detailed case-level comparisons and complete model responses are provided in the Appendix (Figure 5).

5 Limitations and Conclusion

PathTIDE adds pathology-specific evidence to a pretrained VLM while preserving its native visual pathway. Across four closed-ended pathology benchmarks, full expert injection improved accuracy, and ablations favored retaining both expert branches with learned gating. Limitations include related cohorts, tissue-token cost, correlated judge errors, and closed-ended evaluation. Independent cohorts, additional PFMs, pathologist review, and prospective studies remain important next steps.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. doi: 10.52202/068431-1723.

Harish Babu Arunachalam, Rashika Mishra, Ovidiu Daescu, Kevin Cederberg, Dinesh Rakheja, Anita Sengupta, David Leonard, Rami Hallac, and Patrick Leavey. Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *PloS one*, 14(4):e0210706, 2019. doi: 10.1371/journal.pone.0210706.

Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019. doi: 10.48550/arXiv.1912.12142.

Otso Brummer, Petri Pölönen, Satu Mustjoki, and Oscar Brück. Computational textural mapping harmonises sampling variation and reveals multidimensional histopathological fingerprints. *British Journal of Cancer*, 129(4):683–695, 2023. doi: 10.1038/s41416-023-02329-4.

Chengkuan Chen, Luca L Weishaupt, Drew FK Williamson, Richard J Chen, Tong Ding, Bowen Chen, Anurag Vaidya, Long Phi Le, Guillaume Jaume, Ming Y Lu, and Faisal Mahmood. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. *arXiv preprint arXiv:2506.20964*, 2025a. doi: 10.48550/arXiv.2506.20964.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5134–5143. IEEE, 2025b. doi: 10.1109/CVPR52734.2025.00484.

Dawei Dai, Yuanhui Zhang, Long Xu, Qianlan Yang, Xiaojing Shen, Shuyin Xia, and Guoyin Wang. Pa-llava: A large language-vision assistant for human pathology image understanding. In *2024 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pp. 3138–3143. IEEE, 2024. doi: 10.1109/BIBM62325.2024.10821785.

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. doi: 10.1038/s41591-025-03982-3.

Zeyu Gao, Kai He, Weiheng Su, Xiaobo Pang, Ines P. Machado, Mercedes Jimenez-Linan, Brian Rous, Chunbao Wang, Chengzu Li, William McGough, Shangqi Gao, Di Zhang, Tieliang Gong, Ming Y. Lu, Faisal Mahmood, Mengling Feng, Chen Li, and Mireia Crispin-Ortuzar. Alpaca: Adapting llama for pathology context analysis to enable slide-level question answering. *Nature Communications*, 2026. doi: 10.1038/s41467-026-76372-z.

Chu Han, Xipeng Pan, Lixu Yan, Huan Lin, Bingbing Li, Su Yao, Shanshan Lv, Zhenwei Shi, Jinhai Mai, Jiatai Lin, Bingchao Zhao, Zeyan Xu, Zhizhen Wang, Yumeng Wang, Yuan Zhang, Huihui Wang, Chao Zhu, Chunhui Lin, Lijian Mao, Min Wu, Luwen Duan, Jingsong Zhu, Dong Hu, Zijie Fang, Yang Chen, Yongbing Zhang, Yi Li, Yiwen Zou, Yiduo Yu, Xiaomeng Li, Haiming Li, Yanfen Cui, Guoqiang Han, Yan Xu, Jun Xu, Huihua Yang, Chunming Li, Zhenbing Liu, Cheng Lu, Xin Chen, Changhong Liang, Qingling Zhang, and Zaiyi Liu. Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. *arXiv preprint arXiv:2204.06455*, 2022. doi: 10.48550/arXiv.2204.06455.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo, 2018.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564, 2023. doi: 10.52202/075280-1240.

Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, and Linlin Shen. Wsi-llava: A multimodal large language model for whole slide image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22718–22727, 2025. doi: 10.1109/ICCV51701.2025.02109.

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. doi: 10.1038/s41591-024-02856-4.

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. doi: 10.1038/s41586-024-07618-3.

Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13183–13192, 2024. doi: 10.1109/CVPR52733.2024.01252.

George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024. doi: 10.48550/arXiv.2405.10254.

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. doi: 10.48550/arXiv.2402.03300.

Julio Silva-Rodríguez, Adrián Colomer, María A Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine*, 195: 105637, 2020. doi: 10.1016/j.cmpb.2020.105637.

Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021. doi: 10.1016/j.media.2020.101813.

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*, volume 15120 of *Lecture Notes in Computer Science*, pp. 56–73. Springer, 2024a. doi: 10.1007/978-3-031-73033-7_4.

Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5): 5034–5042, 2024b. doi: 10.1609/aaai.v38i5.28308.

Yuxuan Sun, Yixuan Si, Chenglu Zhu, Xuan Gong, Kai Zhang, Pingyi Chen, Ye Zhang, Zhongyi Shui, Tao Lin, and Lin Yang. Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10360–10371, 2025a. doi: 10.1109/CVPR52734.2025.00969.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Jingxiong Li, Xuan Gong, Xinheng Lyu, Tao Lin, and Lin Yang. PathGen-1.6M: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=rFpZnn11gj.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, 2024. doi: 10.1038/s41591-024-03141-0.

Eugene Vorontsov, George Shaikovski, Adam Casson, Julian Viret, Eric Zimmermann, Neil Tenenholtz, Yi Kan Wang, Jan H. Bernhard, Ran A. Godrich, Juan A. Retamero, Jinru Shia, Mithat Gonen, Martin R. Weiser, David S. Klimstra, Razik Yousfi, Nicolò Fusi, Thomas J. Fuchs, Kristen Severson, and Siqi Liu. End-to-end multimodal pathology foundation model with clinical dialogue. *Nature Medicine*, 2026. doi: 10.1038/s41591-026-04521-4.

Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *Artificial Intelligence in Medicine*, volume 12721 of *Lecture Notes in Computer Science*, pp. 11–24. Springer, 2021. doi: 10.1007/978-3-030-77211-6_2.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025. doi: 10.1038/s41586-024-08378-w.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. doi: 10.1038/s41586-024-07441-w.

Zhe Xu, Ziyi Liu, Junlin Hou, Jiabo Ma, Cheng Jin, Yihui Wang, Zhixuan Chen, Zhengyu Zhang, Fuxiang Huang, Zhengrui Guo, Fengtao Zhou, Yingxue Xu, Xi Wang, Ronald Cheong Kin Chan, Li Liang, and Hao Chen. A versatile pathology co-pilot via reasoning enhanced multimodal large language model. *arXiv preprint arXiv:2507.17303*, 2025. doi: 10.48550/arXiv.2507.17303.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Wenchuan Zhang, Penghao Zhang, Jingru Guo, Tao Cheng, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner. *arXiv preprint arXiv:2505.11404*, 2025. doi: 10.48550/arXiv.2505.11404.

APPENDIX

AI ASSISTANCE

We used generative AI tools to assist with manuscript editing and polishing, title refinement, LaTeX formatting, and figure presentation. All AI-assisted changes were reviewed and verified by the authors. The authors take full responsibility for the final manuscript, including all claims, methods, experiments, results, and artifacts.

A REPRODUCIBILITY STATEMENTS

A.1 DATA AVAILABILITY

The datasets used in this study are available from their cited releases under source-specific licenses. Training data comprise PathGen-1.6M (Sun et al., 2025b), SlideInstruct (Chen et al., 2025b) and the WSI-Bench training release (Liang et al., 2025). Linear-probe evaluation uses CRC (Kather et al., 2018), MHIST (Wei et al., 2021), Osteo (Arunachalam et al., 2019), WSSS4LUAD (Han et al., 2022), RenalCell (Brummer et al., 2023), LC25000 (Borkowski et al., 2019) and SICAPv2 (Silva-Rodríguez et al., 2020). Reasoning evaluation uses PathMMU (Sun et al., 2024a), SlideBench-VQA (Chen et al., 2025b) and WSI-Bench (Liang et al., 2025). Supplementary Table 1 summarizes the evaluated sets and task counts.

A.2 TRAINING-SET COMPOSITION AND SLIDE-LEVEL DECONTAMINATION

The SlideInstruct release contains 4,181 caption records and 175,734 VQA records over 4,181 slides. The local author-version WSI-LLaVA training release contains 177,173 records over 9,836 slides. We normalized slide identifiers and formed an exclusion manifest from the union of all SlideBench-VQA-TCGA, SlideBench-VQA-CPTAC and WSI-Bench test slides, comprising 2,169 unique slides. Before filtering, the WSI-LLaVA data shared 1,191 slides with SlideBench-VQA-TCGA and 616 with WSI-Bench, while SlideInstruct shared 306 with WSI-Bench. We removed every training record whose normalized slide identifier matched the exclusion manifest.

The retained slide-level SFT pool contains 188,963 records over 8,811 slides, including 185,088 open-ended VQA and 3,875 caption instances. The RL pool contains 124,036 closed-ended VQA records over 8,812 slides, comprising 123,424 multiple-choice and 612 Yes/No questions. Because SlideInstruct lacks explicit open/closed labels, questions with ordered A–D options enter the RL candidate pool and all others enter the SFT candidate pool. The original source files, exclusion manifest and retained-slide manifests are preserved with the training configuration. Neither retained slide-level pool contains an exact evaluation slide-identifier match. Patient identifiers are unavailable for some sources, so this audit establishes slide-level disjointness rather than patient-level independence.

A.3 EVALUATION-SET COMPOSITION

Supplementary Table 1 reports the sample or question counts and category-level composition of each evaluation set.

Supplementary Table 1: Composition of the patch/ROI-level and slide-level evaluation datasets used in this work. Patch/ROI rows report complete datasets used for linear-probe evaluation; slide-level rows report evaluated question counts.

