HYPERBOLIC VISION-LANGUAGE MODEL FOR PATHOLOGICAL PATCH-SLIDE REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Gigapixel whole-slide image (WSI) analysis requires capturing localized pathological evidence at the patch level and integrating these fine-grained cues into a slide-level representation. Although slide-level foundation models automate holistic tissue analysis, their Euclidean representations discard feature-norm information during normalization and patch aggregation, even though this information could help organize hierarchical relationships and retain clinically meaningful fine-grained spatial cues. In this work, we propose **HYPERION**, a hyperbolic vision-language model for pathological patch-slide representation that leverages the inherent norm structure of pretrained Euclidean foundation models. HYPERION constructs text-conditioned Einstein midpoints over patch embeddings to summarize the patches most similar to each caption and uses entailment cones to enforce hierarchical inclusion among caption, midpoint, and slide representations. Confidence-filtered caption-patch correspondences provide pseudo-labeled patch-level contrastive supervision from slide-level normal or abnormal labels, while intermediate feature norms are propagated to the radii of final patch embeddings to expose norm information. In zero-shot tumor segmentation, HYPERION has the highest Dice on five benchmarks, while prediction evaluations show that HYPERION preserves or enhances pathological knowledge. Integrated into a WSI multimodal large language model, HYPERION improves overall report-generation metrics and yields more lesion-focused visual grounding in qualitative analyses, verifying the utility of hyperbolic patch-slide representations for pathology.

1 INTRODUCTION

Recent advances in computational pathology have shown the potential of deep learning for clinical-grade whole-slide image (WSI) analysis, cancer subtyping, mutation prediction, and prognosis prediction (Campanella et al., 2019; Coudray et al., 2018; Mobadersany et al., 2018). Foundation models like vision-language models (VLMs) and multimodal large language models (MLLMs), are transforming gigapixel slide analysis (Chen et al., 2024c; Vorontsov et al., 2024; Lu et al., 2024; Huang et al., 2023). While early approaches focused on patch-level encoders that learn local tissue representations, recent slide-level foundation models (Xu et al., 2024; Wang et al., 2024; Shaikovski et al., 2024; Ding et al., 2025) aggregate thousands of patch features into a slide representation and align it with clinical text. This transition has enabled holistic slide-level applications, including cancer subtyping, prognosis prediction, visual question answering, and report generation (Chen et al., 2025; Liang et al., 2025; Chen et al., 2024b), moving beyond isolated local recognition toward understanding of entire tissue sections.

However, matching a slide to a report does not necessarily establish which visual evidence supports each histopathological concept. WSI interpretation requires connecting cellular morphology and tissue organization to increasingly abstract diagnostic descriptions, with relevant evidence distributed unevenly across a gigapixel image. A representation useful for this process should preserve local morphological details while organizing their relationships to provide richer contexts of diagnoses. Unfortunately, existing slide models usually discard norm information during normalization for representation alignment, even though this norm information could help model hierarchical text–slide relationships and capture fine-grained spatial information.

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.

To address this challenge, we propose **HYPERION**, a hyperbolic fine-tuning framework that grounds histopathological concepts for patch-slide representations while preserving and enhancing their representation capability. In hyperbolic space, we can simultaneously expose norm information during patch-to-slide aggregation and explicitly align patch-to-slide entailment relationships. Specifically, HYPERION combines three complementary mechanisms: (1) hyperbolic contrastive learning and entailment alignment using text-conditioned Einstein midpoints to adaptively aggregate semantically relevant patch embeddings toward text embeddings, (2) pseudo-labeling that strengthens patch-to-patch relationships by leveraging the hyperbolic geometry, and (3) norm propagation that enhances the patch representations by exposing the intermediate feature norms to the hyperbolic space. Together, these enable fine-tuning of pretrained representations to ground fine-grained patch-level evidence while maintaining or improving slide-level understanding.

We evaluate HYPERION across segmentation, slide-level prediction, and report generation tasks. Across diverse segmentation tasks such as CAMELYON16, CAMELYON17, PANDA, and PAIP, HYPERION achieves a substantial improvement over baselines. Evaluations on 36 prediction benchmarks under supervised and zero-shot classification settings further demonstrate that the refined representations preserve or enhance existing histopathological knowledge. When integrated into a WSI-level MLLM, HYPERION improves report generation. These results highlight the value of fine-grained visual-semantic alignment for strengthening slide representations across both histopathological prediction and language generation.

Our contributions are summarized as follows:

- We propose **HYPERION**, a hyperbolic fine-tuning framework that organizes pretrained slide and text representations into a shared visual–semantic hierarchy through hyperbolic contrastive alignment and entailment guidance, strengthening histopathological grounding.
- We use text-conditioned Einstein midpoint aggregation to integrate fine-grained visual evidence, confidence-filtered patch-level contrastive supervision to align textual concepts with relevant regions, and intermediate norm propagation to improve radial encoding.
- Experiments demonstrate substantially improved segmentation and preserved or enhanced slide-level prediction performance on various benchmarks. Integrated into a WSI-level MLLM, HYPERION further improves report generation, demonstrating the value of hierarchy-aware grounding across discriminative and generative tasks.

2 Related Work

Foundation models for WSI analysis. Computational pathology aims to automate gigapixel whole-slide image (WSI) analysis for cancer subtyping, prognosis prediction, and biomarker assessment. Multiple instance learning (MIL) (Ilse et al., 2018; Li et al., 2021; Tang et al., 2024; Lu et al., 2021; Zhang et al., 2022; Shao et al., 2021; Chen et al., 2022; Lee et al., 2025) provides the prevailing framework by extracting patch-level embeddings and aggregating them into slide representations, motivating foundation models at both levels. At the patch level, vision-only, vision-language, and multimodal foundation models have substantially advanced local histopathology representation and reasoning (Chen et al., 2024c; Vorontsov et al., 2024; Huang et al., 2023; Ikezogwo et al., 2023; Lu et al., 2024; Sun et al., 2025; Seyfioglu et al., 2024; Zhang et al., 2026b). To capture global context across thousands of tissue regions, slide-level models such as Prov-GigaPath (Xu et al., 2024), CHIEF (Wang et al., 2024), PRISM (Shaikovski et al., 2024), and TITAN (Ding et al., 2025) aggregate patch features into holistic representations, while WSI-level MLLMs including WSI-VQA (Chen et al., 2024b), SlideChat (Chen et al., 2025), and WSI-LLaVA (Liang et al., 2025) extend them toward language understanding. However, most existing approaches rely on slide-level aggregation under Euclidean representations, which can obscure patch-level information when thousands of local regions are compressed into a single slide representation.

Hyperbolic representation learning. Hyperbolic space is well suited for modeling hierarchical structures because its volume expands exponentially with radius, enabling efficient representation of tree-like relationships (Nickel & Kiela, 2017; Ganea et al., 2018; Sala et al., 2018). This property has motivated hierarchical representation learning across modalities (Tifrea et al., 2019; Dhingra et al., 2018; Le et al., 2019; Khrulkov et al., 2020; Yan et al., 2023; Atigh et al., 2022; Liu et al.,

2019; Chami et al., 2019; Sinha et al., 2024). MERU (Desai et al., 2023) introduces hyperbolic contrastive learning and entailment constraints for image–text hierarchy, followed by HyCoCLIP (Pal et al., 2025) for intra-modal hierarchies and UNCHA (Kim et al., 2026) for one-to-many alignment between whole and part image regions. In pathology, HyperPath (Huang et al., 2025) learns multi-level WSI representations with angular alignment and entailment constraints, while HyperST (Zhang et al., 2026a) aligns gene embeddings and WSI features in hyperbolic space. HYPERION differs from prior hyperbolic works by leveraging local information already encoded in Euclidean patch representations to structure visual evidence jointly with slide-level semantics in hyperbolic space.

3 PRELIMINARIES

Lorentz hyperbolic space. In this work, we adopt the Lorentz hyperbolic space since it offers numerically stable distance computation and closed-form geometric operations. Specifically, it is defined as a Riemannian manifold of constant negative curvature $-\kappa$, with $\kappa \in \mathbb{R}^+$. Then, it is equipped with the Lorentzian inner product under the Minkowski space $\mathbb{R}^{n+1}$ as:

$$\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}} = -p_0 q_0 + \sum_{j=1}^{n} p_j q_j. \tag{1}$$

The n-dimensional Lorentz manifold is defined as the upper sheet of the two-sheeted hyperboloid, $\mathbb{L}_\kappa^n = \left\{ \mathbf{p} \in \mathbb{R}^{n+1} \,\middle|\, \langle \mathbf{p}, \mathbf{p} \rangle_{\mathbb{L}} = -\frac{1}{\kappa},\ p_0 > 0 \right\}$. For a point $\mathbf{p} \in \mathbb{L}^n$, we decompose $\mathbf{p}$ into its time and spatial components, denoted by p_0 and $\mathbf{p}_{1:n}$, respectively. These components satisfy $p_0 = \sqrt{\frac{1}{\kappa} + \|\mathbf{p}_{1:n}\|_2^2}$. Then, the geodesic distance between two points $\mathbf{p}, \mathbf{q} \in \mathbb{L}_\kappa^n$ is defined as $d_{\mathbb{L}}(\mathbf{p}, \mathbf{q})$, whose detailed formulations can be found in App. B. We also define the hyperbolic origin as $\mathbf{o} = [\sqrt{1/\kappa}, \mathbf{0}]$ and the radius of $\mathbf{p}$ as $r(\mathbf{p}) = d_{\mathbb{L}}(\mathbf{p}, \mathbf{o})$.

Tangent space and exponential mapping. The tangent space for each point $\mathbf{z} \in \mathbb{L}_\kappa^n$ is $T_{\mathbf{z}}\mathbb{L}_\kappa^n = \left\{ \mathbf{u} \in \mathbb{R}^{n+1} : \langle \mathbf{z}, \mathbf{u} \rangle_{\mathbb{L}} = 0 \right\}$. A tangent vector $\mathbf{u} \in T_{\mathbf{z}}\mathbb{L}_\kappa^n$ can be mapped to the Lorentz manifold using the exponential map, $\exp_{\mathbf{z}}^\kappa(\mathbf{u})$ and the logarithmic map $\log_{\mathbf{z}}^\kappa(\mathbf{p})$ projects a point $\mathbf{p} \in \mathbb{L}_\kappa^n$ onto $T_{\mathbf{z}}\mathbb{L}_\kappa^n$. Since $T_{\mathbf{o}}\mathbb{L}_\kappa^n$ is isomorphic to Euclidean space, pretrained Euclidean slide and text embeddings can be lifted into the Lorentz manifold by first treating them as tangent vectors at the origin o and then applying $\exp_{\mathbf{z}}^\kappa(\mathbf{u})$. The detailed formulations of $\exp_{\mathbf{z}}^\kappa(\mathbf{u})$ and $\log_{\mathbf{z}}^\kappa(\mathbf{p})$ are provided in App. B.

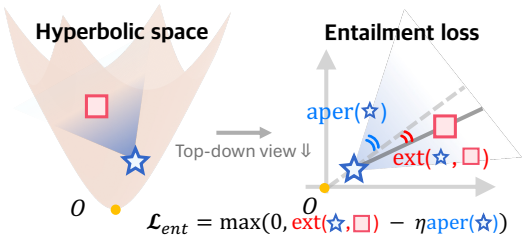

Figure 1: **Entailment loss in hyperbolic space.** A child (red square) is penalized when it falls outside the entailment cone of its parent (blue star).