Benchmark	Scope or category	N	Composition or split breakdown
Patch/ROI	CRC-100K	100,000	9-class patch classification
Patch/ROI	MHIST	3,152	2-class patch classification
Patch/ROI	Osteo	1,144	3-class patch classification
Patch/ROI	WSSS4LUAD	10,211	3-class patch classification
Patch/ROI	RenalCell	52,713	6-class patch classification
Patch/ROI	LC25000-Lung	15,000	3-class patch classification
Patch/ROI	LC25000-Colon	10,000	2-class patch classification
Patch/ROI	SICAPv2	10,340	4-class patch classification
PathMMU	Overall (test + test-tiny)	9,583	Test: 8,445; test-tiny: 1,138

Continued on next page

Benchmark	Scope or category	N	Composition or split breakdown
PathMMU	Atlas	1,007	Test: 799; test-tiny: 208
PathMMU	EduContent	1,938	Test: 1,683; test-tiny: 255
PathMMU	PathCLS	1,809	Test: 1,632; test-tiny: 177
PathMMU	PubMed	3,068	Test: 2,787; test-tiny: 281
PathMMU	SocialPath	1,761	Test: 1,544; test-tiny: 217
SlideBench-VQA-TCGA	Overall closed-ended VQA	3,176	Four-option MCQ
SlideBench-VQA-TCGA	Microscopy	937	Tissue Architecture and Arrangement: 362; Histopathological Changes: 328; Tumor Characteristics: 163; Cytomorphological Characteristics: 84
SlideBench-VQA-TCGA	Diagnosis	2,077	Disease Detection: 43; Disease Classification: 828; Grading: 608; Staging: 475; Differential Diagnosis: 123
SlideBench-VQA-TCGA	Clinical	162	Treatment Guidance: 68; Biomarker Analysis: 27; Prognostic Assessment: 39; Risk Factors: 28
SlideBench-VQA-CPTAC	Overall closed-ended VQA	240	Four-option MCQ
SlideBench-VQA-CPTAC	CM	60	60 questions
SlideBench-VQA-CPTAC	LSCC	60	60 questions
SlideBench-VQA-CPTAC	LUAD	60	60 questions
SlideBench-VQA-CPTAC	UCEC	60	60 questions
WSI-Bench	Full test set	4,174	Open: 2,838; closed: 1,336
WSI-Bench	Closed portion used	1,336	MCQ: 1,288; Yes/No: 48
WSI-Bench	Morphological Analysis	682	Regional/Specific Feature Description (Close): 682
WSI-Bench	Diagnosis	606	Histological Typing: 208; Grading: 205; Molecular Subtyping: 48; Staging: 145
WSI-Bench	Treatment Planning	48	Prognosis (Yes/No): 48

Notes. Patch/ROI rows report samples, whereas reasoning-benchmark rows report questions. LC25000 is evaluated as separate Lung and Colon tasks, and only the closed-ended portion of WSI-Bench is scored. PathMMU answer-option counts vary across subsets. SlideBench-VQA-TCGA sample ID 1577 asks about microvessel invasion and is semantically a Microscopy/Tumor Characteristics item, although the source labels it as Disease Detection. Reported category-level metrics retain the source grouping for comparability.

A.4 Code availability

The source code, training configurations, and trained PathTIDE model checkpoints will be released upon acceptance.

B Extended method and training details

This section provides implementation-level settings and optimization details omitted from the main Method for brevity. It does not introduce an additional model component or training stage.

B.1 Stage 1 alignment losses and optimization

The three loss definitions and the complete Stage 1 objective are given in Section 3.2. This appendix records the implementation settings needed to reproduce that optimization.

Only the query and value projections in MUSK attention layers receive LoRA adapters, with rank 8, scaling factor 16 and dropout 0.05 (Hu et al., 2022). The alignment head

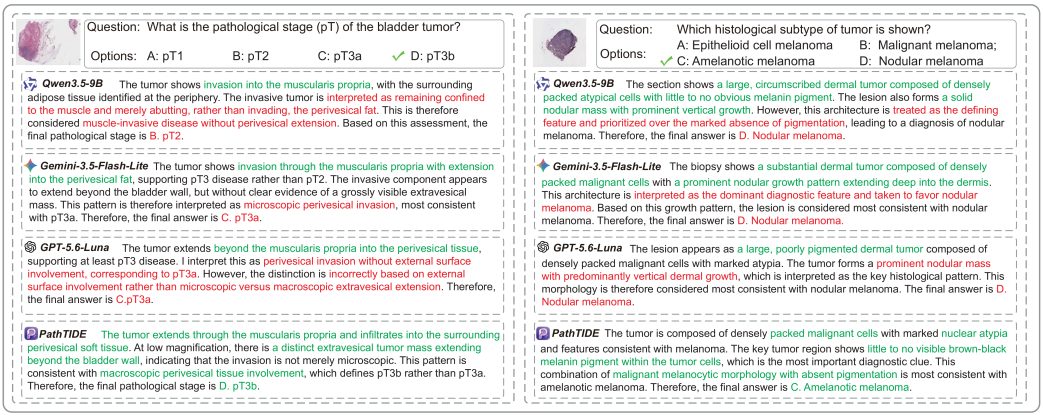


Figure 5: **Qualitative case studies on whole-slide slides.** The left panel presents pathological staging of a TCGA bladder tumor, and the right panel presents histological subtyping of a CPTAC melanoma. Each panel shows the slide thumbnail, question and answer options, followed by the complete responses from three baseline models and PathTIDE. Green text marks correct evidence or conclusions, whereas red text marks erroneous evidence, criteria or conclusions. All three baselines are incorrect in both cases, whereas PathTIDE selects the reference answer.

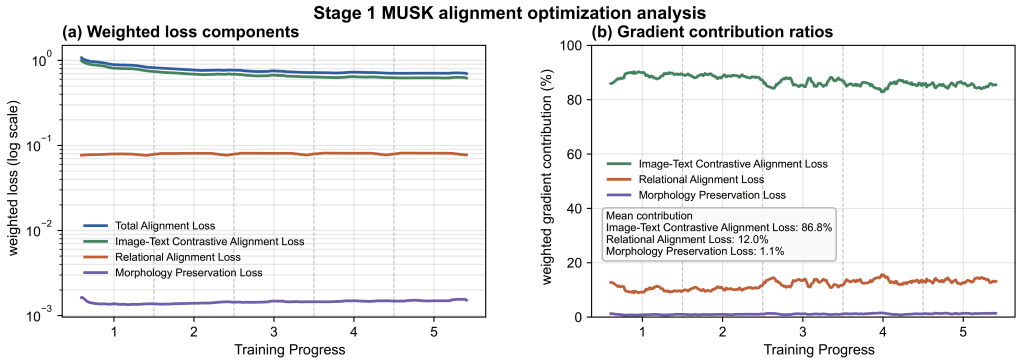


Figure 6: **Stage 1 MUSK alignment optimization analysis.** The curves summarize the recorded training trajectory over five logical epochs. **(a)** Weighted total and component losses, including the image–text contrastive alignment, relational alignment, and morphology preservation terms. **(b)** Normalized weighted gradient contributions measured on the final MUSK LoRA layer. The image–text contrastive, relational, and morphology-preservation terms contribute mean gradient shares of 86.8%, 12.0%, and 1.1%, respectively.

is `Linear(1024,2048)`–GELU–Dropout–`Linear(2048,2048)` with dropout 0.1. The Qwen3-VL-Embedding-2B teacher is frozen throughout alignment, and its image and text embeddings are gathered across devices for the contrastive objective.

We use AdamW with a fixed learning rate of 2×10^{-4}, weight decay 0.01, no warmup and no learning-rate scheduler. The per-device batch size is 8 on two GPUs, giving a global contrastive batch size of 16. After training, the completed aligned MUSK checkpoint is used to pre-extract all downstream expert features; the temporary alignment head is discarded.

B.2 Multi-level expert feature bank and native inputs

Patch-level samples contain one 384×384 patch. ROI and slide samples are partitioned using the same patch size, stride and tissue-filtering protocol. The aligned MUSK encodes every retained patch independently, and the complete CLS sequences are stored with sample identifiers, the Stage 1 checkpoint identifier and patch coordinates. Post-training reads these cached tensors and does not execute or back-propagate through MUSK.

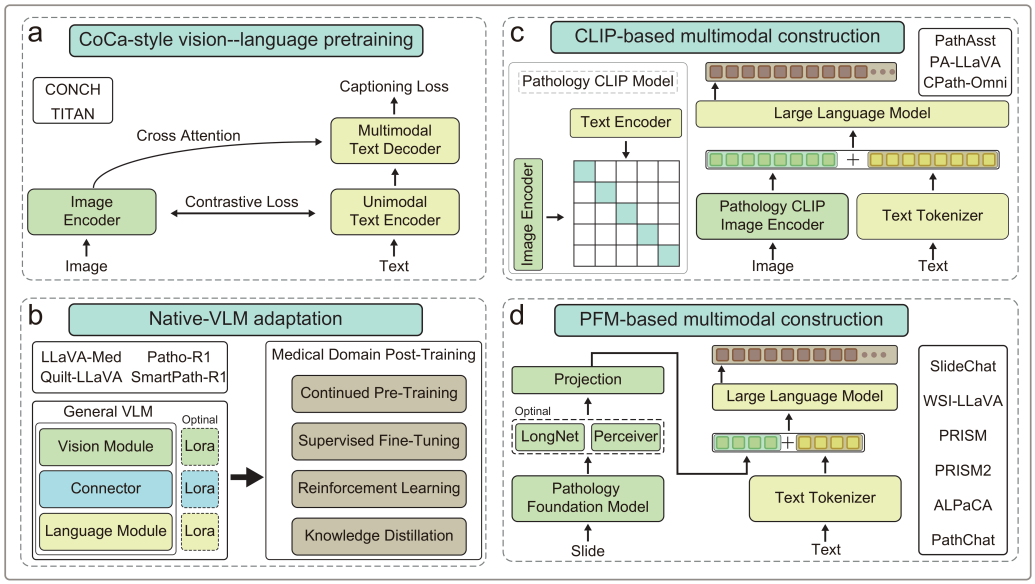


Figure 7: **Existing generative vision–language model designs for pathology.** The panels summarize four recurring architectural routes, categorized by how their visual and language pathways are assembled: **(a)** CoCa-style vision–language pretraining, **(b)** native-VLM adaptation, **(c)** CLIP-based multimodal construction, and **(d)** PFM-based multimodal construction.

For ROI and slide inputs, let $(x_i^{(\ell)}, y_i^{(\ell)})$ be a patch center and (W_ℓ, H_ℓ) the corresponding spatial extent. Coordinates are normalized as

$$u_i^{(\ell)} = \left(\frac{x_i^{(\ell)}}{W_\ell}, \frac{y_i^{(\ell)}}{H_\ell} \right) \in [0,1]^2, \tag{20}$$

and encoded by a fixed two-dimensional sine–cosine function. The positional term is zero for patch samples and $\mathrm{PE}_{2\mathrm{D}}(u_i)$ for ROI and slide samples. Patch and ROI native inputs are the complete input images. Slide native inputs are aspect-ratio-preserving thumbnails with a 1,536-pixel short side; no additional pre-transformer spatial merger is used.

Fusion operates in a $d_b = 1024$ bottleneck. Native queries and expert keys and values are masked according to the valid length of each expert sequence and projected back to $d = 4096$ before gated residual addition. The same projector, attention projections, gate and injection LayerNorm are shared across all three visual levels. Variable-length expert sequences are padded only within a mini-batch and accompanied by an attention mask; no valid expert token is discarded by this batching operation.

B.3 Macro-level ablation architectures

The macro-level variants use the same Qwen3.5-9B language backbone, aligned and frozen MUSK Pathology Expert, pathology post-training data, objectives and optimization budget. Shared components follow the same trainability settings as PathTIDE; a component omitted by a variant is not optimized. Let

$$\begin{aligned} T_{\mathrm{N}} &= E_{\mathrm{N}}(x_{\mathrm{N}}) \in \mathbb{R}^{N \times d}, \\ T_{\mathrm{P}} &= \Pi(F_{\mathrm{M}}(\mathcal{P}(x))) \in \mathbb{R}^{M \times d}, \qquad d = 4096. \end{aligned} \tag{21}$$

where T_{N} and T_{P} are the General Expert and projected Pathology Expert sequences, respectively. Padding in T_{P} is excluded by the Pathology Expert attention mask.

Replace+PT. This variant bypasses the General Expert Branch and uses the projected Pathology Expert sequence as the complete visual input to the language backbone:

$$T_{\mathrm{Replace}} := \mathrm{LN}_{\mathrm{R}}(T_{\mathrm{P}}), \qquad y \sim p_{\theta_{\mathrm{R}}}(y \mid q, T_{\mathrm{Replace}}) \,. \tag{22}$$

The visual-token span and its attention mask follow the M valid Pathology Expert tokens. During post-training, MUSK remains frozen, whereas the shared expert projector, replacement LayerNorm and language-backbone LoRA adapters are optimized. This construction tests whether pathology features alone can replace the model's native visual interface.

Add+PT. This variant retains both expert branches and uses the same cross-attention alignment as PathTIDE:

$$\Delta T := \mathrm{CrossAttn}(T_{\mathrm{N}} W_Q, T_{\mathrm{P}} W_K, T_{\mathrm{P}} W_V)\, W_O \in \mathbb{R}^{N \times d}. \tag{23}$$

Because the cross-attention update has the same shape as the General Expert sequence, the two can be added directly. Add+PT removes the learned token-wise gate and applies an ungated residual update:

$$T_{\mathrm{Add}} := \mathrm{LN}(T_{\mathrm{N}} + \Delta T)\,, \qquad y \sim p_{\theta_{\mathrm{A}}}(y \mid q, T_{\mathrm{Add}})\,. \tag{24}$$

Equivalently, Add+PT fixes the PathTIDE gate to $\alpha \equiv 1$. It retains the same expert mask, attention bottleneck, post-training data, base objectives, optimization budget and expert-active/expert-free batch schedule as PathTIDE, but omits the gate and its regularizer. Reported Add+PT results use the expert-active route. This comparison therefore isolates learned token-wise modulation from uniform residual injection.

B.4 Stage 2 supervised warm-start and parameter schedule

Stage 2 begins with supervised fine-tuning on the available instruction–response dataset

$$\mathcal{D}_{\mathrm{SFT}} := \{(x_n, y_n^\star)\}_{n=1}^{N}\,, \tag{25}$$

where x_n contains the pathology visual input, question and answer choices when present, and $y_n^\star = (y_{n,1}^\star, \ldots, y_{n,T_n}^\star)$ is the target response with T_n response tokens. Under teacher forcing, the model likelihood of the target response is

$$\pi_\theta(y_n^\star \mid x_n) := \prod_{t=1}^{T_n} \pi_\theta\big(y_{n,t}^\star \mid x_n, y_{n,<t}^\star\big)\,, \tag{26}$$

where $y_{n,<t}^\star$ denotes the gold prefix before token t. We optimize the response-token negative log-likelihood

$$\mathcal{L}_{\mathrm{SFT}}(\theta) := -\frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\sum_{t=1}^{T_n}\log \pi_\theta\big(y_{n,t}^\star \mid x_n, y_{n,<t}^\star\big)\,. \tag{27}$$

The loss is evaluated on response tokens, whereas the visual input, question and answer choices provide conditioning context. The parameter set θ contains the trainable expert-injection modules; the Qwen3.5-9B base parameters and aligned MUSK remain frozen. SFT provides a stable initialization for task behavior and output formatting. After SFT, we save the resulting PathTIDE checkpoint and create a frozen copy as the GRPO reference policy π_{ref}; this copy is created before the first GRPO update and is never updated.

Stage 2 inherits the Stage 1 MUSK LoRA configuration: query/value projections in all matched MUSK layers, rank 8, scaling factor 16 and dropout 0.05. The accepted aligned MUSK checkpoint is frozen and its CLS features are cached, so these adapters are not updated during Stage 2. Stage 2 optimization uses AdamW with learning rate 2×10^{-4}, weight decay 0.01, no warmup and no learning-rate scheduler, matching Stage 1. The expert projector, cross-attention, token-wise gate and injection LayerNorm are optimized during the native-frozen warm-start. Its regularized objective uses $\mathcal{L}_{\mathrm{SFT}} + 0.05\mathcal{L}_{\Delta} + 0.01\mathcal{L}_{\mathrm{gate}}$; the later native-adaptation/GRPO route uses the $0.02/0.005$ weights in Eq. 28.

The subsequent GRPO run uses all available answer-checkable closed-ended training instances, including multiple-choice, true/false, classification and normalized short-answer tasks. Patch-, ROI- and slide-level examples are combined in a mixed multi-level run with level-homogeneous mini-batches. The Qwen3.5-9B base parameters and aligned MUSK remain frozen. LoRA adapters in the native visual pathway and language backbone are trainable together with the expert projector, attention and output projections, token-wise gate and injection LayerNorm.

The two deployment routes are jointly optimized with a deterministic 10-batch cycle: nine batches that activate the Pathology Expert Branch (PFM-Full) followed by one batch that bypasses it (PFM-Free). The schedule is applied independently of visual level. On a PFM-Full batch, the model

reads the cached MUSK sequence and updates the native LoRA and full injection path. On a PFM-Free batch, it bypasses the complete gated residual fusion module—the expert projector, cross-attention, gate and injection LayerNorm—and updates only the native visual and language LoRA adapters. Thus, both PFM-Free training and PFM-Free inference compute $T_{\text{free}} = T_{\text{N}}$; this is distinct from setting the gate to zero while retaining the injection LayerNorm. This interleaving keeps the native pathway as an optimization anchor while introducing Pathology Expert evidence gradually. At inference, PFM-Free is reported as PathTIDE (PFM-Free) and PFM-Full as PathTIDE (PFM-Full); both use the same final checkpoint, and their contrast measures the complete expert-enabled fusion route rather than the residual update alone.

B.5 Stage 2 total loss and GRPO regularization

The construction in Eq. 18 is the core policy–KL objective. Implementation adds regularization for the expert-induced update and the token-wise gate. The total loss for the expert-active training route is

$$\mathcal{L}_{\text{total}} := \mathcal{L}_{\text{GRPO}} + 0.02\mathcal{L}_{\Delta} + 0.005\mathcal{L}_{\text{gate}}. \tag{28}$$

Here, $\mathcal{L}_{\Delta}$ regularizes the change induced by the expert injection path and $\mathcal{L}_{\text{gate}}$ regularizes the token-wise injection gate. The coefficient 0.02 on $\mathcal{L}_{\Delta}$ is independent of the KL coefficient $\beta = 0.02$ in Eq. 18. On PFM-Free batches, the complete injection path is bypassed as described in Appendix B.4.

With batch index b and native-token index n, the regularizers implemented in the CoAdapt training code are

$$\mathcal{L}_{\Delta} := \frac{1}{BN}\sum_{b=1}^{B}\sum_{n=1}^{N}\frac{\|\alpha_{b,n}\Delta T_{b,n}\|_2^2}{\|\text{sg}(T_{\text{N},b,n})\|_2^2 + \varepsilon}, \qquad \varepsilon = 10^{-6}, \tag{29}$$

and

$$\mathcal{L}_{\text{gate}} := \frac{1}{BN}\sum_{b=1}^{B}\sum_{n=1}^{N}\alpha_{b,n}. \tag{30}$$

Here $\alpha_{b,n}$ is broadcast over the feature dimension when multiplying $\Delta T_{b,n}$, and $\text{sg}(\cdot)$ denotes stop-gradient. The first term therefore measures gated injected energy relative to the detached native-token energy, while the second encourages conservative gate values.

B.6 Evaluation prompts

Two output contracts are used. During GRPO, each policy rollout contains a rationale and a final answer in the `<RATIONALE>...</RATIONALE>` and `<FINAL_ANSWER>...</FINAL_ANSWER>` blocks. Benchmark evaluation uses the separate answer-only contract shown below; it requires exactly `<answer>X</answer>` and no explanation. The canonical benchmark prompt is reproduced below.

Model-specific evaluation settings. We use fixed model versions and task-consistent decoding settings throughout evaluation. Models with native reasoning retain that capability, whereas models without it are instructed to reason before returning the required answer field. The exact model identifiers and decoding settings are retained with the evaluation scripts.