Hyperbolic contrastive loss. Let $\mathbf{t}, \mathbf{v} \in \mathbb{L}_\kappa^n$ denote the hyperbolic text and visual embeddings of a positive pair, respectively, and let $\{\mathbf{t}_k\}$ and $\{\mathbf{v}_k\}$ denote their respective candidate embeddings. Using the Lorentz geodesic distance $d_{\mathbb{L}}$ and temperature parameter τ, we define the symmetric hyperbolic InfoNCE loss as

$$\mathcal{L}_{\text{con}}(\mathbf{t}, \mathbf{v}; \tau) = -\frac{1}{2}\left[\log \frac{\exp(-d_{\mathbb{L}}(\mathbf{v}, \mathbf{t})/\tau)}{\sum_k \exp(-d_{\mathbb{L}}(\mathbf{v}, \mathbf{t}_k)/\tau)} + \log \frac{\exp(-d_{\mathbb{L}}(\mathbf{t}, \mathbf{v})/\tau)}{\sum_k \exp(-d_{\mathbb{L}}(\mathbf{t}, \mathbf{v}_k)/\tau)} \right]. \tag{2}$$

Hyperbolic entailment cones. To encode hierarchical inclusion, we adopt hyperbolic entailment cones (Ganea et al., 2018), following prior hyperbolic vision–language models (Desai et al., 2023; Pal et al., 2025; Kim et al., 2026). Unlike contrastive alignment, entailment is defined over a modality-independent parent–child pair $(\mathbf{p}, \mathbf{q})$, encouraging the child $\mathbf{q}$ to lie within the cone rooted at the parent $\mathbf{p}$. This enables HYPERION to model hierarchical relationships across visual and text representations. Let $\phi(\mathbf{p}, \mathbf{q})$ be the angle between the outward radial direction at $\mathbf{p}$ and the geodesic toward $\mathbf{q}$, $\omega(\mathbf{p})$ the cone aperture, and η the constraint tightness. We define the entailment loss

$$\mathcal{L}_{\text{ent}}(\mathbf{p}, \mathbf{q}; \eta) = \max\left(0, \phi(\mathbf{p}, \mathbf{q}) - \eta\omega(\mathbf{p})\right). \tag{3}$$

More details on hyperbolic geometry can be found in App. B.

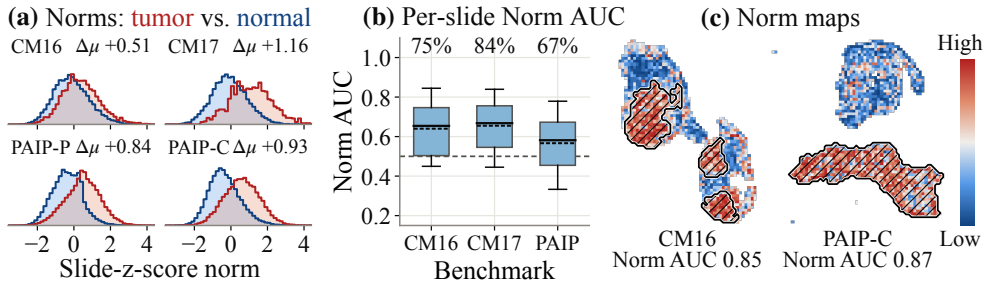

Figure 2: **Patch-embedding norms of TITAN.** (a) Slide-standardized norms of tumor and normal tiles on the slides whose per-slide norm AUC is at least 0.8 ($\Delta\mu$: difference of their means). (b) Per-slide norm AUC over all slides; percentages give the fraction of slides above 0.5. (c) Norm maps of a CAMELYON16 and a PAIP Colon slide; the tumor annotation is hatched with a black outline.

4 Hyperbolic Fine-Tuning for Visual-Semantic Grounding

In this section, we first empirically verify an intriguing observation: although the pretrained model's vision–language alignment is invariant to feature magnitude, pathological spatial cues are already inherently encoded in its feature norms. We then introduce HYPERION, a framework for fine-tuning a pretrained slide foundation model to organize visual and text representations into a shared visual–semantic hierarchy in $\mathbb{L}^n_\kappa$ (Fig. 3). The visual encoder produces a slide feature and its corresponding T patch features, while the text encoder produces a feature for each caption. Following Sec. 3, these Euclidean features are mapped to the Lorentz manifold via $\exp^\kappa_{\mathbf{o}}$. We denote the resulting slide embedding, patch embeddings, and caption embedding by $\mathbf{s}$, $\{\mathbf{x}_j\}_{j=1}^T$, and $\mathbf{t}$, lying in $\mathbb{L}^n_\kappa$.

4.1 Pretrained Feature Norms Encode Pathological Spatial Cues

In the Lorentz model, the radius of an embedding lifted by $\exp^\kappa_{\mathbf{o}}$ grows with the norm of its Euclidean feature (Sec. 3). We therefore examine whether the norms of a pretrained slide encoder already carry lesion information, comparing the L2 norm of each final patch embedding of the original TITAN (Ding et al., 2025) with tile-level tumor annotations (Fig. 2). Interestingly, tumor tiles consistently have larger norms across CAMELYON16, CAMELYON17, and PAIP, and the per-slide norm AUC exceeds 0.5 on 75.1% of slides on average. Because TITAN received no lesion-level supervision, this trend is consistent with the hypothesis that pretraining assigns larger norms to tissue that departs from the dominant slide background, such as dense and atypical tumor cells, and smaller norms to repetitive patterns of normal tissue. This ordering matches the hierarchy that we aim to learn, in which normal tissue lies close to the origin as a generic concept and lesions lie farther out as more specific ones. The cue is, however, weak and uneven: the tumor and normal distributions overlap substantially, which leaves room for fine-tuning to sharpen it.

4.2 Slide–Caption Alignment with Einstein Midpoint Aggregation

Text-conditioned Einstein midpoint. Based on the observation from Sec. 4.1, we now organize these local visual representations with textual semantics in the shared hyperbolic space. Specifically, whereas a slide embedding summarizes thousands of patches, each caption typically pertains to only a subset of the tissue. We therefore ground each caption in its relevant visual evidence by constructing a text-conditioned Einstein midpoint (Ungar, 2005), which aggregates patch embeddings weighted by their similarity to the caption, following Jang et al. (2026). The weighted Einstein midpoint $\boldsymbol{\mu}_\mathbf{t}$ is computed in the Lorentz model as

$$\boldsymbol{\mu}_\mathbf{t} = \frac{\mathbf{h}_\mathbf{t}}{\sqrt{-\kappa\langle \mathbf{h}_\mathbf{t}, \mathbf{h}_\mathbf{t}\rangle_\mathbb{L}}}, \quad \mathbf{h}_\mathbf{t} = \sum_{j=1}^{T} \mathrm{softmax}_j\left(\frac{a_j(\mathbf{t})}{\tau_{\mathrm{em}}}\right)\mathbf{x}_j, \tag{4}$$

where $a_j(\mathbf{t})$ denotes the cosine similarity between patch j and caption $\mathbf{t}$, and τ_{em} is a fixed temperature. By assigning greater weight to caption-relevant patches, this aggregation produces a caption-specific visual representation $\boldsymbol{\mu}_\mathbf{t}$ on the hyperbolic manifold. Thus, the visual representation can correspond to either the global slide embedding $\mathbf{s}$ or its caption-conditioned visual evidence $\boldsymbol{\mu}_\mathbf{t}$.

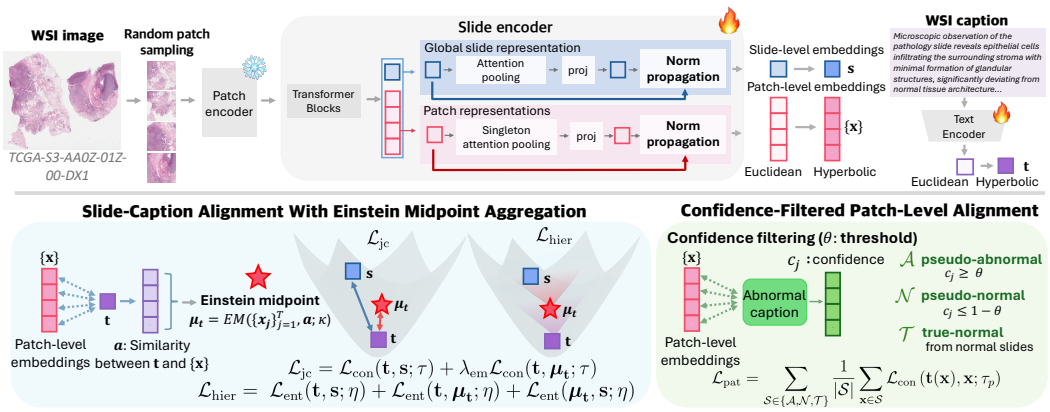

Figure 3: **Overview of HYPERION.** Slide, patch, and caption features of a pretrained slide encoder are lifted to the Lorentz model. Each caption is aligned with the slide embedding and with a text-conditioned Einstein midpoint of the patches (Sec. 4.2), confident caption–patch pairs supervise individual patches (Sec. 4.3), and block-output norms set the patch radius (Sec. 4.4).

Contrastive alignment. To establish visual–semantic correspondence at both global and local levels, we jointly align the slide embedding, its text-conditioned Einstein midpoint and its corresponding caption. While slide–caption alignment captures global semantic correspondence, midpoint–caption alignment encourages the caption to be associated with its most relevant visual evidence. Using the hyperbolic contrastive loss in Eq. (2), we formulate the joint contrastive loss as

$$\mathcal{L}_{\text{jc}} = \mathcal{L}_{\text{con}}(\mathbf{t}, \mathbf{s}; \tau) + \lambda_{\text{em}} \mathcal{L}_{\text{con}}(\mathbf{t}, \boldsymbol{\mu}_\mathbf{t}; \tau), \tag{5}$$

where λ_{em} balances global and local alignment. To avoid false negatives, other captions and midpoints associated with the same slide are excluded from the negative candidate sets. The contrastive objective is computed over abnormal slides to avoid ambiguity arising from generic normal captions.

Entailment guidance. While contrastive alignment establishes visual–semantic correspondence, it does not explicitly capture hierarchy among the aligned representations. We therefore impose *caption* $\supset$ *midpoint* $\supset$ *slide*, from a general concept to its visual evidence and a specific slide instance. Using the entailment loss in Eq. (3), we enforce parent–child hierarchical relationships:

$$\mathcal{L}_{\text{hier}} = \mathcal{L}_{\text{ent}}(\mathbf{t}, \mathbf{s}; \eta) + \mathcal{L}_{\text{ent}}(\mathbf{t}, \boldsymbol{\mu}_\mathbf{t}; \eta) + \mathcal{L}_{\text{ent}}(\boldsymbol{\mu}_\mathbf{t}, \mathbf{s}; \eta). \tag{6}$$

The ordering follows specificity rather than spatial extent: a caption is a concept shared by many slides, its midpoint is that concept's evidence within one slide, and the full slide embedding is the most specific instance, which additionally encodes concepts absent from the caption. Together with contrastive alignment, the hyperbolic learning objective is

$$\mathcal{L}_{\text{hyp}} = \mathcal{L}_{\text{jc}} + \lambda_{\text{ent}} \mathcal{L}_{\text{hier}}, \tag{7}$$

4.3 Confidence-Filtered Patch-Level Alignment

Text-conditioned Einstein midpoints aggregate caption-relevant visual evidence but do not explicitly align individual patches with captions. We therefore use confident caption–patch pairs as pseudo-labels for patch-level contrastive learning.

Confidence filtering. For an abnormal caption $\mathbf{t}$, we first spatially smooth the patch similarities $a_j(\mathbf{t})$ from Eq. (4) using a local averaging window to suppress isolated responses. The smoothed similarities $\bar{a}_j$ are then normalized within each slide to obtain confidence scores, from which we select pseudo-abnormal and pseudo-normal patches:

$$c_j = \frac{\bar{a}_j - \min_k \bar{a}_k}{\max_k \bar{a}_k - \min_k \bar{a}_k + \epsilon}, \quad \mathcal{A} = \{\mathbf{x}_j : c_j \geq \theta\}, \quad \mathcal{N} = \{\mathbf{x}_j : c_j \leq 1 - \theta\}, \tag{8}$$

where ϵ ensures numerical stability and $\theta \in (0.5, 1)$ controls the confidence threshold. Here, $\mathcal{A}$ and $\mathcal{N}$ denote patches that are confidently associated and confidently unassociated with the abnormal caption, respectively, while $\mathcal{T}$ comprises true-normal patches from normal slides. To prevent excessive sampling, we randomly subsample $\mathcal{A}$ and $\mathcal{N}$ to fixed fractions of the slide's patches.

Patch-level contrastive loss. Using the confidence-filtered patches, we extend visual–semantic alignment to individual patch embeddings. Pseudo-abnormal patches are paired with their corresponding abnormal captions, while pseudo-normal and true-normal patches are paired with normal captions. Let $\mathbf{t}(\mathbf{x}) \in \{\mathbf{t}^a, \mathbf{t}^n\}$ denote the caption assigned to a selected patch $\mathbf{x}$, where $\mathbf{t}^a$ is the pseudo-abnormal caption, and $\mathbf{t}^n$ the (pseudo-)normal caption. Using the hyperbolic contrastive loss in Eq. (2), we define the patch-level objective as

$$\mathcal{L}_{\text{pat}} = \sum_{\mathcal{S}\in\{\mathcal{A},\mathcal{N},\mathcal{T}\}} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} \mathcal{L}_{\text{con}}\left(\mathbf{t}(\mathbf{x}), \mathbf{x}; \tau_p\right), \tag{9}$$

where τ_p is the patch-level temperature. This objective provides direct patch-level supervision.