Evaluation prompt

System message
You are an expert in pathology, with specialized knowledge in histopathology and visual pathology. Answer each multiple-choice question based on the provided pathology image, question, and answer options.
Your entire final response must use exactly the following structure:
`<answer>X</answer>`
Replace `X` with one uppercase option letter from the provided choices. Do not output any text outside the `<answer>...</answer>` tags.
A pathology visual input, prepared using the evaluation preprocessing, is attached.
`Question: <QUESTION>`
`Answer choices: <OPTIONS>`

Evidence-reward judge prompt. The evidence judge is called only after the deterministic format and answer checks pass. It receives the same visual input, question, answer choices, verified reference answer and complete training rollout. The rollout contains the rationale and final answer

blocks defined above; unlike benchmark evaluation, the rationale is intentionally exposed to the evidence judge. The reference answer is a correctness constraint, not visual evidence: the judge must not invent pathology merely to agree with it. It does not re-evaluate answer correctness; it scores whether the rollout is grounded in visible pathology. The returned score is used directly as $R_{\mathrm{evi},g}$ in Eq. 15. The canonical judge prompt is:

Evidence-reward prompt

System message

You are a pathology evidence-grounding judge. A candidate training rollout contains one `<RATIONALE>...</RATIONALE>` block and one `<FINAL_ANSWER>...</FINAL_ANSWER>` block. It has already passed external checks for output format and answer correctness. The verified reference answer is supplied as a correctness constraint, not as visual evidence. Do not reassess whether the selected answer is correct. Judge only whether the rollout's visual observations, pathological claims and reasoning are supported by the attached image and the stated question.

Use only the supplied image, question, answer choices and reference answer. Do not use outside knowledge to invent findings, and do not treat the reference answer as proof that a finding is visible. Treat claims requiring unavailable clinical history, laboratory results, molecular tests, special stains, higher magnification or information absent from the image as unsupported. Distinguish directly visible findings from reasonable interpretations, unsupported hallucinations and claims that contradict the image. Do not reward confidence, length or fluent wording.

Assign one signed evidence-grounding score $S \in [-1, 1]$:

`+1`: all material claims are visually supported and no material unsupported finding is introduced;

`+0.5`: the response is mostly grounded, with only minor overclaiming or a small omission;

`0`: evidence support is mixed, indeterminate or largely absent;

`-0.5`: substantial claims are unsupported or the reasoning relies on information not available in the supplied inputs;

`-1`: material claims are contradicted by the image or are fabricated.

Use intermediate values only when the response falls between these anchors. Return exactly one numeric score, with no explanation and no additional text:

`<evidence_score>S</evidence_score>`

User message

A pathology visual input, prepared using the evaluation preprocessing, is attached.

`Question: <QUESTION>`
`Answer choices: <OPTIONS>`
`Reference answer: <REFERENCE_ANSWER>`
`Candidate rollout: <ROLLOUT>`

C SUPPLEMENTARY TABLES

Unless a table states otherwise, PathTIDE (PFM-Free) and PathTIDE (PFM-Full) are two inference modes of the same final checkpoint. PathTIDE (PFM-Free) bypasses the complete gated residual fusion module and forwards the retained native visual tokens directly to the language backbone, whereas PathTIDE (PFM-Full) activates the Stage 1-aligned MUSK feature bank and the full injection module.

Supplementary Table 2: **Overall performance on PathMMU multiple-choice questions.** Accuracy (95% CI) is reported on combined test + test-tiny, test and test-tiny subsets; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Overall (test + test-tiny) (N=9,583)	Overall (test) (N=8,445)	Overall (test-tiny) (N=1,138)
LLaVA-Med-v1.5-7B	0.283 (0.274, 0.292)	0.282 (0.273, 0.291)	0.287 (0.262, 0.310)
MedGemma-1.5-4B-it	0.495 (0.485, 0.505)	0.491 (0.481, 0.502)	0.522 (0.495, 0.552)
Patho-R1-7B	0.625 (0.616, 0.635)	0.619 (0.609, 0.630)	0.671 (0.644, 0.699)
GLM-4.6V-Flash	0.568 (0.559, 0.578)	0.564 (0.554, 0.574)	0.598 (0.569, 0.625)
GLM-4.6V	0.617 (0.607, 0.627)	0.613 (0.602, 0.623)	0.652 (0.626, 0.679)
Qwen3.5-2B	0.585 (0.576, 0.595)	0.580 (0.570, 0.591)	0.618 (0.587, 0.646)
Qwen3.5-4B	0.675 (0.666, 0.685)	0.670 (0.661, 0.680)	0.714 (0.686, 0.739)
Qwen3.5-9B	0.702 (0.693, 0.711)	0.698 (0.687, 0.707)	0.738 (0.712, 0.764)
Qwen3.7-Flash	*0.730 (0.721, 0.738)*	*0.726 (0.717, 0.735)*	*0.758 (0.734, 0.781)*
Gemini-3.5-Flash-Lite	0.689 (0.679, 0.699)	0.685 (0.675, 0.694)	0.720 (0.693, 0.746)
GPT-5.6-Luna	0.653 (0.643, 0.662)	0.648 (0.638, 0.658)	0.690 (0.664, 0.719)
GPT-5.6-Luna-Pro	0.663 (0.653, 0.674)	0.658 (0.649, 0.668)	0.703 (0.677, 0.728)
PathChat 1	0.609 (0.599, 0.619)	–	0.640 (0.613, 0.668)
PathChat +	0.728 (0.719, 0.736)	–	0.750 (0.725, 0.777)
Expert Performance	–	–	0.718
PathTIDE (PFM-Free)	0.800 (0.792, 0.808)	0.796 (0.787, 0.804)	0.830 (0.808, 0.852)
PathTIDE (PFM-Full)	**0.869 (0.862, 0.875)**	**0.865 (0.858, 0.872)**	**0.895 (0.877, 0.912)**

Supplementary Table 3: **Performance on the Atlas subset of PathMMU.** Accuracy (95% CI) is reported on combined test + test-tiny, test and test-tiny subsets; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Atlas (test + test-tiny) (N=1,007)	Atlas (test) (N=799)	Atlas (test-tiny) (N=208)
LLaVA-Med-v1.5-7B	0.271 (0.244, 0.298)	0.262 (0.232, 0.292)	0.308 (0.245, 0.370)
MedGemma-1.5-4B-it	0.515 (0.486, 0.547)	0.516 (0.482, 0.551)	0.514 (0.447, 0.577)
Patho-R1-7B	0.767 (0.741, 0.792)	0.760 (0.730, 0.787)	0.793 (0.740, 0.846)
GLM-4.6V-Flash	0.609 (0.577, 0.639)	0.594 (0.561, 0.627)	0.663 (0.601, 0.721)
GLM-4.6V	0.686 (0.658, 0.716)	0.685 (0.651, 0.718)	0.692 (0.630, 0.755)
Qwen3.5-2B	0.647 (0.618, 0.676)	0.643 (0.611, 0.677)	0.663 (0.601, 0.731)
Qwen3.5-4B	0.720 (0.691, 0.749)	0.711 (0.678, 0.742)	0.755 (0.702, 0.817)
Qwen3.5-9B	0.774 (0.748, 0.799)	0.775 (0.747, 0.802)	0.769 (0.707, 0.832)
Qwen3.7-Flash	0.801 (0.778, 0.825)	*0.802 (0.777, 0.827)*	0.798 (0.745, 0.851)
Gemini-3.5-Flash-Lite	0.745 (0.717, 0.772)	0.737 (0.708, 0.767)	0.774 (0.712, 0.832)
GPT-5.6-Luna	0.739 (0.711, 0.765)	0.735 (0.703, 0.763)	0.755 (0.697, 0.808)
GPT-5.6-Luna-Pro	0.767 (0.739, 0.791)	0.767 (0.740, 0.796)	0.764 (0.707, 0.827)
PathChat 1	0.753 (0.727, 0.779)	–	0.788 (0.731, 0.841)
PathChat +	*0.811 (0.786, 0.834)*	–	*0.817 (0.764, 0.870)*
Expert Performance	–	–	0.683
PathTIDE (PFM-Free)	0.833 (0.809, 0.856)	0.829 (0.802, 0.855)	0.851 (0.803, 0.894)
PathTIDE (PFM-Full)	**0.867 (0.846, 0.888)**	**0.862 (0.839, 0.885)**	**0.885 (0.837, 0.928)**

Supplementary Table 4: **Performance on the EduContent subset of PathMMU.** Accuracy (95% CI) is reported on combined test + test-tiny, test and test-tiny subsets; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	EduContent (test + test-tiny) (N=1,938)	EduContent (test) (N=1,683)	EduContent (test-tiny) (N=255)
LLaVA-Med-v1.5-7B	0.300 (0.278, 0.320)	0.306 (0.284, 0.328)	0.263 (0.208, 0.318)
MedGemma-1.5-4B-it	0.525 (0.502, 0.546)	0.517 (0.492, 0.540)	0.580 (0.518, 0.639)
Patho-R1-7B	0.669 (0.648, 0.689)	0.660 (0.638, 0.684)	0.725 (0.671, 0.780)
GLM-4.6V-Flash	0.599 (0.576, 0.620)	0.589 (0.564, 0.611)	0.663 (0.600, 0.718)
GLM-4.6V	0.634 (0.613, 0.657)	0.626 (0.602, 0.650)	0.686 (0.631, 0.745)
Qwen3.5-2B	0.575 (0.554, 0.598)	0.564 (0.541, 0.586)	0.651 (0.588, 0.706)
Qwen3.5-4B	0.669 (0.647, 0.689)	0.657 (0.633, 0.677)	0.749 (0.694, 0.800)
Qwen3.5-9B	0.690 (0.669, 0.711)	0.676 (0.654, 0.698)	0.780 (0.725, 0.831)
Qwen3.7-Flash	*0.732 (0.711, 0.752)*	*0.720 (0.698, 0.742)*	*0.808 (0.757, 0.855)*
Gemini-3.5-Flash-Lite	0.678 (0.656, 0.698)	0.672 (0.649, 0.695)	0.714 (0.663, 0.769)
GPT-5.6-Luna	0.693 (0.673, 0.715)	0.683 (0.660, 0.704)	0.765 (0.714, 0.812)
GPT-5.6-Luna-Pro	0.700 (0.679, 0.720)	0.690 (0.668, 0.711)	0.765 (0.714, 0.820)
PathChat 1	0.631 (0.609, 0.651)	–	0.678 (0.616, 0.737)
PathChat + Expert Performance	0.709 (0.689, 0.728) –	– –	0.722 (0.663, 0.780) 0.690
PathTIDE (PFM-Free)	0.798 (0.780, 0.814)	0.787 (0.766, 0.806)	0.875 (0.835, 0.910)
PathTIDE (PFM-Full)	**0.857 (0.841, 0.873)**	**0.846 (0.828, 0.863)**	**0.929 (0.894, 0.961)**