Total objective. The overall training objective combines hyperbolic visual–semantic alignment and hierarchy supervision with confidence-filtered patch-level contrastive learning:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{hyp}} + \lambda_{\text{pat}}\mathcal{L}_{\text{pat}}, \tag{10}$$

where λ_{pat} balances the contribution of patch-level supervision. $\mathcal{L}_{\text{hyp}}$ promotes visual–semantic alignment and radial organization, while $\mathcal{L}_{\text{pat}}$ strengthens fine-grained caption–patch alignment. At inference, localization is performed by comparing the Lorentz geodesic distances between patch embeddings and tumor or normal text prototypes.

4.4 Radial hierarchy encoding via intermediate norm propagation

In hyperbolic space, radial distance naturally encodes semantic specificity, with more specific representations lying farther from the origin. However, pretrained slide encoders can obscure this signal through repeated normalization, as patch features are layer-normalized after the transformer blocks and again in the attentional pooler. Because the hyperbolic radius depends on the norm of the lifted feature, these operations prevent contrastive and entailment objectives from exploiting the norm of pre-normalization patch features to encode radial specificity.

We therefore introduce **intermediate norm propagation**, which combines the direction of the pooled patch embedding with the magnitude of the last transformer block output. Unlike the final embedding, this intermediate feature is measured before the attentional pooler's LayerNorm and projection, allowing its norm to remain trainable and encode radial specificity. Let $\mathbf{f}_j$ denote the last-block feature of patch j. We redefine the patch embedding as

$$\mathbf{x}_j \longmapsto \exp_{\mathbf{o}}^{\kappa}\left(\gamma \frac{\|\mathbf{f}_j\|_2}{\bar{r}} r(\mathbf{s}) \frac{\log_{\mathbf{o}}^{\kappa}(\mathbf{x}_j)}{r(\mathbf{x}_j)}\right), \tag{11}$$

where $\bar{r}$ is the running mean of block-output token norms, $\gamma > 0$ is a fixed scaling factor, and $r(\cdot)$ denotes the hyperbolic radius defined in Sec. 3. Using running statistics rather than slide-wise normalization preserves magnitude differences across slides. This allows the contrastive and entailment objectives to shape radial specificity through intermediate transformer representations.

5 Experiments

5.1 Experimental Setup

Baselines. In this work, we fine-tune TITAN (Ding et al., 2025), one of the representative foundational slide encoders for pathology. We compare with patch-level vision–language models like CONCH (Lu et al., 2024), PathGen-CLIP (Sun et al., 2025), and slide-level models such as CHIEF (Wang et al., 2024) with its supervised tumor head, PRISM (Shaikovski et al., 2024), and TITAN. We also consider Prov-GigaPath (Xu et al., 2024) for classification.

Datasets. HYPERION is fine-tuned on TCGA (Weinstein et al., 2013) and REG2025 (REG2025, 2025) slides with captions and reports and on TCGA normal-tissue slides. Segmentation uses the tumor-containing slides of CAMELYON16 (CM16) (Bejnordi et al., 2017), CAMELYON17 (CM17) (Bandi et al., 2019), PAIP Prostate, PAIP Colon (Kang et al., 2021), and the five-class Radboud subset of PANDA (Bulten et al., 2022). Report generation uses WSI-Bench (Liang et al., 2025), and classification uses 36 benchmarks with labels from Patho-Bench (Zhang et al., 2025), grouped into diagnosis (9), biomarker status (9), and mutation prediction (18).

Table 1: **Segmentation performance on five datasets.** CM16 and CM17 denote CAMELYON16 and CAMELYON17. For the four binary datasets, Dice and intersection over union (IoU) are computed per slide and averaged across slides. For the five-class PANDA dataset, mean Dice (mDice) and mean IoU (mIoU) are averages of class-wise scores computed from the global confusion matrix. Best results are bold and second-best results are underlined across all methods with reported results. – denotes an unreported result.

Encoding level	Method	CM16		CM17		PAIP Prostate		PAIP Colon		PANDA	
		Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	mDice (↑)	mIoU (↑)
Patch-level	CONCH (patch) (Lu et al., 2024)	0.155	0.103	0.112	0.072	0.291	0.194	0.503	0.364	<u>0.293</u>	<u>0.185</u>
	PathGen-CLIP (Sun et al., 2025)	<u>0.265</u>	<u>0.189</u>	<u>0.250</u>	<u>0.188</u>	0.263	0.170	<u>0.586</u>	<u>0.440</u>	0.263	0.154
Slide-level	CHIEF (Wang et al., 2024)	0.146	0.103	0.085	0.054	<u>0.402</u>	<u>0.287</u>	0.571	0.430	–	–
	PRISM (Shaikovski et al., 2024)	0.231	0.165	0.156	0.106	0.284	0.190	0.521	0.385	0.067	0.036
	TITAN (Ding et al., 2025)	0.143	0.098	0.098	0.063	0.248	0.162	0.563	0.422	0.201	0.114
	HYPERION (Ours)	**0.463**	**0.362**	**0.404**	**0.318**	**0.773**	**0.658**	**0.783**	**0.691**	**0.411**	**0.271**

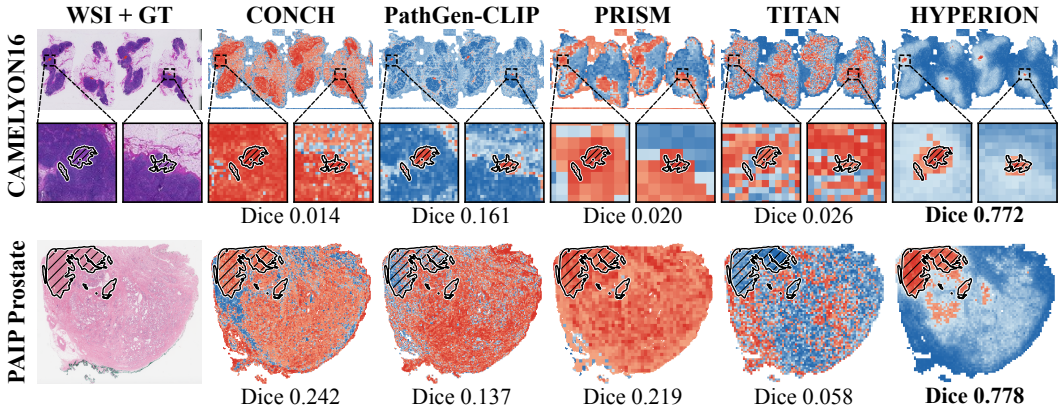

Figure 4: **Qualitative zero-shot tumor segmentation.** Predicted tumor (red) and normal (blue) tiles on a CM16 slide with small metastases (top; boxed regions are magnified below each panel) and a PAIP Prostate slide (bottom). Tumor annotations are hatched with a black outline (filled red in the CM16 overview); more slides are in App. E.1.

5.2 Zero-shot Tumor Segmentation

We first evaluate whether HYPERION improves fine-grained tumor localization without task-specific supervision. As shown in Tab. 1, HYPERION achieves the best Dice and IoU on all five benchmarks, substantially improving over its TITAN backbone, e.g., from 0.143 to 0.463 Dice on CM16. Notably, the improvement is particularly pronounced for small tumor regions: on the slides of the same dataset where tumor occupies less than 2% of the tiles, HYPERION achieves a Dice of 0.246, compared with at most 0.089 for the baselines (Tab. A15). This advantage is also evident qualitatively from Fig. 4, where HYPERION accurately localizes small metastatic foci that are largely missed or only weakly captured by competing methods, highlighting its ability to preserve fine-grained pathological evidence within gigapixel WSIs.

5.3 WSI-level MLLM Evaluation

To examine whether HYPERION-refined representations remain useful for downstream language generation, we integrate them into a WSI-level MedGemma MLLM (Sellergren et al., 2025). Following the WSI-LLaVA training pipeline, slide-level patch features are converted into 256 visual tokens using a learned-query cross-attention resampler, with HYPERION- and TITAN-based models sharing the same architecture and training scheme (App. A.7). On the 208 report-generation questions of WSI-Bench (Tab. 2), HYPERION features improve all six lexical metrics, including BLEU-1–4, ROUGE-L, and METEOR, over TITAN features, showing that the refined representations transfer to language generation without sacrificing slide-level semantic information. More importantly, Fig. 5 shows improved visual grounding: in the illustrated CM16 macrometastasis case,

Table 2: **Report generation on WSI-Bench (Liang et al., 2025)** (↑). Upper rows are from WSI-LLaVA; the MedGemma models differ only in the slide encoder (bold represents the best).

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
GPT-4o (Hurst et al., 2024)	0.202	0.069	0.030	0.016	0.132	0.167
Quilt-LLaVA (Seyfioglu et al., 2024)	0.474	0.351	0.282	0.236	0.475	0.460
MI-Gen (Chen et al., 2024a)	0.403	0.306	0.248	0.209	0.446	0.407
Hist-Gen (Guo et al., 2024)	0.406	0.307	0.248	0.208	0.448	0.416
WSI-LLaVA (Liang et al., 2025)	0.480	0.358	0.287	0.240	0.490	0.465
MedGemma+TITAN	0.507	0.385	0.312	0.264	0.504	0.483
MedGemma+**HYPERION (Ours)**	**0.514**	**0.390**	**0.317**	**0.269**	**0.506**	**0.487**

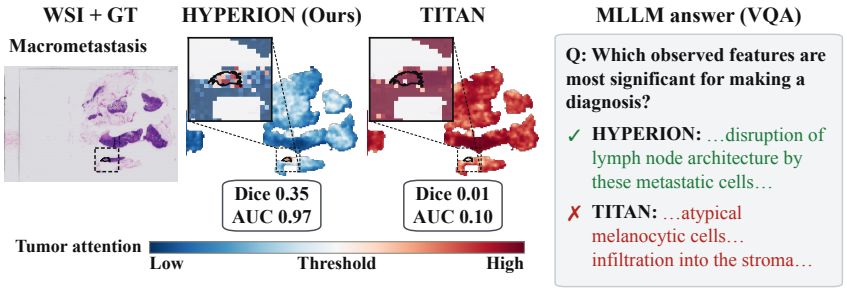

Figure 5: **Visual grounding inside the MLLM.** A CM16 slide with a 2.0 mm macrometastasis, same MedGemma model with HYPERION or TITAN features. Tumor attention = Attention lens (Jiang et al., 2025) of the visual tokens (tumor minus normal word log-odds) mapped to tiles and normalized to measure Dice and AUC score. The inset enlarges the dashed box (5 mm wide); black line = tumor annotation. Right: a VQA item with both answers (correct, incorrect). A micrometastasis and two macrometastasis examples are in App. E.2.

the HYPERION-based model concentrates tumor evidence on the annotated lesion region (AUC 0.97, Dice 0.35), whereas the TITAN-based counterpart shows substantially weaker localization (AUC 0.10, Dice 0.01), with its VQA response also reflecting less pathology-consistent evidence. Additional micro- and macrometastasis examples in App. E.2 exhibit similar qualitative trends.

5.4 Slide-level Classification

We further evaluate whether introducing fine-grained visual–semantic grounding comes at the expense of the slide-level knowledge already encoded by the pretrained foundation model. Under linear probing (Tab. 3), HYPERION largely preserves this predictive capability while achieving improved overall performance. A similar trend is observed in zero-shot classification, where HYPERION improves the overall AUC from 0.638 for TITAN to 0.661. These results suggest that the proposed hyperbolic fine-tuning largely preserves slide-level predictive performance on average with localization ability; that is, it retains the existing predictive information while yielding an overall improvement across diverse downstream prediction tasks (App. D.2).

5.5 Ablation Studies

Tab. 4 evaluates the contribution of each component of HYPERION. Eliminating the hyperbolic formulation reduces the overall Dice from 0.606 to 0.501, indicating the benefit of organizing visual and text representations in a hierarchical geometry. The entailment loss enforces the caption–midpoint–slide hierarchy, whereas the midpoint contrastive term aligns captions with their visual evidence. Removing them reduces the overall Dice to 0.526 and 0.540, respectively. Removing the confidence-filtered patch-level objective $\mathcal{L}_{\text{pat}}$, which provides caption–patch supervision, causes the largest drop in overall Dice to 0.399, highlighting the importance of fine-grained visual grounding. Finally, removing intermediate norm propagation, which transfers feature magnitudes to the hyperbolic radius for radial hierarchy encoding, reduces the overall Dice to 0.489, with large degradation on PAIP Prostate. Together, these results show that patch-level supervision, radial norm encoding, and hierarchical visual–semantic alignment provide complementary contributions to HYPERION.

Table 3: **Slide-level classification on 36 benchmarks.** AUC and support-weighted F1 averaged over tasks and folds; Overall averages the groups. Best in bold, second-best underlined per setting.

Evaluation Setting	Method	Diagnosis (9 tasks)		Biomarker (9 tasks)		Mutation (18 tasks)		Overall (36 tasks)	
		AUC (↑)	F1 (↑)	AUC (↑)	F1 (↑)	AUC (↑)	F1 (↑)	AUC (↑)	F1 (↑)
Linear Probing	Prov-GigaPath (Xu et al., 2024)	0.8445	0.6877	0.6998	0.5967	0.6497	0.6856	0.7313	0.6567
	CHIEF (Wang et al., 2024)	0.9338	0.8059	<u>0.7428</u>	0.6332	0.7060	<u>0.7167</u>	0.7942	0.7186
	PRISM (Shaikovski et al., 2024)	0.9557	0.8313	0.7107	0.6127	0.6974	0.7028	0.7879	0.7156
	TITAN (Ding et al., 2025)	<u>0.9563</u>	<u>0.8528</u>	0.7427	<u>0.6387</u>	<u>0.7101</u>	0.7166	<u>0.8030</u>	<u>0.7360</u>
	HYPERION (Ours)	**0.9575**	**0.8596**	**0.7688**	**0.6508**	**0.7163**	**0.7244**	**0.8142**	**0.7449**
Zero-shot	PRISM (Shaikovski et al., 2024)	<u>0.7982</u>	<u>0.5036</u>	0.4564	0.3540	**0.5454**	<u>0.3410</u>	0.6000	<u>0.3995</u>
	TITAN (Ding et al., 2025)	0.7784	0.4235	<u>0.6135</u>	<u>0.4098</u>	0.5211	0.3321	<u>0.6376</u>	0.3885
	HYPERION (Ours)	**0.8043**	**0.5730**	**0.6372**	**0.4299**	<u>0.5419</u>	**0.4686**	**0.6611**	**0.4905**

Table 4: **Component ablation.** Each row removes one component from the full objective. Dice and IoU are computed per slide and averaged across slides. Overall scores are computed as unweighted averages across the four datasets. Best results are bold and second-best results are underlined.

Ablation	Setting	CM16		CM17		PAIP Prostate		PAIP Colon		Overall	
		Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)
Default	**HYPERION (Ours)**	**0.463**	**0.362**	**0.404**	**0.318**	**0.773**	**0.658**	**0.783**	**0.691**	**0.606**	**0.507**
Removal	w/o hyperbolic space	0.271	0.209	0.286	0.239	0.700	0.596	0.746	0.635	0.501	0.420
	w/o entailment ($\mathcal{L}_{\text{hier}}$)	0.364	0.296	0.306	0.264	<u>0.734</u>	<u>0.626</u>	0.701	0.610	0.526	<u>0.449</u>
	w/o midpoint contrastive	<u>0.408</u>	<u>0.320</u>	<u>0.375</u>	<u>0.306</u>	0.689	0.585	0.688	0.581	<u>0.540</u>	0.448
	w/o $\mathcal{L}_{\text{pat}}$	0.296	0.235	0.238	0.188	0.421	0.315	0.641	0.504	0.399	0.311
	w/o norm propagation	0.330	0.251	0.256	0.193	0.613	0.500	<u>0.756</u>	<u>0.646</u>	0.489	0.398

6 Discussion

Fig. 6 shows how the norm ratio $\|\mathbf{f}_j\|_2/\bar{r}$ from the last transformer block outputs separates the pseudo-abnormal set $\mathcal{A}$ from the pseudo-normal set $\mathcal{N}$ during fine-tuning. Pseudo-abnormal patches initially have smaller norm ratios (AUC 0.35), but with norm propagation, this ordering reverses near the end of the first epoch and the AUC reaches over 0.75, despite no loss directly supervising the norms (Eq. (10)). While other components also weaken the final separation, the AUC based on pseudo-labels falls to 0.00 without norm propagation. Consistently, removing norm propagation reduces overall Dice from 0.606 to 0.489 (Tab. 4), showing that norm propagation is critical for exposing informative feature magnitudes to the hyperbolic objective and turning them into discriminative lesion cues.

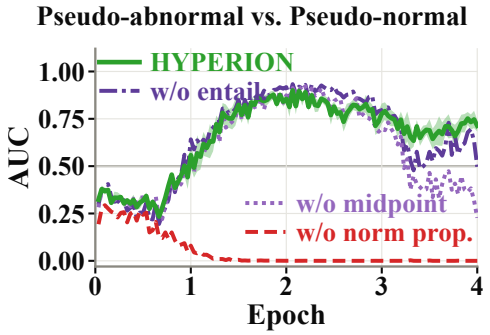

Figure 6: **Norm separation of pseudo-labeled patches during fine-tuning.** AUC of the norm ratio (Eq. (11)) for $\mathcal{A}$ vs. $\mathcal{N}$ (Eq. (8)).

7 Conclusion

We introduced HYPERION, a hyperbolic fine-tuning framework that organizes pretrained slide, patch, and text representations into a shared visual–semantic hierarchy while preserving the knowledge already encoded by the foundation model. HYPERION combines text-conditioned Einstein midpoint aggregation, hyperbolic contrastive and entailment objectives, confidence-filtered patch-level supervision, and intermediate norm propagation to ground fine-grained visual evidence while exploiting feature magnitude as a radial cue. We further showed that pretrained feature norms already contain pathological spatial information that HYPERION strengthens during fine-tuning. Across zero-shot tumor segmentation, slide-level prediction, and WSI-level multimodal language modeling, HYPERION improves fine-grained localization while preserving or enhancing slide-level predictive knowledge and providing more spatially grounded evidence for downstream language generation.

AI USE STATEMENT

We used generative AI tools as supporting utilities for programming, checking mathematical derivations, managing experiments, and editing the manuscript. Generative AI was not involved in formulating the research ideas presented in this work. All AI-assisted outputs, including generated code, were reviewed and validated by the authors. The authors retain full responsibility for the accuracy and integrity of the final work, including its text, claims, code, and other AI-assisted artifacts.

REPRODUCIBILITY STATEMENT

We provide the model formulation, data construction and preprocessing, prompts, evaluation protocols, and complete training hyperparameters in the main text and appendix. We also specify the datasets and splits, model initialization, optimization settings, and computational resources used for HYPERION and the downstream WSI-level MLLM. Classification experiments use fixed stratified folds, with per-benchmark results and variations across folds reported in the appendix. To facilitate reproducibility, we will release the implementation, training and evaluation scripts, and configuration files used to produce the reported results.

REFERENCES

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *CVPR*, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

P. Bandi et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *TMI*, 2019.

Babak Ehteshami Bejnordi, Mitko Veta, Paul J Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 2017.

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28(1):154–163, 2022.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 2019.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *NeurIPS*, 2019.

Pingyi Chen, Honglin Li, Chenglu Zhu, Sunyi Zheng, Zhongyi Shui, and Lin Yang. Wsicaption: Multiple instance generation of pathology reports for gigapixel whole-slide images. In *MICCAI*, 2024a.

Pingyi Chen, Chenglu Zhu, Sunyi Zheng, Honglin Li, and Lin Yang. Wsi-vqa: Interpreting whole slide images by generative visual question answering. In *ECCV*, 2024b.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *CVPR*, 2022.

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), 2024c.

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 *CVPR*, 2025.

Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 2018.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, 2023.

Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. Embedding text in hyperbolic spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, 2018.

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*, 2025.

Nathan J. Edwards, Mauricio Oberti, Ratna R. Thangudu, Shuang Cai, Peter B. McGarvey, Shine Jacob, Subha Madhavan, and Karen A. Ketchum. The CPTAC data portal: A resource for cancer proteomics research. *Journal of Proteome Research*, 14(6):2707–2713, 2015.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *ICML*, 2018.

Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, and Nando de Freitas. Hyperbolic attention networks. In *International Conference on Learning Representations*, 2019.

Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Liansheng Wang, and Hao Chen. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *MICCAI*, 2024.

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.

Peixiang Huang, Yanyan Huang, Weiqin Zhao, Junjun He, and Lequan Yu. Hyperpath: Knowledge-guided hyperbolic semantic hierarchy modeling for wsi analysis. In *MICCAI*, 2025.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9), 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. In *NeurIPS*, 2023.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, 2018.

Ji Ha Jang, Hayeon Kim, Chulwon Lee, Junghun James Kim, and Se Young Chun. Hyfl-clip: Hyperbolic fine-tuning of clip for robust long-context understanding. In *ECCV*. Springer, 2026.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *CVPR*. IEEE, 2025.

Yunsook Kang, Y. Kim, Seongkeun Park, Gun Ro, C. Hong, Hyungjoon Jang, Sungduk Cho, Wonjae Hong, D. Kang, Jonghoon Chun, Kyoungbun Lee, G. Kang, et al. Development and operation of a digital platform for sharing pathology image data. *BMC Medical Informatics and Decision Making*, 21, 2021.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *CVPR*, 2020.

Hayeon Kim, Ji Ha Jang, Junghun James Kim, and Se Young Chun. Uncertainty-guided compositional alignment with part-to-whole semantic representativeness in hyperbolic vision-language models. In *CVPR*, 2026.

Marc Law, Renjie Liao, Jake Snell, and Richard Zemel. Lorentzian distance learning for hyperbolic representations. In *International Conference on Machine Learning*, pp. 3672–3681, 2019.

Matthew Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In *ACL*, 2019.

Byung Hyun Lee, Wongi Jeong, Woojae Han, Kyoungbun Lee, and Se Young Chun. Continual multiple instance learning with enhanced localization for histopathological whole slide image analysis. In *ICCV*, 2025.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *CVPR*, 2021.

Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, et al. Wsi-llava: A multimodal large language model for whole slide image. In *ICCV*, 2025.

Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In *NeurIPS*, volume 32, 2019.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6), 2021.

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), 2024.

Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13), 2018.

Pedro C. Neto, Diana Montezuma, Sara P. Oliveira, Domingos Oliveira, João Fraga, Ana Monteiro, João Monteiro, Liliana Ribeiro, Sofia Gonçalves, Stefan Reinhard, et al. An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. *npj Precision Oncology*, 8:56, 2024.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *NeurIPS*, 2017.

Avik Pal, Max Van Spengler, Guido D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *ICLR*, 2025.

REG2025. Report generation in pathology using pan-asia giga-pixel wsis. MICCAI 2025 Grand Challenge, 2025. URL `https://reg2025.grand-challenge.org`.

Thomas Roetzer-Pejrimovsky, Anna-Christina Moser, Baran Atli, Clemens Christian Vogel, Petra A. Mercea, Romana Prihoda, Ellen Gelpi, Christine Haberler, Romana Höftberger, Johannes A. Hainfellner, et al. The Digital Brain Tumour Atlas, an open histopathology resource. *Scientific Data*, 9:55, 2022.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *ICML*, 2018.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

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 *CVPR*, 2024.

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.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *NeurIPS*, 2021.

Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. Learning structured representations with hyperbolic embeddings. *NeurIPS*, 2024.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Jingxiong Li, Xuan Gong, Xinheng Lyu, Tao Lin, et al. Pathgen-1.6m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In *ICLR*, 2025.

Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *CVPR*, 2024.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. In *ICLR*, 2019.

Abraham A. Ungar. *Analytic Hyperbolic Geometry: Mathematical Foundations and Applications*. World Scientific, 2005.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2024.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634, 2024.

Jason W. Wei, Laura J. Tafe, Yevgeniy A. Linnik, Louis J. Vaickus, Naofumi Tomita, and Saeed Hassanpour. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific Reports*, 9:3358, 2019.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in Oncology*, 11:759007, 2021.

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, 2024.

Shiyang Yan, Zongxuan Liu, and Lin Xu. Hyp-uml: Hyperbolic image retrieval with uncertainty-aware metric learning. *arXiv preprint arXiv:2310.08390*, 2023.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.

Chen Zhang, Yilu An, Ying Chen, Hao Li, Xitong Ling, Lihao Liu, Junjun He, Yuxiang Lin, Zihui Wang, and Rongshan Yu. Hyperst: Hierarchical hyperbolic learning for spatial transcriptomics prediction. In *CVPR*, 2026a.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *CVPR*, 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. In *AAAI*, 2026b.