Supplementary Table 5: **Performance on the PathCLS subset of PathMMU.** Accuracy (95% CI) is reported on combined test + test-tiny, test and test-tiny subsets; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	PathCLS (test + test-tiny) (N=1,809)	PathCLS (test) (N=1,632)	PathCLS (test-tiny) (N=177)
LLaVA-Med-v1.5-7B	0.217 (0.197, 0.235)	0.216 (0.197, 0.236)	0.220 (0.164, 0.282)
MedGemma-1.5-4B-it	0.390 (0.369, 0.413)	0.390 (0.367, 0.415)	0.390 (0.316, 0.469)
Patho-R1-7B	0.431 (0.408, 0.453)	0.430 (0.404, 0.453)	0.446 (0.373, 0.520)
GLM-4.6V-Flash	0.353 (0.332, 0.375)	0.357 (0.333, 0.379)	0.322 (0.254, 0.395)
GLM-4.6V	0.421 (0.398, 0.443)	0.422 (0.399, 0.445)	0.412 (0.345, 0.480)
Qwen3.5-2B	0.552 (0.531, 0.574)	0.555 (0.531, 0.578)	0.525 (0.452, 0.599)
Qwen3.5-4B	0.641 (0.620, 0.662)	0.641 (0.616, 0.664)	0.638 (0.571, 0.712)
Qwen3.5-9B	0.658 (0.637, 0.680)	0.658 (0.634, 0.682)	0.655 (0.588, 0.729)
Qwen3.7-Flash	0.656 (0.634, 0.678)	*0.661 (0.637, 0.683)*	0.616 (0.542, 0.689)
Gemini-3.5-Flash-Lite	0.633 (0.611, 0.653)	0.634 (0.610, 0.659)	0.621 (0.548, 0.689)
GPT-5.6-Luna	0.442 (0.420, 0.464)	0.444 (0.418, 0.468)	0.424 (0.350, 0.497)
GPT-5.6-Luna-Pro	0.460 (0.439, 0.484)	0.463 (0.438, 0.487)	0.435 (0.362, 0.514)
PathChat 1	0.452 (0.430, 0.475)	–	0.424 (0.356, 0.497)
PathChat + Expert Performance	*0.753 (0.734, 0.774)* –	– –	0.751 (0.689, 0.814) *0.789*
PathTIDE (PFM-Free)	0.824 (0.805, 0.841)	0.826 (0.807, 0.845)	0.802 (0.740, 0.853)
PathTIDE (PFM-Full)	**0.905 (0.892, 0.919)**	**0.908 (0.894, 0.922)**	**0.881 (0.831, 0.927)**

Supplementary Table 6: **Performance on the PubMed subset of PathMMU.** Accuracy (95% CI) is reported on combined test + test-tiny, test and test-tiny subsets; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	PubMed (test + test-tiny) (N=3,068)	PubMed (test) (N=2,787)	PubMed (test-tiny) (N=281)
LLaVA-Med-v1.5-7B	0.312 (0.296, 0.328)	0.309 (0.292, 0.326)	0.345 (0.288, 0.402)
MedGemma-1.5-4B-it	0.520 (0.502, 0.537)	0.516 (0.498, 0.535)	0.555 (0.495, 0.612)
Patho-R1-7B	0.650 (0.634, 0.667)	0.645 (0.628, 0.663)	0.698 (0.641, 0.751)
GLM-4.6V-Flash	0.636 (0.619, 0.653)	0.635 (0.615, 0.654)	0.651 (0.594, 0.708)
GLM-4.6V	0.674 (0.656, 0.691)	0.672 (0.654, 0.689)	0.698 (0.648, 0.747)
Qwen3.5-2B	0.592 (0.575, 0.610)	0.590 (0.570, 0.610)	0.609 (0.548, 0.669)
Qwen3.5-4B	0.689 (0.674, 0.706)	0.685 (0.667, 0.703)	0.730 (0.676, 0.783)
Qwen3.5-9B	0.716 (0.699, 0.731)	0.712 (0.694, 0.727)	0.758 (0.705, 0.808)
Qwen3.7-Flash	*0.754 (0.739, 0.769)*	*0.750 (0.733, 0.766)*	*0.797 (0.751, 0.847)*
Gemini-3.5-Flash-Lite	0.723 (0.707, 0.738)	0.722 (0.706, 0.738)	0.733 (0.680, 0.786)
GPT-5.6-Luna	0.701 (0.685, 0.718)	0.700 (0.682, 0.718)	0.715 (0.658, 0.765)
GPT-5.6-Luna-Pro	0.706 (0.689, 0.722)	0.701 (0.686, 0.718)	0.758 (0.708, 0.808)
PathChat 1	0.636 (0.620, 0.654)	–	0.633 (0.580, 0.687)
PathChat +	0.723 (0.707, 0.739)	–	0.776 (0.730, 0.822)
Expert Performance	–	–	0.729
PathTIDE (PFM-Free)	0.763 (0.747, 0.777)	0.755 (0.738, 0.770)	0.843 (0.797, 0.886)
PathTIDE (PFM-Full)	**0.846 (0.832, 0.858)**	**0.839 (0.825, 0.851)**	**0.918 (0.883, 0.950)**

Supplementary Table 7: **Performance on the SocialPath subset of PathMMU.** Accuracy (95% CI) is reported on combined test + test-tiny, test and test-tiny subsets; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	SocialPath (test + test-tiny) (N=1,761)	SocialPath (test) (N=1,544)	SocialPath (test-tiny) (N=217)
LLaVA-Med-v1.5-7B	0.287 (0.265, 0.309)	0.289 (0.267, 0.311)	0.276 (0.221, 0.336)
MedGemma-1.5-4B-it	0.514 (0.492, 0.538)	0.513 (0.488, 0.539)	0.525 (0.461, 0.594)
Patho-R1-7B	0.653 (0.631, 0.677)	0.655 (0.632, 0.677)	0.641 (0.576, 0.705)
GLM-4.6V-Flash	0.612 (0.591, 0.635)	0.612 (0.589, 0.635)	0.613 (0.548, 0.677)
GLM-4.6V	0.663 (0.640, 0.685)	0.656 (0.631, 0.679)	0.710 (0.650, 0.765)
Qwen3.5-2B	0.581 (0.557, 0.605)	0.575 (0.551, 0.600)	0.622 (0.562, 0.687)
Qwen3.5-4B	0.669 (0.647, 0.691)	0.668 (0.645, 0.692)	0.673 (0.613, 0.733)
Qwen3.5-9B	0.698 (0.676, 0.719)	0.698 (0.674, 0.720)	0.700 (0.636, 0.760)
Qwen3.7-Flash	*0.720 (0.700, 0.740)*	*0.719 (0.696, 0.742)*	0.728 (0.673, 0.788)
Gemini-3.5-Flash-Lite	0.668 (0.644, 0.691)	0.659 (0.635, 0.683)	*0.737 (0.677, 0.793)*
GPT-5.6-Luna	0.690 (0.668, 0.711)	0.685 (0.663, 0.709)	0.724 (0.668, 0.783)
GPT-5.6-Luna-Pro	0.698 (0.676, 0.720)	0.696 (0.674, 0.718)	0.719 (0.659, 0.779)
PathChat 1	0.614 (0.591, 0.637)	–	0.638 (0.576, 0.699)
PathChat +	0.684 (0.660, 0.706)	–	0.690 (0.629, 0.747)
Expert Performance	–	–	0.715
PathTIDE (PFM-Free)	0.825 (0.806, 0.842)	0.833 (0.813, 0.852)	0.765 (0.705, 0.820)
PathTIDE (PFM-Full)	**0.886 (0.872, 0.901)**	**0.892 (0.878, 0.908)**	**0.843 (0.793, 0.885)**

Supplementary Table 8: **Performance on SlideBench-VQA-TCGA.** Accuracy (95% CI) is reported overall and for the Microscopy, Diagnosis, and Clinical question categories; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Overall (N=3,176)	Microscopy (N=937)	Diagnosis (N=2,077)	Clinical (N=162)
LLaVA-Med-v1.5-7B	0.325 (0.309, 0.341)	0.425 (0.394, 0.458)	0.260 (0.242, 0.279)	0.593 (0.519, 0.667)
MedGemma-1.5-4B-it	0.388 (0.371, 0.405)	0.464 (0.431, 0.495)	0.338 (0.318, 0.358)	0.586 (0.518, 0.660)
GLM-4.6V-Flash	0.492 (0.474, 0.509)	0.529 (0.496, 0.561)	0.461 (0.440, 0.484)	0.673 (0.599, 0.741)
GLM-4.6V	0.534 (0.515, 0.551)	0.569 (0.536, 0.602)	0.505 (0.483, 0.526)	0.698 (0.630, 0.759)
Qwen3.5-2B	0.364 (0.348, 0.381)	0.467 (0.436, 0.499)	0.301 (0.283, 0.320)	0.568 (0.494, 0.642)
Qwen3.5-4B	0.483 (0.466, 0.499)	0.610 (0.581, 0.642)	0.411 (0.390, 0.432)	0.667 (0.593, 0.735)
Qwen3.5-9B	0.504 (0.486, 0.521)	0.594 (0.562, 0.622)	0.449 (0.429, 0.472)	0.691 (0.617, 0.765)
Qwen3.7-Flash	0.558 (0.541, 0.573)	0.677 (0.647, 0.708)	0.492 (0.471, 0.512)	0.716 (0.642, 0.778)
Gemini-3.5-Flash-Lite	0.521 (0.504, 0.537)	0.625 (0.594, 0.657)	0.463 (0.443, 0.483)	0.667 (0.593, 0.735)
GPT-5.6-Luna	0.558 (0.540, 0.575)	0.648 (0.617, 0.680)	0.506 (0.483, 0.528)	0.704 (0.636, 0.766)
GPT-5.6-Luna-Pro	0.575 (0.559, 0.591)	0.671 (0.642, 0.700)	0.521 (0.500, 0.545)	0.716 (0.648, 0.784)
PRISM2	0.556 (0.539, 0.574)	0.553 (0.520, 0.583)	0.554 (0.533, 0.576)	0.605 (0.537, 0.673)
WSI-LLaVA	0.629 (0.614, 0.646)	0.663 (0.631, 0.694)	0.617 (0.597, 0.639)	0.586 (0.512, 0.654)
SlideChat	*0.685 (0.668, 0.701)*	*0.762 (0.734, 0.788)*	*0.645 (0.623, 0.665)*	*0.741 (0.673, 0.802)*
PathTIDE (PFM-Free)	0.695 (0.678, 0.710)	0.773 (0.746, 0.799)	0.655 (0.634, 0.675)	0.753 (0.685, 0.815)
PathTIDE (PFM-Full)	**0.793 (0.778, 0.807)**	**0.856 (0.834, 0.878)**	**0.760 (0.741, 0.778)**	**0.846 (0.784, 0.901)**