Mengdan Zhu, Bing Ren, Ryland Richards, Matthew Suriawinata, Naofumi Tomita, and Saeed Hassanpour. Development and evaluation of a deep neural network for histologic classification of renal cell carcinoma on biopsy and surgical resection slides. *Scientific Reports*, 11:7080, 2021.

A Experimental Setups

A.1 Dataset Details

Training data of the slide encoder. Tab. A1 summarizes the fine-tuning data. TCGA slides are paired with descriptive captions derived from the WSI-Bench (Liang et al., 2025) exclusively splitted into training and test dataset, annotations of each slide (morphology, diagnosis, and staging sentences) and with mutation captions derived from the GDC masked somatic mutation files. Slides belonging to the WSI-Bench test split are excluded from all HYPERION fine-tuning sets, so the encoder never sees a test report sentence. REG2025 slides come from the REG2025 challenge[1], whose cases were collected at six medical centers in Korea, Japan, India, Turkey, and Germany, and are paired with the sentences of their pathology reports, which are fully separated in REG2025 dataset. TCGA normal-tissue slides are paired with a fixed normal caption and serve as the true-normal set $\mathcal{T}$ in Eq. (9); they are excluded from the slide–caption and entailment objectives. In each epoch, every TCGA slide and an equal number of REG2025 slides are sampled once, together with normal slides at a ratio of 4:4:1, and eight captions are drawn per slide.

Table A1: **Training data of HYPERION.** Slides are counted after removing those without extracted features.

Source	Slides	Caption rows	Text
TCGA (30 cancer types)	6,834	115,032	descriptive captions (WSI-Bench)
TCGA mutation captions	5,573	23,421	gene-level mutation summaries (GDC)
REG2025 (7 organs)	8,205	36,690	pathology report sentences
TCGA normal tissue	1,807	–	fixed normal caption

Segmentation benchmarks. Tab. A2 lists the five segmentation benchmarks. We use only slides that contain annotated tumor, since Dice is undefined on slides without tumor. A tile is labeled as tumor when the annotated tumor covers at least 1% of its area. For PANDA, we use the Radboud slides, whose masks distinguish stroma, benign epithelium, and Gleason patterns 3, 4, and 5; these masks were generated semi-automatically (Bulten et al., 2022) and are noisy at the tile level. A tile takes the Gleason pattern with the largest area if any pattern covers at least 1% of it, and otherwise benign epithelium or stroma.

Table A2: **Zero-shot segmentation benchmarks.**

Benchmark	Organ	Slides	Classes
CAMELYON16 (Bejnordi et al., 2017)	lymph node (breast)	48	tumor / normal
CAMELYON17 (Bandi et al., 2019)	lymph node (breast)	49	tumor / normal
PAIP Prostate (Kang et al., 2021)	prostate	600	tumor / normal
PAIP Colon (Kang et al., 2021)	colon	437	tumor / normal
PANDA (Radboud) (Bulten et al., 2022)	prostate	5,060	stroma / benign / G3 / G4 / G5

CAMELYON16 uses the 48 tumor-containing slides of the official test set. CAMELYON17 uses the 49 slides with lesion annotations and extracted features, which contain 16 isolated-tumor-cell, 17 micro-metastasis, and 16 macro-metastasis lesions. PAIP Prostate and PAIP Colon are whole-slide images from three Korean hospitals with tumor areas annotated by pathologists, released by the PAIP data platform (Kang et al., 2021)[2]. One prostate slide could not be read by the CONCH and CHIEF preprocessing, so these two baselines use 599 slides. Analyses that need both tumor and normal tiles in a slide (the norm AUCs of App. C) use the 599 prostate and 435 colon slides that contain both.

Prediction benchmarks. Tab. A3 lists the 36 classification benchmarks. Labels and task definitions come from Patho-Bench (Zhang et al., 2025) for EBRAINS (Roetzer-Pejrimovsky et al., 2022), BCNB (Xu et al., 2021), PANDA, DHMC LUAD (Wei et al., 2019), and all CPTAC (Edwards et al., 2015) tasks. For DHMC RCC (Zhu et al., 2021) and IMP-CRC (Neto et al., 2024), they come from

[1] https://reg2025.grand-challenge.org

[2] http://wisepaip.org

Table A3: **The 36 prediction benchmarks.** Evaluated samples are those with extracted features.

Group	Benchmark	Task	Classes	Unit	Samples	Folds
Diagnosis	EBRAINS Dx (30-cls)	Fine-grained CNS tumor diagnosis	30 WHO diagnoses	slide	2,309	10
	EBRAINS Dx group (12-cls)	Coarse CNS tumor diagnosis group	12 diagnosis groups	slide	2,309	10
	DHMC LUAD	LUAD predominant growth pattern	Acinar/Lepidic/Micropapillary/Papillary/Solid	slide	143*	5
	DHMC RCC	Renal tumor subtype	5 classes	slide	563	10
	CM16 Tumor	Lymph-node metastasis detection (breast)	normal/tumor	slide	399	10
	CM17 pN	Node-level metastasis size category	ITC/macro/micro/negative	slide	499	10
	PANDA ISUP	Prostate biopsy ISUP grade	ISUP 0-5	slide	9,555	10
	IMP-CRC	Colorectal biopsy dysplasia grade	non-neoplastic/low-grade/high-grade	slide	900	10
	CPTAC NSCLC	Lung cancer subtype	LUAD/LSCC	patient	206	10
Biomarker	BCNB ER	ER status	neg/pos	slide	1,058	10
	BCNB HER2	HER2 status	neg/pos	slide	1,058	10
	BCNB PR	PR status	neg/pos	slide	1,058	10
	EBRAINS IDH	IDH mutation status (glioma)	wildtype/mutant	slide	846	10
	CPTAC BRCA Immune	Immune infiltration class	low/medium/high	patient	102	10
	CPTAC LUAD Immune	Immune infiltration class	low/medium/high	patient	108	10
	CPTAC LSCC Immune	Immune infiltration class	low/medium/high	patient	108	10
	CPTAC LSCC Grade	Histologic grade	well (G1)/moderate (G2)	patient	104	10
	CPTAC COAD Immune	Immune infiltration class	low/medium/high	patient	55	10
Mutation	CPTAC GBM EGFR	EGFR mutation	WT/MUT	patient	99	10
	CPTAC GBM TP53	TP53 mutation	WT/MUT	patient	99	10
	CPTAC LUAD EGFR	EGFR mutation	WT/MUT	patient	108	10
	CPTAC LUAD KRAS	KRAS mutation	WT/MUT	patient	102	10
	CPTAC LUAD STK11	STK11 mutation	WT/MUT	patient	108	10
	CPTAC LUAD TP53	TP53 mutation	WT/MUT	patient	108	10
	CPTAC LSCC ARID1A	ARID1A mutation	WT/MUT	patient	108	10
	CPTAC LSCC KEAP1	KEAP1 mutation	WT/MUT	patient	108	10
	CPTAC BRCA PIK3CA	PIK3CA mutation	WT/MUT	patient	102	10
	CPTAC BRCA TP53	TP53 mutation	WT/MUT	patient	102	10
	CPTAC COAD APC	APC mutation	WT/MUT	patient	55	10
	CPTAC COAD KRAS	KRAS mutation	WT/MUT	patient	55	10
	CPTAC COAD TP53	TP53 mutation	WT/MUT	patient	55	10
	CPTAC COAD PIK3CA	PIK3CA mutation	WT/MUT	patient	55	10
	CPTAC COAD ARID1A	ARID1A mutation	WT/MUT	patient	55	9
	CPTAC COAD ACVR2A	ACVR2A mutation	WT/MUT	patient	55	10
	CPTAC COAD SETD1B	SETD1B mutation	WT/MUT	patient	55	10
	CPTAC COAD MSI-H	MSI-high status	MSS/MSI-H	patient	54	10

the dataset releases, and for CAMELYON16 and CAMELYON17 from the official annotations. CPTAC tasks are defined per patient, and the embeddings of all tumor slides of a patient are averaged. We build our own stratified 10-fold splits with a fixed seed; a task uses fewer folds when its rarest class has fewer than ten samples (DHMC LUAD 5 folds, CPTAC COAD ARID1A 9 folds). Nineteen slides without extracted tissue coordinates are excluded for all models.

MLLM data. The MLLM is trained and evaluated on WSI-Bench (Liang et al., 2025), which contains 9,850 WSIs and 179,569 visual question–answer pairs. Tab. A4 lists the training data: the connector is pretrained on 17,081 TCGA slides with report captions, and instruction tuning uses 8,610 WSI-Bench training slides. Report generation is evaluated on the 208 report questions of the WSI-Bench test set, disjoint to the training dataset.

Table A4: **Training data of the WSI-level MLLM.**

Training stage	Samples	Slides	Sample type
Stage 2 (connector pretraining)	17,081	17,081	one report per slide
Stage 3 (LoRA instruction tuning)	154,632	8,610	multiple QA pairs per slide

A.2 Evaluation Protocol

Segmentation. For each tile, a method produces a tumor score and a normal score from the tile embedding and the two text prototypes. HYPERION scores a tile by its Lorentz inner product with each prototype, which gives the same prediction as comparing Lorentz distances, and the baselines use cosine similarity. Each method is scored on the tile grid of its own patch encoder (Tab. A9): 512×512 tiles for HYPERION and TITAN and 224×224 patches for the other methods. The tile ground truth is built on that grid with the rule of App. A.1, so it differs slightly between grids; PRISM assigns one score to each block of 4×4 patches (App. A.5). Dice and IoU are computed for each slide and averaged over slides.

The AUC in App. D.1 ranks the tiles of each slide by the difference between the two scores and is averaged over slides that contain both classes. For PANDA, the prediction is the class with the highest score among five prototypes.

Class-wise Dice and IoU on PANDA are computed from the confusion matrix accumulated over all slides, since most slides contain only some of the five classes, and tiles with less than 5% tissue are excluded.

Classification. Linear probing trains a single linear layer with softmax cross-entropy on frozen slide embeddings. We use Adam with a learning rate of 10^{-3} for 1,000 full-batch steps, without feature normalization or hyperparameter search. All models use the same splits, classifier, and training schedule. For HYPERION, the linear probe uses the Euclidean slide embedding before the Lorentz lift.

Zero-shot classification scores each slide against the class prompts of App. A.3 without training and reuses the test folds of linear probing. TITAN uses the cosine similarity, HYPERION uses the negative Lorentz distance between the lifted slide and prompt embeddings, and PRISM uses its own slide–text similarity with its paraphrase marginalization. We report the AUC (one-vs-rest macro AUC for multi-class tasks), support-weighted F1, and balanced accuracy per fold, and average them over folds. A group score is the mean over the tasks of the group, and the overall score is the mean of the three group scores.

Report generation. Generated reports are compared with the reference reports using BLEU-1 to BLEU-4, ROUGE-L, and METEOR, following WSI-LLaVA (Liang et al., 2025).

A.3 PROMPTS

Segmentation.

Tabs. A5–A7 list the four pairs used in the main results The same prompts are used for every text-based method.

Table A5: **Segmentation prompts for CAMELYON16 and CAMELYON17.**

Tumor	Normal
metastatic carcinoma in a sentinel lymph node	normal sentinel lymph node tissue
lymph node metastasis of breast carcinoma	lymph node without metastatic carcinoma
abundant eosinophilic cytoplasm	scant cytoplasm with a high nuclear density
a sharp border between tumor cells and surrounding lymphocytes	no boundary between different cell populations

Table A6: **Segmentation prompts for PAIP Prostate and PAIP Colon.**

	Tumor	Normal
Prostate	cancerous glands infiltrating between benign glands	benign glands arranged in lobules
	a single layer of malignant cells lining each gland	two cell layers lining each gland
	an absent basal cell layer	a continuous basal cell layer
	a dark crowded area of small glands	evenly spaced pale glands
Colon	glands infiltrating through the muscularis mucosae	crypts confined to the mucosa above the muscularis mucosae
	stratified crowded epithelial cells losing polarity	a single layer of polarized columnar cells
	loss of nuclear polarity along the gland	nuclei aligned at the base of every cell
	a dark irregular mass replacing the mucosa	an orderly row of parallel crypts

Classification. Each class is described by two to four class names. For EBRAINS, DHMC LUAD, BCNB, EBRAINS IDH, and CPTAC NSCLC, the names are taken from the Patho-Bench label definitions. Following TITAN (Ding et al., 2025), TITAN and HYPERION insert each class name into the 23 templates of Tab. A8. TITAN averages the normalized text embeddings, and HYPERION averages the text embeddings without normalization before the Lorentz lift, because normalizing them places all prompts at almost the same hyperbolic radius. PRISM uses the class names without templates, since its text model is a report decoder.

Table A7: **Segmentation prompts for the five PANDA classes.**

Class	Name	Gland size and shape	Basal layer	Negative definition
Stroma	stroma	no epithelial gland profiles	no epithelium and no basal cell layer	supporting tissue only, not epithelium
Benign	benign glands	large glands with papillary infoldings	glands with an intact basal cell layer	glands that are not cancerous
Gleason 3	Gleason 3	small to medium uniform round glands	discrete glands lacking a basal cell layer	cancerous glands that are still individually formed, not fused
Gleason 4	Gleason 4	irregular glands with ill-defined lumens	fused glands lacking a basal cell layer	cancerous glands that are fused or cribriform, not individually formed
Gleason 5	Gleason 5	single cells and cords without lumens	solid tumor without glands or basal cells	cancer that forms no glands at all

Table A8: **Prompt templates for zero-shot classification**, reproduced from the TITAN release.

#	Template	#	Template
1	"CLASSNAME."	13	"shows CLASSNAME."
2	"an image of CLASSNAME."	14	"contains CLASSNAME."
3	"the image shows CLASSNAME."	15	"presence of CLASSNAME."
4	"the image displays CLASSNAME."	16	"CLASSNAME is present."
5	"the image exhibits CLASSNAME."	17	"CLASSNAME is observed."
6	"an example of CLASSNAME."	18	"the pathology image reveals CLASSNAME."
7	"CLASSNAME is shown."	19	"a microscopic image of showing CLASSNAME."
8	"this is CLASSNAME."	20	"histology shows CLASSNAME."
9	"I observe CLASSNAME."	21	"CLASSNAME can be seen."
10	"the pathology image shows CLASSNAME."	22	"the tissue shows CLASSNAME."
11	"a pathology image shows CLASSNAME."	23	"CLASSNAME is identified."
12	"the pathology slide shows CLASSNAME."		

A.4 Training Text Construction

Each training row pairs a slide with one sentence. For TCGA, the sentences are the morphology, diagnosis, and staging statements of the slide's WSI-Bench annotations. Mutation captions state the mutated genes of the slide's case in a fixed sentence form derived from the GDC somatic mutation files, excluding single-nucleotide polymorphisms. For REG2025, the sentences of the pathology report are used as captions. At each step, eight captions are drawn for each abnormal slide from its descriptive and mutation captions, and all of them are treated as positives in Eq. (2), with other captions of the same slide removed from the negatives. Normal slides use the single caption "Normal benign tissue, no tumor identified."

A.5 Baselines

We apply each baseline to segmentation without training, as follows. CONCH (Lu et al., 2024) and PathGen-CLIP (Sun et al., 2025) score each 224×224 patch against the prompts, since they are patch-level vision–language models. TITAN (Ding et al., 2025) and HYPERION run the slide encoder over all patches of the slide and pass each patch token separately through the attention pooler and projection. This gives one embedding per patch in the space of the text encoder; since each token is pooled alone, the class token does not enter the per-patch embedding. PRISM (Shaikovski et al., 2024) encodes each block of 4×4 neighboring tiles as a small slide and assigns its score to all tiles in the block, because its slide encoder does not produce per-tile embeddings aligned with text.

CHIEF (Wang et al., 2024) predicts tumor tiles with its supervised tumor head, weighted by its attention scores, and uses the anatomical site index of the organ. It therefore does not use prompts, and its five-class PANDA Dice is undefined; the PANDA AUC in Tab. A14 ranks tiles by its tumor probability separately for each class. Prov-GigaPath (Xu et al., 2024) has neither a text encoder nor a tumor head and is used only for linear probing. For linear probing, every baseline uses the slide embedding released with the model.

A.6 WSI Preprocessing

Tissue regions are segmented and tiled with a CLAM-based pipeline (Lu et al., 2021) at a spacing of $0.5\,\mu$m/px ($20\times$). Each encoder is tiled at its own tile size (Tab. A9), so the tile coordinates differ

between the 512×512 and 224×224 encoders. TITAN and HYPERION use CONCH v1.5 features of 512×512 tiles together with the tile coordinates. All patch features are extracted once and cached before fine-tuning and evaluation.

Table A9: **Patch encoders used by each model.**

Model	Patch encoder	Tile size	Feature dim.
TITAN, HYPERION	CONCH v1.5	512×512	768
CONCH	CONCH (ViT-B/16)	224×224	512
PathGen-CLIP	PathGen-CLIP	224×224	512
PRISM	Virchow	224×224	2560
CHIEF	CTransPath	224×224	768
Prov-GigaPath	Prov-GigaPath tile encoder	256×256	1536

A.7 Implementation Details

HYPERION training. We initialize HYPERION from TITAN and fine-tune the slide encoder, the slide projection, and the text encoder, while the attention pooler is kept frozen. Patch, slide, and text features are lifted to the Lorentz model with learnable scales, and the curvature κ and the temperature τ are learned, starting from 1.0 and 0.02; κ ends at 1.0025. Normal slides are excluded from the entailment terms. Each training step samples between 256 and 2,304 patches per slide, and the details of the objective are given at the end of this section. We train with AdamW for 4 epochs (120 steps) on 4 H100 GPUs, which takes about 31 minutes. Tab. A10 lists the hyperparameters, and the training arguments are released with the code.

Table A10: **Training hyperparameters of HYPERION.** Symbols follow Sec. 4.

Hyperparameter	Value	Hyperparameter	Value
Optimizer	AdamW	λ_{em}	0.1
Learning rate	5×10^{-5}	τ_{em}	0.75
Schedule	warm-up (5%) + cosine	λ_{ent}	0.1
Weight decay / grad. clip	0.01 / 1.0	η	0.7
Batch size	128 per GPU (512 negatives)	λ_{pat}	0.2
Epochs (steps)	4 (120)	τ_p	0.10
Curvature κ	learned, init. 1.0	θ	0.90
Temperature τ	learned, init. 0.02	Smoothing window	4×4
Captions per slide	8	Sampled fraction of $\mathcal{A}$, $\mathcal{N}$	8%
Patches per slide	256–2,304	γ	0.65

MLLM integration. We follow the two-stage training of WSI-LLaVA (Liang et al., 2025) using MedGemma-4B (Sellergren et al., 2025). Slide-level patch features are mapped to 256 visual tokens by a Qwen-VL (Bai et al., 2023) -style cross-attention with learned queries. The resampler uses query-only RMS normalization with unnormalized key-value features. In Stage 2 (Tab. A4), we freeze the language model and pretrain the entire resampler for 5 epochs with a learning rate of 10^{-3}. In stage 3 (instruction tuning), we jointly fine-tune the resampler and LoRA adapters (Hu et al., 2022) (rank 128, scaling factor 256) of MedGemma-4B for 3 epochs. We use learning rate of 2×10^{-4} for the language-model LoRA adapters and 5×10^{-5} for the resampler. TITAN and HYPERION features are processed with the same architecture, data, and schedule, and reports are generated with greedy decoding.

B Hyperbolic Geometry Details

Hyperbolic representation. We summarize the geometric operations used in our hyperbolic formulation, following MERU (Desai et al., 2023). We consider the Lorentz model $\mathbb{L}^n_\kappa$ equipped with the Minkowski inner product. For two vectors $\mathbf{p}, \mathbf{q} \in \mathbb{R}^{n+1}$, where $\mathbf{p} = (p_0, p_1, \ldots, p_n)$ and

$\mathbf{q} = (q_0, q_1, \ldots, q_n)$, the Lorentzian inner product is defined as

$$\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}} = -p_0 q_0 + \sum_{j=1}^{n} p_j q_j. \tag{12}$$

For a point $\mathbf{z} \in \mathbb{L}_\kappa^n$, its tangent space consists of all vectors that are Lorentz-orthogonal to $\mathbf{z}$:

$$T_{\mathbf{z}}\mathbb{L}_\kappa^n = \left\{ \mathbf{v} \in \mathbb{R}^{n+1} : \langle \mathbf{z}, \mathbf{v} \rangle_{\mathbb{L}} = 0 \right\}. \tag{13}$$

Given a tangent vector $\mathbf{v} \in T_{\mathbf{z}}\mathbb{L}_\kappa^n$, the exponential map moves $\mathbf{v}$ from the tangent space back to the Lorentz manifold:

$$\exp_{\mathbf{z}}^{\kappa}(\mathbf{v}) = \cosh\left(\sqrt{\kappa}\, \|\mathbf{v}\|_{\mathbb{L}}\right) \mathbf{z} + \frac{\sinh(\sqrt{\kappa}\, \|\mathbf{v}\|_{\mathbb{L}})}{\sqrt{\kappa}\, \|\mathbf{v}\|_{\mathbb{L}}} \mathbf{v}. \tag{14}$$

Conversely, the logarithmic map represents a point $\mathbf{p} \in \mathbb{L}_\kappa^n$ as a tangent vector at $\mathbf{z}$. It is given by

$$\log_{\mathbf{z}}^{\kappa}(\mathbf{p}) = \frac{\cosh^{-1}(-\kappa \langle \mathbf{z}, \mathbf{p} \rangle_{\mathbb{L}})}{\sqrt{(\kappa \langle \mathbf{z}, \mathbf{p} \rangle_{\mathbb{L}})^2 - 1}} \operatorname{Proj}_{\mathbf{z}}(\mathbf{p}), \tag{15}$$

where the tangent projection term is

$$\operatorname{Proj}_{\mathbf{z}}(\mathbf{p}) = \mathbf{p} + \kappa \langle \mathbf{z}, \mathbf{p} \rangle_{\mathbb{L}} \mathbf{z}. \tag{16}$$

This projection removes the component of $\mathbf{p}$ that is normal to the tangent space at $\mathbf{z}$, ensuring that $\operatorname{Proj}_{\mathbf{z}}(\mathbf{p}) \in T_{\mathbf{z}}\mathbb{L}_\kappa^n$.

Geodesic distance and entailment cones. The geodesic distance used in Eq. (2) is

$$d_{\mathbb{L}}(\mathbf{p}, \mathbf{q}) = \frac{1}{\sqrt{\kappa}} \cosh^{-1}(-\kappa \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}}). \tag{17}$$

For the entailment loss in Eq. (3), the half-aperture of the cone at $\mathbf{p}$ and the exterior angle between the cone axis at $\mathbf{p}$ and the geodesic to $\mathbf{q}$ are

$$\omega(\mathbf{p}) = \sin^{-1}\left(\frac{2K}{\sqrt{\kappa}\, \|\mathbf{p}_{1:n}\|_2} \right), \qquad \phi(\mathbf{p}, \mathbf{q}) = \cos^{-1}\left(\frac{q_0 + p_0\, \kappa \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}}}{\|\mathbf{p}_{1:n}\|_2 \sqrt{(\kappa \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}})^2 - 1}} \right), \tag{18}$$

where $K = 0.1$ defines a small region around the origin in which the aperture is not defined. A child $\mathbf{q}$ satisfies the cone of $\mathbf{p}$ when $\phi(\mathbf{p}, \mathbf{q}) \leq \eta\, \omega(\mathbf{p})$.

To compute the Einstein midpoint, we further make use of the projection between the Lorentz and Klein models. Let $\mathbf{p} = (p_0, p_1, \ldots, p_n) \in \mathbb{L}_\kappa^n$, where p_0 denotes the time-like coordinate and $(p_1, \ldots, p_n)$ denotes the spatial coordinates. The Lorentz-to-Klein projection is defined as

$$\Pi_{\mathbb{L} \to \mathbb{K}}(\mathbf{p}) = \frac{(p_1, \ldots, p_n)}{p_0}. \tag{19}$$

Using the Lorentz hyperboloid constraint, the same projection can be written as

$$\Pi_{\mathbb{L} \to \mathbb{K}}(\mathbf{p}) = \frac{(p_1, \ldots, p_n)}{\sqrt{\frac{1}{\kappa} + \sum_{j=1}^{n} p_j^2}}. \tag{20}$$

The inverse mapping lifts a Klein point $\mathbf{k} = (k_1, \ldots, k_n)$ back to the Lorentz manifold:

$$\Pi_{\mathbb{K} \to \mathbb{L}}(\mathbf{k}) = \frac{(1, k_1, \ldots, k_n)}{\sqrt{\kappa \left(1 - \sum_{j=1}^{n} k_j^2\right)}}. \tag{21}$$

These projections provide a convenient way to perform aggregation in the Klein model and subsequently map the result back to the Lorentz hyperboloid. In particular, we use them to compute the Einstein midpoint while maintaining the underlying hyperbolic geometry.