Supplementary Table 9: **Detailed performance on the Microscopy category of SlideBench-VQA-TCGA.** Accuracy (95% CI) is reported for the four Microscopy subcategories; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Tissue Architecture and Arrangement (N=362)	Tumor Characteristics (N=163)	Cytomorphological Characteristics (N=84)	Histopathological Changes (N=328)
LLaVA-Med-v1.5-7B	0.536 (0.489, 0.583)	0.362 (0.288, 0.436)	0.262 (0.167, 0.357)	0.375 (0.326, 0.430)
MedGemma-1.5-4B-it	0.500 (0.450, 0.550)	0.472 (0.393, 0.546)	0.417 (0.310, 0.524)	0.433 (0.381, 0.485)
GLM-4.6V-Flash	0.558 (0.506, 0.613)	0.503 (0.423, 0.583)	0.571 (0.476, 0.679)	0.500 (0.448, 0.555)
GLM-4.6V	0.608 (0.558, 0.657)	0.534 (0.454, 0.613)	0.619 (0.524, 0.715)	0.530 (0.482, 0.585)
Qwen3.5-2B	0.475 (0.428, 0.528)	0.454 (0.374, 0.528)	0.571 (0.464, 0.679)	0.439 (0.387, 0.494)
Qwen3.5-4B	0.619 (0.569, 0.674)	0.595 (0.515, 0.669)	0.738 (0.643, 0.833)	0.576 (0.524, 0.625)
Qwen3.5-9B	0.707 (0.660, 0.754)	0.595 (0.521, 0.669)	0.071 (0.024, 0.131)	0.604 (0.549, 0.655)
Qwen3.7-Flash	0.746 (0.702, 0.790)	0.583 (0.503, 0.663)	0.738 (0.643, 0.833)	0.631 (0.579, 0.683)
Gemini-3.5-Flash-Lite	0.660 (0.613, 0.710)	0.607 (0.528, 0.681)	0.619 (0.512, 0.726)	0.598 (0.546, 0.649)
GPT-5.6-Luna	0.655 (0.605, 0.702)	0.613 (0.540, 0.687)	0.702 (0.595, 0.798)	0.643 (0.591, 0.692)
GPT-5.6-Luna-Pro	0.677 (0.627, 0.724)	0.663 (0.589, 0.736)	0.714 (0.619, 0.810)	0.659 (0.610, 0.707)
PRISM2	0.575 (0.522, 0.624)	0.558 (0.472, 0.632)	0.512 (0.405, 0.631)	0.537 (0.482, 0.588)
WSI-LLaVA	0.638 (0.591, 0.685)	0.699 (0.626, 0.773)	0.762 (0.667, 0.845)	0.646 (0.595, 0.695)
SlideChat	*0.812 (0.771, 0.851)*	*0.712 (0.638, 0.785)*	*0.774 (0.679, 0.857)*	*0.726 (0.677, 0.771)*
PathTIDE (PFM-Free)	0.823 (0.782, 0.862)	0.724 (0.650, 0.798)	0.786 (0.690, 0.869)	0.738 (0.692, 0.784)
PathTIDE (PFM-Full)	**0.878 (0.843, 0.912)**	**0.773 (0.706, 0.834)**	**0.881 (0.810, 0.940)**	**0.866 (0.826, 0.902)**

Supplementary Table 10: **Detailed performance on the Diagnosis category of SlideBench-VQA-TCGA, part I.** Accuracy (95% CI) is reported for Disease Detection, Disease Classification, and Staging; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Disease Detection (N=43)	Disease Classification (N=828)	Staging (N=475)
LLaVA-Med-v1.5-7B	0.395 (0.256, 0.535)	0.318 (0.286, 0.350)	0.238 (0.200, 0.280)
MedGemma-1.5-4B-it	0.605 (0.465, 0.744)	0.362 (0.330, 0.396)	0.255 (0.217, 0.295)
GLM-4.6V-Flash	0.558 (0.395, 0.721)	0.499 (0.465, 0.530)	0.423 (0.381, 0.467)
GLM-4.6V	0.558 (0.395, 0.698)	0.546 (0.512, 0.576)	0.440 (0.394, 0.486)
Qwen3.5-2B	0.558 (0.418, 0.698)	0.390 (0.357, 0.424)	0.147 (0.116, 0.179)
Qwen3.5-4B	0.535 (0.395, 0.674)	0.478 (0.444, 0.510)	0.316 (0.274, 0.362)
Qwen3.5-9B	0.558 (0.395, 0.698)	0.522 (0.487, 0.553)	0.371 (0.326, 0.415)
Qwen3.7-Flash	0.605 (0.465, 0.744)	0.581 (0.546, 0.616)	0.398 (0.356, 0.438)
Gemini-3.5-Flash-Lite	0.535 (0.372, 0.698)	0.514 (0.481, 0.547)	0.413 (0.364, 0.459)
GPT-5.6-Luna	0.535 (0.395, 0.674)	0.572 (0.536, 0.606)	0.413 (0.371, 0.455)
GPT-5.6-Luna-Pro	0.535 (0.395, 0.674)	0.579 (0.541, 0.610)	0.423 (0.375, 0.469)
PRISM2	0.535 (0.395, 0.674)	0.656 (0.624, 0.691)	0.406 (0.362, 0.448)
WSI-LLaVA	0.698 (0.558, 0.837)	0.714 (0.686, 0.745)	0.495 (0.446, 0.539)
SlideChat	*0.674 (0.535, 0.814)*	*0.729 (0.698, 0.760)*	*0.617 (0.573, 0.661)*
PathTIDE (PFM-Free)	*0.674 (0.535, 0.814)*	0.739 (0.709, 0.766)	0.627 (0.583, 0.672)
PathTIDE (PFM-Full)	**0.767 (0.628, 0.884)**	**0.843 (0.818, 0.867)**	**0.709 (0.669, 0.752)**

Supplementary Table 11: **Detailed performance on the Diagnosis category of SlideBench-VQA-TCGA, part II.** Accuracy (95% CI) is reported for Grading and Differential Diagnosis; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Grading (N=608)	Differential Diagnosis (N=123)
LLaVA-Med-v1.5-7B	0.169 (0.140, 0.199)	0.350 (0.268, 0.439)
MedGemma-1.5-4B-it	0.340 (0.303, 0.377)	0.398 (0.309, 0.488)
GLM-4.6V-Flash	0.414 (0.377, 0.451)	0.545 (0.455, 0.634)
GLM-4.6V	0.488 (0.447, 0.530)	0.545 (0.463, 0.634)
Qwen3.5-2B	0.271 (0.235, 0.306)	0.358 (0.276, 0.447)
Qwen3.5-4B	0.372 (0.334, 0.408)	0.472 (0.390, 0.569)
Qwen3.5-9B	0.391 (0.354, 0.431)	0.512 (0.423, 0.602)
Qwen3.7-Flash	0.419 (0.382, 0.461)	0.577 (0.488, 0.675)
Gemini-3.5-Flash-Lite	0.426 (0.386, 0.465)	0.472 (0.390, 0.553)
GPT-5.6-Luna	0.479 (0.436, 0.520)	0.537 (0.455, 0.618)
GPT-5.6-Luna-Pro	0.502 (0.460, 0.544)	0.602 (0.520, 0.691)
PRISM2	0.543 (0.505, 0.586)	0.504 (0.415, 0.593)
WSI-LLaVA	0.566 (0.530, 0.604)	0.667 (0.585, 0.748)
SlideChat	0.539 (0.498, 0.577)	*0.699 (0.618, 0.780)*
PathTIDE (PFM-Free)	*0.549 (0.508, 0.587)*	0.707 (0.626, 0.789)
PathTIDE (PFM-Full)	**0.678 (0.640, 0.714)**	**0.805 (0.732, 0.870)**

Supplementary Table 12: **Detailed performance on the Clinical category of SlideBench-VQA-TCGA.** Accuracy (95% CI) is reported for each Clinical subcategory; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Treatment Guidance (N=68)	Biomarker Analysis (N=27)	Risk Factors (N=28)	Prognostic Assessment (N=39)
LLaVA-Med-v1.5-7B	0.588 (0.471, 0.721)	0.556 (0.370, 0.741)	0.679 (0.500, 0.857)	0.564 (0.410, 0.718)
MedGemma-1.5-4B-it	0.632 (0.515, 0.735)	0.519 (0.333, 0.704)	0.571 (0.393, 0.750)	0.564 (0.410, 0.718)
GLM-4.6V-Flash	0.750 (0.647, 0.853)	0.519 (0.333, 0.704)	0.786 (0.607, 0.929)	0.564 (0.410, 0.718)
GLM-4.6V	0.750 (0.632, 0.853)	0.593 (0.407, 0.778)	**0.821 (0.679, 0.929)**	0.590 (0.436, 0.744)
Qwen3.5-2B	0.691 (0.574, 0.809)	0.333 (0.148, 0.519)	0.571 (0.393, 0.750)	0.513 (0.359, 0.667)
Qwen3.5-4B	0.706 (0.588, 0.809)	0.704 (0.519, 0.852)	0.607 (0.429, 0.786)	*0.615 (0.462, 0.769)*
Qwen3.5-9B	0.735 (0.618, 0.838)	0.704 (0.519, 0.852)	0.679 (0.500, 0.857)	*0.615 (0.462, 0.769)*
Qwen3.7-Flash	0.765 (0.647, 0.868)	*0.667 (0.481, 0.852)*	0.786 (0.643, 0.929)	*0.615 (0.462, 0.769)*
Gemini-3.5-Flash-Lite	0.691 (0.574, 0.794)	*0.667 (0.481, 0.852)*	0.786 (0.643, 0.929)	0.538 (0.385, 0.692)
GPT-5.6-Luna	0.809 (0.706, 0.897)	0.704 (0.519, 0.852)	0.643 (0.464, 0.821)	0.564 (0.410, 0.718)
GPT-5.6-Luna-Pro	0.809 (0.706, 0.897)	0.704 (0.519, 0.852)	*0.714 (0.536, 0.857)*	0.564 (0.410, 0.718)
PRISM2	0.706 (0.588, 0.809)	0.481 (0.296, 0.667)	0.536 (0.357, 0.714)	0.564 (0.410, 0.718)
WSI-LLaVA	0.662 (0.544, 0.765)	0.407 (0.222, 0.593)	0.607 (0.429, 0.786)	0.564 (0.410, 0.718)
SlideChat	*0.838 (0.750, 0.926)*	**0.778 (0.593, 0.926)**	0.607 (0.429, 0.786)	0.667 (0.513, 0.821)
PathTIDE (PFM-Free)	0.853 (0.765, 0.926)	**0.778 (0.593, 0.926)**	0.607 (0.429, 0.786)	0.667 (0.513, 0.821)
PathTIDE (PFM-Full)	**0.897 (0.824, 0.956)**	**0.778 (0.593, 0.926)**	**0.821 (0.679, 0.964)**	**0.821 (0.692, 0.923)**