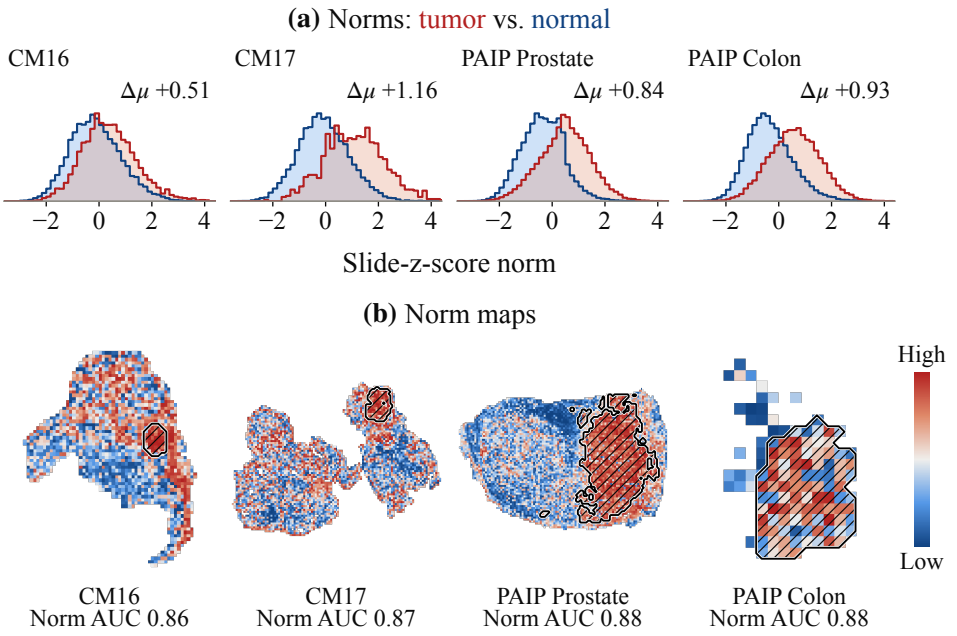

Figure A1: **Patch-embedding norms of the original TITAN on four benchmarks.** (a) Slide-standardized norms of tumor and normal tiles on the slides with a per-slide norm AUC of at least 0.8, as in Fig. 2(a) (8 of 48 CAMELYON16, 7 of 49 CAMELYON17, 14 of 599 PAIP Prostate, and 17 of 435 PAIP Colon slides; $\Delta\mu$: difference of their means). (b) Norm maps of one slide per benchmark, chosen among the five highest norm AUCs with at least 50 tumor tiles, so they illustrate the cue where it is present rather than a typical slide; the tumor annotation is hatched with a black outline.

Einstein midpoint. In the Klein model, the Einstein midpoint of points $\mathbf{k}_1, \ldots, \mathbf{k}_T$ with nonnegative weights $w_1, \ldots, w_T$ is the weighted average

$$\mathrm{EM}(\{\mathbf{k}_j\}, \{w_j\}) = \frac{\sum_{j=1}^{T} w_j \, \zeta_j \, \mathbf{k}_j}{\sum_{j=1}^{T} w_j \, \zeta_j}, \qquad \zeta_j = \frac{1}{\sqrt{1 - \|\mathbf{k}_j\|_2^2}}, \tag{22}$$

where ζ_j is the Lorentz factor (Ungar, 2005). Points far from the origin receive larger factors, so the midpoint respects the hyperbolic geometry rather than the Euclidean geometry of the coordinates. For a Lorentz point $\mathbf{x}_j$ with Klein coordinates $\mathbf{k}_j = \Pi_{\mathbb{L}\to\mathbb{K}}(\mathbf{x}_j)$, the hyperboloid constraint gives $\zeta_j = \sqrt{\kappa}\, x_{j,0}$ and hence $\zeta_j \mathbf{k}_j = \sqrt{\kappa}\,(x_{j,1}, \ldots, x_{j,n})$. Substituting these into Eq. (22) shows that the Klein midpoint equals $(h_1, \ldots, h_n)/h_0$ for $\mathbf{h} = \sum_j w_j \mathbf{x}_j$. Mapping it back with $\Pi_{\mathbb{K}\to\mathbb{L}}$ gives

$$\Pi_{\mathbb{K}\to\mathbb{L}}\left(\frac{(h_1, \ldots, h_n)}{h_0}\right) = \frac{\mathbf{h}}{\sqrt{-\kappa \langle \mathbf{h}, \mathbf{h} \rangle_{\mathbb{L}}}}, \tag{23}$$

which is Eq. (4) with $w_j = \mathrm{softmax}_j(a_j(\mathbf{t})/\tau_{\mathrm{em}})$. The midpoint can therefore be computed directly in the Lorentz model, without converting to Klein coordinates. This closed form is the Lorentzian centroid of Law et al. (2019), and weighting it with attention scores follows the hyperbolic attention of Gulcehre et al. (2019).

C Additional Analyses

C.1 Lesion Cues in the Norms of the Pretrained TITAN

Sec. 4.1 reports that the patch-embedding norms of the original all four benchmarks. Tab. A11 gives the complete statistics, and Fig. A1 shows the norm distributions and one norm map per benchmark. Although TITAN is trained with cosine-based alignment, its pretrained feature norms already encode lesion-related cues.

Table A11: **Norm of the final patch embeddings of the original TITAN.** Per-slide AUC of the norm for tumor vs. normal tiles, fraction of slides with AUC > 0.5, mean per-slide ratio of mean tumor to mean normal norm and the fraction of slides with a ratio above one, and one-sided Wilcoxon signed-rank test of AUC > 0.5. PAIP pools the prostate and colon slides. Slides with only one class are excluded.

Benchmark	Slides	Mean AUC	Median AUC	AUC > 0.5	Norm ratio	Ratio > 1	p-value
CAMELYON16	48	0.640	0.654	75.0%	1.050	77.1%	4.1×10^{-8}
CAMELYON17	49	0.655	0.668	83.7%	1.053	83.7%	1.4×10^{-9}
PAIP	1,034	0.566	0.582	66.7%	1.024	67.0%	2.5×10^{-42}

C.2 Robustness to the Choice of Prompts

The main results use four-pair prompt combination per dataset. To examine whether the conclusions depend on this choice, we evaluate HYPERION, TITAN, CONCH, and PRISM with the 300 combinations (Fig. A2). HYPERION remains the best method for every combination size on CAMELYON16/17 and PAIP. From Tab. A12, HYPERION is the best method on all three benchmarks, with a Dice of 0.426 on CAMELYON16/17 and 0.781 on PAIP and an mDice of 0.426 on PANDA..

Table A12: **Held-out segmentation Dice** (mean $\pm$ standard deviation over 200 random selected four to six prompt pairs). CM16/17 is the mean of CM16 and CM17, and PAIP the mean of the two organs.

Method	CM16/17	PAIP	PANDA (mDice)
CONCH	0.174 ± 0.037	0.420 ± 0.012	0.286 ± 0.002
PathGen-CLIP	0.255 ± 0.053	0.475 ± 0.008	0.358 ± 0.002
PRISM	0.202 ± 0.043	0.427 ± 0.008	0.109 ± 0.002
TITAN	0.129 ± 0.021	0.401 ± 0.008	0.276 ± 0.002
HYPERION	$\mathbf{0.426 \pm 0.032}$	$\mathbf{0.781 \pm 0.006}$	$\mathbf{0.426 \pm 0.004}$

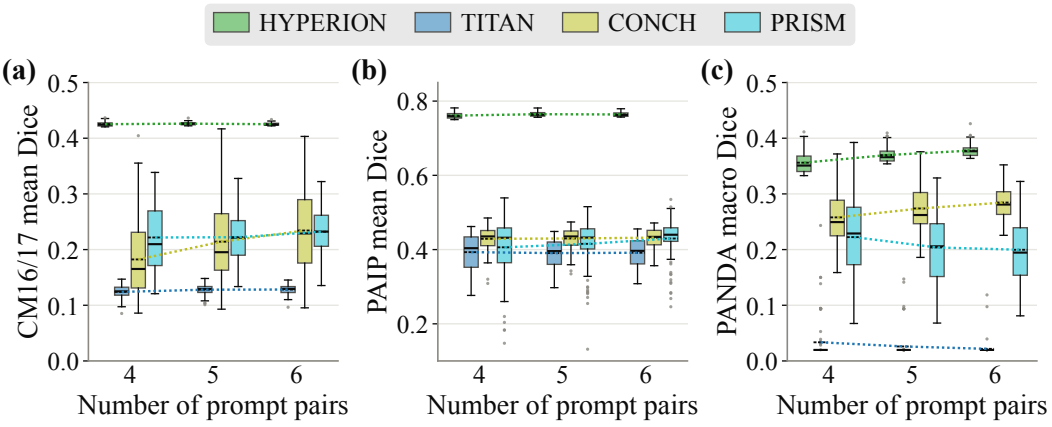

Figure A2: **Segmentation over the 300 prompt combinations.** Dice of each method for combinations with four, five, and six prompt pairs (100 each); the dotted lines connect the means. CAMELYON16/17 and PAIP show the mean of their two datasets, and PANDA shows mDice.

C.3 Additional Ablation Studies

From Tab. A13, we can observe that the patch-level weight λ_{pat} is the most sensitive. A smaller value weakens the patch supervision (0.298 Dice at 0.05), whereas a larger value reduces the Dice to 0.117, which we attribute to the patch objective dominating the slide-level alignment. At 0.05, the CAMELYON16 Dice (0.021) is even lower than without $\mathcal{L}_{\text{pat}}$ (Tab. 4), so training is unstable at small weights.

HYPERION is insensitive to λ_{em} up to the default and to η over the tested range, with all values within 0.016 overall Dice. Larger values of λ_{em} reduce the Dice, mainly on PAIP Prostate.

A smaller smoothing window reduces the Dice on all four datasets, consistent with isolated high-similarity patches entering the pseudo-labels. A larger window reduces the CAMELYON16/17 Dice, consistent with small lesions being blurred. The confidence threshold θ trades the purity of $\mathcal{A}$ and $\mathcal{N}$ against their size; a slightly lower threshold of 0.85 is comparable to the default, while 0.80 and 0.95 are worse. For the norm scale γ, a value of 0.60 is comparable to the default overall and higher on CAMELYON16, whereas 0.50 and 0.80 reduce the Dice.

Table A13: **Hyperparameter sweep.** Each block varies one hyperparameter while the others keep their default values (λ_{em}=0.1, λ_{ent}=0.1, λ_{pat}=0.2, smoothing k=4, θ=0.9, γ=0.65, η=0.7). The default row is the mean of five runs and each sweep setting is a single run. Overall averages the four datasets. Best results are in bold and second-best results are underlined within each hyperparameter block.