Supplementary Table 13: **Performance on SlideBench-VQA-CPTAC.** Accuracy (95% CI) is reported overall and separately for the CM, LSCC, LUAD, and UCEC tumor cohorts; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Overall (N=240)	CM (N=60)	LSCC (N=60)	LUAD (N=60)	UCEC (N=60)
LLaVA-Med-v1.5-7B	0.188 (0.142, 0.242)	0.217 (0.117, 0.333)	0.083 (0.017, 0.167)	0.250 (0.150, 0.367)	0.200 (0.100, 0.317)
MedGemma-1.5-4B-it	0.246 (0.192, 0.300)	0.300 (0.200, 0.417)	0.067 (0.017, 0.133)	0.200 (0.100, 0.317)	0.417 (0.283, 0.550)
GLM-4.6V-Flash	0.312 (0.258, 0.375)	0.267 (0.167, 0.383)	0.250 (0.150, 0.367)	0.283 (0.167, 0.400)	0.450 (0.317, 0.583)
GLM-4.6V	0.396 (0.338, 0.458)	0.467 (0.350, 0.600)	0.383 (0.267, 0.517)	0.333 (0.217, 0.450)	0.400 (0.283, 0.517)
Qwen3.5-2B	0.179 (0.133, 0.229)	0.050 (0.000, 0.117)	0.200 (0.100, 0.300)	0.233 (0.133, 0.334)	0.233 (0.133, 0.333)
Qwen3.5-4B	0.267 (0.217, 0.325)	0.117 (0.050, 0.200)	0.167 (0.083, 0.250)	0.300 (0.183, 0.433)	0.483 (0.350, 0.617)
Qwen3.5-9B	0.308 (0.250, 0.367)	0.183 (0.083, 0.283)	0.267 (0.166, 0.383)	0.300 (0.183, 0.417)	0.483 (0.350, 0.617)
Qwen3.7-Flash	0.421 (0.358, 0.483)	0.233 (0.133, 0.350)	0.667 (0.550, 0.783)	0.317 (0.200, 0.450)	0.467 (0.333, 0.583)
Gemini-3.5-Flash-Lite	0.417 (0.354, 0.479)	0.250 (0.150, 0.367)	0.550 (0.417, 0.667)	0.400 (0.283, 0.533)	0.467 (0.350, 0.583)
GPT-5.6-Luna	0.417 (0.354, 0.479)	0.100 (0.033, 0.183)	0.467 (0.333, 0.584)	0.500 (0.383, 0.634)	*0.600 (0.483, 0.717)*
GPT-5.6-Luna-Pro	0.383 (0.321, 0.446)	0.100 (0.033, 0.183)	0.367 (0.250, 0.483)	0.533 (0.417, 0.667)	0.533 (0.417, 0.650)
PRISM2	0.554 (0.496, 0.617)	0.233 (0.133, 0.333)	*0.717 (0.600, 0.833)*	0.583 (0.467, 0.700)	0.683 (0.567, 0.800)
WSI-LLaVA	0.571 (0.512, 0.629)	**0.917 (0.850, 0.983)**	0.183 (0.083, 0.283)	0.483 (0.350, 0.617)	0.683 (0.567, 0.800)
SlideChat	*0.600 (0.537, 0.662)*	0.583 (0.467, 0.717)	*0.717 (0.600, 0.817)*	*0.617 (0.500, 0.733)*	0.483 (0.350, 0.617)
PathTIDE (PFM-Free)	0.671 (0.613, 0.729)	*0.650 (0.517, 0.767)*	0.800 (0.700, 0.900)	0.683 (0.567, 0.800)	0.550 (0.417, 0.667)
PathTIDE (PFM-Full)	**0.812 (0.758, 0.863)**	0.767 (0.650, 0.867)	**0.867 (0.767, 0.950)**	**0.817 (0.717, 0.917)**	**0.800 (0.700, 0.900)**

Supplementary Table 14: **Performance on the closed-ended questions of WSI-Bench.** Accuracy (95% CI) is reported for Overall, Morphological Analysis, Diagnosis, and Treatment Planning. Overall accuracy is calculated over all 1,336 closed-ended questions; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Overall (N=1,336)	Morphological Analysis (N=682)	Diagnosis (N=606)	Treatment Planning (N=48)
LLaVA-Med-v1.5-7B	0.279 (0.255, 0.305)	0.324 (0.289, 0.358)	0.236 (0.203, 0.271)	0.188 (0.083, 0.312)
MedGemma-1.5-4B-it	0.580 (0.555, 0.609)	0.652 (0.616, 0.688)	0.480 (0.444, 0.518)	0.812 (0.688, 0.917)
GLM-4.6V-Flash	0.674 (0.651, 0.702)	0.733 (0.699, 0.767)	0.591 (0.553, 0.627)	*0.896 (0.812, 0.979)*
GLM-4.6V	0.682 (0.658, 0.710)	0.740 (0.707, 0.774)	0.602 (0.564, 0.639)	0.854 (0.750, 0.958)
Qwen3.5-2B	0.585 (0.561, 0.613)	0.708 (0.673, 0.742)	0.444 (0.406, 0.483)	0.625 (0.479, 0.750)
Qwen3.5-4B	0.657 (0.632, 0.685)	0.751 (0.717, 0.783)	0.550 (0.512, 0.587)	0.688 (0.542, 0.812)
Qwen3.5-9B	0.660 (0.635, 0.688)	0.762 (0.730, 0.793)	0.540 (0.502, 0.579)	0.729 (0.583, 0.833)
Qwen3.7-Flash	0.707 (0.683, 0.732)	0.801 (0.770, 0.830)	0.586 (0.548, 0.624)	*0.896 (0.812, 0.979)*
Gemini-3.5-Flash-Lite	0.652 (0.627, 0.680)	0.708 (0.673, 0.742)	0.574 (0.538, 0.611)	0.833 (0.729, 0.938)
GPT-5.6-Luna	0.730 (0.707, 0.755)	0.796 (0.765, 0.826)	0.645 (0.609, 0.682)	0.854 (0.750, 0.958)
GPT-5.6-Luna-Pro	0.722 (0.698, 0.747)	0.805 (0.774, 0.834)	0.620 (0.583, 0.657)	0.833 (0.729, 0.938)
PRISM2	0.688 (0.664, 0.714)	0.742 (0.707, 0.776)	0.614 (0.576, 0.650)	0.854 (0.750, 0.958)
WSI-LLaVA	**0.882 (0.865, 0.899)**	**0.924 (0.903, 0.943)**	0.825 (0.792, 0.855)	**1.000 (1.000, 1.000)**
SlideChat	0.733 (0.710, 0.758)	0.801 (0.770, 0.830)	0.601 (0.563, 0.637)	0.854 (0.750, 0.958)
PathTIDE (PFM-Free)	*0.824 (0.803, 0.843)*	*0.862 (0.836, 0.887)*	*0.772 (0.739, 0.805)*	0.938 (0.854, 1.000)
PathTIDE (PFM-Full)	0.869 (0.851, 0.887)	0.899 (0.875, 0.921)	**0.830 (0.800, 0.860)**	0.938 (0.854, 1.000)

Supplementary Table 15: **Detailed performance on the Diagnosis category of closed-ended WSI-Bench.** Accuracy (95% CI) is reported for Histological Typing, Grading, Molecular Subtyping, and Staging; CIs were estimated using 1,000 non-parametric bootstrap resamples. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Histological Typing (N=208)	Grading (N=205)	Molecular Subtyping (N=48)	Staging (N=145)
LLaVA-Med-v1.5-7B	0.101 (0.062, 0.144)	0.127 (0.083, 0.176)	0.458 (0.312, 0.604)	0.510 (0.427, 0.593)
MedGemma-1.5-4B-it	0.438 (0.375, 0.510)	0.322 (0.258, 0.385)	0.542 (0.396, 0.688)	0.745 (0.669, 0.814)
GLM-4.6V-Flash	0.567 (0.500, 0.635)	0.429 (0.361, 0.498)	0.521 (0.375, 0.667)	0.876 (0.821, 0.924)
GLM-4.6V	0.611 (0.543, 0.678)	0.395 (0.327, 0.463)	0.583 (0.438, 0.729)	0.890 (0.841, 0.938)
Qwen3.5-2B	0.471 (0.404, 0.543)	0.224 (0.166, 0.288)	0.458 (0.312, 0.604)	0.710 (0.634, 0.779)
Qwen3.5-4B	0.572 (0.505, 0.644)	0.337 (0.273, 0.400)	0.646 (0.500, 0.771)	0.786 (0.717, 0.848)
Qwen3.5-9B	0.519 (0.452, 0.591)	0.322 (0.258, 0.385)	0.583 (0.438, 0.729)	0.862 (0.800, 0.917)
Qwen3.7-Flash	0.635 (0.567, 0.697)	0.337 (0.273, 0.400)	0.583 (0.438, 0.729)	0.869 (0.814, 0.924)
Gemini-3.5-Flash-Lite	0.553 (0.486, 0.620)	0.380 (0.312, 0.449)	0.646 (0.500, 0.771)	0.855 (0.800, 0.910)
GPT-5.6-Luna	0.678 (0.615, 0.740)	0.439 (0.371, 0.507)	0.708 (0.562, 0.833)	0.869 (0.814, 0.924)
GPT-5.6-Luna-Pro	0.630 (0.562, 0.697)	0.424 (0.356, 0.493)	0.688 (0.542, 0.812)	0.862 (0.800, 0.917)
PRISM2	0.769 (0.712, 0.827)	0.556 (0.488, 0.620)	0.333 (0.208, 0.479)	0.566 (0.483, 0.648)
WSI-LLaVA	**0.942 (0.904, 0.971)**	0.649 (0.580, 0.712)	*0.729 (0.583, 0.833)*	*0.938 (0.897, 0.972)*
SlideChat	0.678 (0.615, 0.740)	0.420 (0.351, 0.488)	0.625 (0.479, 0.750)	0.731 (0.655, 0.800)
PathTIDE (PFM-Free)	*0.793 (0.736, 0.846)*	*0.610 (0.541, 0.678)*	0.792 (0.667, 0.896)	0.966 (0.931, 0.993)
PathTIDE (PFM-Full)	0.865 (0.817, 0.909)	**0.683 (0.619, 0.746)**	**0.875 (0.771, 0.958)**	**0.972 (0.945, 0.993)**

Supplementary Table 16: **Ablation study on the multiple-choice questions of PathMMU.** Accuracy (95% CI) is reported on the combined test + test-tiny, test, and test-tiny subsets; CIs were estimated using 1,000 non-parametric bootstrap resamples. *+PT* denotes matched pathology post-training; *Replace* uses projected Pathology Expert tokens without the General Expert Branch; *Add* uses ungated residual addition after cross-attention alignment. Full definitions are provided in Appendix B.3. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Overall (test + test-tiny) (N=9,583)	Overall (test) (N=8,445)	Overall (test-tiny) (N=1,138)
Qwen3.5-9B	0.702 (0.693, 0.711)	0.698 (0.687, 0.707)	0.738 (0.712, 0.764)
Qwen3.5-9B+PT	0.755 (0.745, 0.763)	0.750 (0.740, 0.759)	0.786 (0.763, 0.809)
Qwen3.5-9B(Replace)+PT	0.718 (0.708, 0.727)	0.713 (0.703, 0.723)	0.756 (0.730, 0.780)
Qwen3.5-9B(Add)+PT	*0.781 (0.772, 0.789)*	*0.777 (0.768, 0.786)*	*0.811 (0.788, 0.833)*
PathTIDE (PFM-Free)	0.800 (0.792, 0.808)	0.796 (0.787, 0.804)	0.830 (0.808, 0.852)
PathTIDE (PFM-Full)	**0.869 (0.862, 0.875)**	**0.865 (0.858, 0.872)**	**0.895 (0.877, 0.912)**

Supplementary Table 17: **Ablation study on SlideBench-VQA-TCGA.** Accuracy (95% CI) is reported overall and for the Microscopy, Diagnosis, and Clinical categories; CIs were estimated using 1,000 non-parametric bootstrap resamples. *+PT* denotes matched pathology post-training; *Replace* uses projected Pathology Expert tokens without the General Expert Branch; *Add* uses ungated residual addition after cross-attention alignment. Full definitions are provided in Appendix B.3. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Overall (N=3,176)	Microscopy (N=937)	Diagnosis (N=2,077)	Clinical (N=162)
Qwen3.5-9B	0.504 (0.486, 0.521)	0.594 (0.562, 0.622)	0.449 (0.429, 0.472)	0.691 (0.617, 0.765)
Qwen3.5-9B+PT	0.609 (0.591, 0.625)	0.693 (0.662, 0.721)	0.562 (0.540, 0.583)	0.722 (0.648, 0.790)
Qwen3.5-9B(Replace)+PT	0.548 (0.530, 0.565)	0.658 (0.628, 0.687)	0.488 (0.467, 0.511)	0.685 (0.611, 0.759)
Qwen3.5-9B(Add)+PT	*0.661 (0.643, 0.676)*	*0.740 (0.711, 0.766)*	*0.618 (0.596, 0.638)*	*0.747 (0.679, 0.809)*
PathTIDE (PFM-Free)	0.695 (0.678, 0.710)	0.773 (0.746, 0.799)	0.655 (0.634, 0.675)	0.753 (0.685, 0.815)
PathTIDE (PFM-Full)	**0.793 (0.778, 0.807)**	**0.856 (0.834, 0.878)**	**0.760 (0.741, 0.778)**	**0.846 (0.784, 0.901)**

Supplementary Table 18: **Ablation study on SlideBench-VQA-CPTAC.** Accuracy (95% CI) is reported overall and separately for the CM, LSCC, LUAD, and UCEC tumor cohorts; CIs were estimated using 1,000 non-parametric bootstrap resamples. *+PT* denotes matched pathology post-training; *Replace* uses projected Pathology Expert tokens without the General Expert Branch; *Add* uses ungated residual addition after cross-attention alignment. Full definitions are provided in Appendix B.3. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Overall (N=240)	CM (N=60)	LSCC (N=60)	LUAD (N=60)	UCEC (N=60)
Qwen3.5-9B	0.308 (0.250, 0.367)	0.183 (0.083, 0.283)	0.267 (0.166, 0.383)	0.300 (0.183, 0.417)	0.483 (0.350, 0.617)
Qwen3.5-9B+PT	0.508 (0.446, 0.571)	0.433 (0.316, 0.567)	0.567 (0.433, 0.684)	0.517 (0.383, 0.650)	0.517 (0.383, 0.650)
Qwen3.5-9B(Replace)+PT	0.408 (0.346, 0.471)	0.317 (0.200, 0.433)	0.417 (0.300, 0.550)	0.400 (0.283, 0.533)	0.500 (0.367, 0.633)
Qwen3.5-9B(Add)+PT	*0.604 (0.542, 0.667)*	*0.567 (0.433, 0.684)*	*0.700 (0.583, 0.817)*	*0.617 (0.483, 0.733)*	*0.533 (0.400, 0.667)*
PathTIDE (PFM-Free)	0.671 (0.613, 0.729)	0.650 (0.517, 0.767)	0.800 (0.700, 0.900)	0.683 (0.567, 0.800)	0.550 (0.417, 0.667)
PathTIDE (PFM-Full)	**0.812 (0.758, 0.863)**	**0.767 (0.650, 0.867)**	**0.867 (0.767, 0.950)**	**0.817 (0.717, 0.917)**	**0.800 (0.700, 0.900)**

Supplementary Table 20: **Linear-probe classification using frozen image embeddings on CRC, MHIST, Osteo and WSSS4LUAD.** Accuracy and 95% confidence intervals (CIs) are reported for each benchmark. Best results are shown in **bold**, and second-best results are underlined.

Model	CRC (9 classes)	MHIST (2 classes)	Osteo (3 classes)	WSSS4LUAD (3 classes)
Qwen3.5-9B	0.910 (0.903, 0.916)	0.800 (0.769, 0.830)	0.909 (0.870, 0.943)	0.920 (0.902, 0.936)
MUSK	0.954 (0.949, 0.958)	0.846 (0.818, 0.873)	0.943 (0.913, 0.970)	0.970 (0.959, 0.980)
MUSK+Qwen3.5-9B	**0.973 (0.969, 0.976)**	**0.916 (0.894, 0.937)**	**0.948 (0.917, 0.974)**	**0.976 (0.965, 0.985)**

Supplementary Table 21: **Linear-probe classification using frozen image embeddings on RenalCell, LC25000 and SICAPv2.** Accuracy and 95% confidence intervals (CIs) are reported for each benchmark. Best results are shown in **bold**, and second-best results are underlined.

Model	RenalCell (6 classes)	LC25000 Lung (3 classes)	LC25000 Colon (2 classes)	SICAPv2 (4 classes)
Qwen3.5-9B	0.910 (0.904, 0.916)	0.985 (0.980, 0.990)	0.997 (0.995, 0.999)	0.749 (0.731, 0.767)
MUSK	0.932 (0.927, 0.938)	0.996 (0.994, 0.999)	**1.000 (1.000, 1.000)**	0.805 (0.788, 0.823)
MUSK+Qwen3.5-9B	**0.953 (0.948, 0.957)**	**0.997 (0.995, 0.999)**	**1.000 (1.000, 1.000)**	**0.883 (0.869, 0.897)**

Supplementary Table 19: **Ablation study on the closed-ended questions of WSI-Bench.** Accuracy (95% CI) is reported for Overall, Morphological Analysis, Diagnosis, and Treatment Planning. Overall accuracy is calculated over all 1,336 closed-ended questions; CIs were estimated using 1,000 non-parametric bootstrap resamples. *+PT* denotes matched pathology post-training; *Replace* uses projected Pathology Expert tokens without the General Expert Branch; *Add* uses ungated residual addition after cross-attention alignment. Full definitions are provided in Appendix B.3. Best result is shown in **bold**, second best is underlined, and third best is *italicized*.

Model	Overall (N=1,336)	Morphological Analysis (N=682)	Diagnosis (N=606)	Treatment Planning (N=48)
Qwen3.5-9B	0.660 (0.635, 0.688)	0.762 (0.733, 0.795)	0.540 (0.502, 0.581)	0.729 (0.583, 0.854)
Qwen3.5-9B+PT	0.754 (0.730, 0.777)	0.821 (0.793, 0.850)	0.668 (0.630, 0.705)	*0.875 (0.771, 0.958)*
Qwen3.5-9B (Replace)+PT	0.700 (0.674, 0.721)	0.806 (0.777, 0.836)	0.583 (0.543, 0.622)	0.667 (0.521, 0.792)
Qwen3.5-9B (Add)+PT	*0.800 (0.778, 0.820)*	*0.850 (0.824, 0.877)*	*0.734 (0.700, 0.769)*	0.917 (0.833, 0.979)
PathTIDE (PFM-Free)	0.824 (0.803, 0.843)	0.862 (0.836, 0.887)	0.772 (0.739, 0.805)	**0.938 (0.854, 1.000)**
PathTIDE (PFM-Full)	**0.869 (0.851, 0.887)**	**0.899 (0.875, 0.921)**	**0.830 (0.800, 0.860)**	**0.938 (0.854, 1.000)**