Hyperparameter	Setting	CM16		CM17		PAIP Prostate		PAIP Colon		Overall	
		Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Default	**HYPERION (Ours)**	0.463	0.362	0.404	0.318	0.773	0.658	0.783	0.691	0.606	0.507
λ_{em}	0.03	**0.472**	**0.364**	**0.391**	**0.306**	0.724	0.602	**0.764**	**0.669**	**0.588**	**0.485**
	0.05	0.446	0.345	0.386	0.300	**0.753**	**0.632**	0.755	0.656	0.585	0.483
	0.3	0.421	0.335	0.374	0.299	0.600	0.467	0.717	0.617	0.528	0.429
	0.5	0.399	0.320	0.344	0.280	0.690	0.557	0.729	0.624	0.540	0.445
λ_{ent}	0.03	0.419	0.344	0.346	0.287	0.692	0.579	**0.771**	**0.678**	0.557	0.472
	0.05	**0.443**	**0.366**	**0.377**	**0.310**	**0.768**	**0.655**	0.758	0.664	**0.586**	**0.499**
	0.3	0.432	0.347	0.363	0.287	0.737	0.619	0.728	0.623	0.565	0.469
	0.5	0.382	0.288	0.314	0.238	0.631	0.504	0.761	0.661	0.522	0.423
λ_{pat}	0.05	0.021	0.011	0.050	0.036	**0.609**	**0.470**	0.511	0.376	0.298	0.223
	0.1	0.365	0.286	0.316	0.251	0.593	0.463	**0.669**	**0.548**	**0.486**	**0.387**
	0.3	**0.385**	**0.291**	**0.352**	**0.275**	0.533	0.398	0.500	0.372	0.443	0.334
	0.5	0.161	0.094	0.134	0.082	0.029	0.015	0.144	0.082	0.117	0.068
smoothing k	1	0.397	0.305	0.349	0.272	0.651	0.516	0.708	0.596	0.526	0.422
	2	**0.403**	**0.319**	**0.358**	**0.291**	0.654	0.522	0.726	0.619	**0.535**	**0.438**
	8	0.344	0.254	0.281	0.203	**0.661**	**0.543**	**0.773**	**0.680**	0.515	0.420
threshold θ	0.80	0.435	0.354	0.361	0.295	0.643	0.519	0.709	0.606	0.537	0.444
	0.85	**0.456**	**0.367**	**0.374**	**0.298**	**0.781**	**0.669**	0.690	0.584	**0.575**	**0.480**
	0.95	0.442	0.335	0.370	0.282	0.625	0.497	**0.756**	**0.657**	0.548	0.443
γ	0.50	0.337	0.274	0.299	0.258	**0.769**	**0.657**	0.723	0.631	0.532	0.455
	0.60	**0.500**	**0.403**	**0.403**	**0.321**	0.741	0.625	0.764	0.669	**0.602**	**0.504**
	0.70	0.403	0.308	0.327	0.249	0.747	0.631	**0.774**	**0.680**	0.563	0.467
	0.80	0.309	0.222	0.298	0.229	0.463	0.337	0.627	0.495	0.424	0.321
η	0.50	0.437	0.348	**0.399**	**0.317**	**0.762**	**0.644**	**0.761**	**0.665**	0.590	0.494
	0.60	0.432	0.346	0.384	0.312	0.733	0.606	0.748	0.644	0.574	0.477
	0.80	**0.456**	0.351	0.377	0.294	0.715	0.591	0.755	0.652	0.576	0.472
	0.90	0.453	**0.359**	0.396	0.314	0.760	0.643	**0.761**	**0.665**	**0.593**	**0.495**

D Additional Quantitative Results

The main text reports Dice and IoU for segmentation and group means for classification. In this section, we report the segmentation AUC, per-class and per-benchmark results, and results under other prompt protocols.

D.1 Segmentation

Localization AUC. Tab. A14 reports the per-slide AUC of the tumor score, which does not depend on the decision boundary. HYPERION achieves the highest AUC on CAMELYON16, PAIP Prostate, and PAIP Colon, and PRISM is higher on CAMELYON17. On PANDA, CHIEF and the patch-level models achieve a higher mean AUC than HYPERION. HYPERION separates the three Gleason patterns well (one-vs-rest AUC above 0.76) but not stroma and benign epithelium (below 0.38), so its lower mAUC comes from the two non-tumor classes. Its Dice for these classes is nevertheless the highest, because Dice uses the argmax over the five prototypes, whereas the one-vs-rest AUC ranks tiles by a single class score. Since the Lorentz inner product grows with the radius of a patch, high-radius tumor tiles also receive high stroma and benign scores.

Table A14: **Localization AUC on five benchmarks.** For the four binary benchmarks, the AUC of the tumor score is computed per slide and averaged over slides; for PANDA, mAUC averages the five one-vs-rest AUCs. The prompts are those of Tab. 1. Best results are in bold and second-best results are underlined.

Encoding level	Method	CM16	CM17	PAIP Prostate	PAIP Colon	PANDA (mAUC)
Patch-level	CONCH (Lu et al., 2024)	0.826	0.842	0.665	0.775	0.646
	PathGen-CLIP (Sun et al., 2025)	0.845	0.841	0.564	<u>0.809</u>	<u>0.649</u>
Slide-level	CHIEF (Wang et al., 2024)	0.819	0.747	<u>0.878</u>	0.797	**0.673**
	PRISM (Shaikovski et al., 2024)	<u>0.950</u>	**0.909**	0.698	0.667	0.438
	TITAN (Ding et al., 2025)	0.711	0.617	0.487	0.738	0.564
	HYPERION (Ours)	**0.951**	<u>0.891</u>	**0.973**	**0.940**	0.610

Lesion size and category. Tab. A15 splits the Dice of Tab. 1 by the official CAMELYON17 categories of isolated tumor cells (ITC), micro-metastases, and macro-metastases, and CAMELYON16 by the fraction of tumor tiles. The gain of HYPERION comes from micro- and macro-metastases. Its Dice on micro-metastases is 0.367, against at most 0.088 for the baselines. On macro-metastases, it is 0.803, against at most 0.655.

Table A15: **Dice by lesion category and size** with the prompts of Tab. 1. CM17 uses the official categories (16 ITC, 17 micro-, and 16 macro-metastasis slides); CM16 splits slides by the fraction of tumor tiles (28 below and 20 at or above 2%). PRISM uses the shared prompts and 4×4 blocks.

	CAMELYON17			CAMELYON16	
Method	ITC	Micro	Macro	$< 2\%$	$\geq 2\%$
CONCH	0.003	0.022	0.316	0.020	0.343
PathGen-CLIP	0.017	0.088	0.655	0.089	0.512
CHIEF	0.001	0.015	0.244	0.015	0.331
PRISM	0.010	0.046	0.419	0.057	0.475
TITAN	0.002	0.027	0.269	0.016	0.320
HYPERION	**0.043**	**0.367**	**0.803**	**0.246**	**0.766**

Lorentz vs. cosine scoring over prompt combinations. Because the Lorentz inner product scales with the radius of the patch embedding, scoring the same HYPERION embeddings with the Lorentz score instead of cosine similarity shifts the decision boundary between the two prototypes. Tab. A16 compares the two scores over the ten prompt combinations of each dataset. The Lorentz score gives the higher Dice on CAMELYON16 and CAMELYON17 for 10 and 9 of the 10 combinations, and the two scores are comparable on PAIP and PANDA. The AUC differs by at most 0.002 in all cases.

PANDA per-class results. Tab. A17 reports the class-wise Dice on PANDA. HYPERION achieves the highest Dice on stroma, benign epithelium, and Gleason 3 and 5, while CONCH and TITAN are

Table A16: **Lorentz vs. cosine scoring of HYPERION.** (Dice ↑; mDice for PANDA). "Wins" counts the combinations for which the Lorentz score has the higher Dice.

Score	CM16	CM17	PAIP Prostate	PAIP Colon	PANDA
Cosine	0.433	0.376	0.768	0.777	**0.399**
Lorentz	**0.465**	**0.402**	**0.771**	**0.781**	0.395
Wins (of 10)	10	9	5	6	4

higher on Gleason 4. Separating Gleason 3 from Gleason 4 remains difficult for all methods, since both consist of tumor glands that differ mainly in their architecture.

Table A17: **Class-wise Dice on PANDA (Radboud)** with the prompts of Tab. A7. Best results are in bold.

Method	Stroma	Benign	Gleason 3	Gleason 4	Gleason 5	mDice
CONCH	0.609	0.249	0.182	**0.324**	0.102	0.293
PathGen-CLIP	0.407	0.245	0.300	0.209	0.154	0.263
PRISM	0.040	0.007	0.020	0.172	0.098	0.067
TITAN	0.288	0.118	0.175	0.276	0.150	0.201
HYPERION	**0.658**	**0.508**	**0.322**	0.267	**0.302**	**0.411**

PRISM with its own prompts. In Tab. 1, PRISM uses the same prompts as the other methods. Its text model is a report decoder, and its scores depend strongly on the wording of the prompts. We therefore also evaluate PRISM with a separate fixed set of five prompts per class and blocks of 4×4 tiles. Its Dice rises to 0.282 on CAMELYON16 and 0.615 on PAIP Colon, and it remains below HYPERION on all four binary benchmarks. On PANDA, however, it reaches an mDice of 0.439, higher than HYPERION (0.411), so the PANDA comparison with PRISM depends on the prompt protocol.

D.2 Classification

Tab. 3 reports the group means of AUC and F1. Tab. A19 reports the balanced accuracy, where HYPERION also achieves the best group and overall scores. The per-benchmark AUC, F1, and balanced accuracy of linear probing, with standard deviations over folds, are given in Tabs. A20–A25, and the per-benchmark zero-shot results in Tabs. A26 and A27.

Comparison with each baseline. Under linear probing, HYPERION has a higher AUC than TITAN on 25 of the 36 benchmarks, a tie on 1, and a lower AUC on 10 (mean difference $+0.010$, Wilcoxon signed-rank test $p = 0.021$). All nine biomarker benchmarks improve (mean $+0.026$), while the diagnosis benchmarks change by $+0.001$ on average. The largest decreases occur on the smallest cohorts, such as CPTAC COAD MSI-H (-0.108 with 54 patients), where the fold-to-fold standard deviation is large. Against PRISM and CHIEF, HYPERION has a higher AUC on 25 and 26 of the 36 benchmarks ($p = 0.028$ and $p = 0.036$). These tests treat the benchmarks as independent, although EBRAINS, BCNB, and CPTAC-COAD each contribute several tasks on the same slides. After a Holm correction for the three comparisons, all three p-values become 0.063.

For F1 and balanced accuracy, HYPERION improves every group mean over TITAN, but the per-benchmark differences are split evenly, so we do not claim an F1 or balanced-accuracy gain for individual benchmarks.

Zero-shot classification without templates.

For a fair comparison, Tab. A18 also reports TITAN and HYPERION with the class names alone. HYPERION leads TITAN in overall AUC and F1 in both conditions. Without templates, the AUC of TITAN increases slightly while its F1 decreases, and HYPERION decreases slightly in both.

Table A18: **Zero-shot classification with and without prompt templates.** Overall scores (mean of the three group means) over 36 benchmarks.

Model	Templates	AUC	F1	BACC
PRISM	none	0.600	0.400	0.473
TITAN	none	0.644	0.375	0.489
TITAN	23	0.638	0.389	0.496
HYPERION	none	0.651	0.483	0.512
HYPERION	23	**0.661**	**0.491**	**0.520**

Table A19: **Slide-level classification by balanced accuracy (BACC, ↑) under linear probing.** Group scores average the tasks of each group, and the overall score averages the three groups. Best results are in bold and second-best results are underlined.

Method	Diagnosis (9 tasks)	Biomarker (9 tasks)	Mutation (18 tasks)	Overall (36 tasks)
Prov-GigaPath (Xu et al., 2024)	0.5839	0.5584	0.5803	0.5742
CHIEF (Wang et al., 2024)	0.7204	0.6014	0.6195	0.6471
PRISM (Shaikovski et al., 2024)	0.7655	0.5937	0.6151	0.6581
TITAN (Ding et al., 2025)	<u>0.7918</u>	<u>0.6173</u>	<u>0.6244</u>	<u>0.6778</u>
HYPERION (Ours)	**0.7974**	**0.6287**	**0.6326**	**0.6862**

Table A20: **Slide-level prediction by AUC (↑) — diagnosis and biomarker groups.** Each cell is mean ± standard deviation over multiple folds. CM16 and CM17 denote CAMELYON16 and CAMELYON17. Best and second-best means are bold and underlined respectively, including ties at the displayed precision. HYPERION is highlighted in yellow.

Diagnosis and morphology (9 tasks) – 1/2

Method	EBRAINS		DHMC	DHMC	CM16
	30-cls	12-cls	LUAD	RCC	Tumor
Prov-GigaPath	0.9451 ±0.0065	0.9463 ±0.0168	0.5774 ±0.0764	0.9628 ±0.0234	0.7009 ±0.0775
CHIEF	0.9744 ±0.0057	0.9854 ±0.0062	**0.8669** ±0.0401	0.9782 ±0.0115	0.8525 ±0.0726
PRISM	0.9760 ±0.0038	0.9874 ±0.0053	0.8588 ±0.0267	0.9455 ±0.0190	**0.9871** ±0.0146
TITAN	0.9862 ±0.0018	0.9954 ±0.0026	0.8643 ±0.0401	0.9864 ±0.0078	0.9689 ±0.0294
HYPERION (Ours)	**0.9874** ±0.0021	**0.9958** ±0.0022	0.8493 ±0.0604	**0.9878** ±0.0082	0.9713 ±0.0261

Diagnosis and morphology (9 tasks) – 2/2

Method	CM17	PANDA	IMP-CRC	CPTAC
	pN	ISUP	Dysplasia	NSCLC
Prov-GigaPath	0.6934 ±0.0576	0.8206 ±0.0061	0.9703 ±0.0148	0.9840 ±0.0168
CHIEF	0.8283 ±0.0477	0.9374 ±0.0023	0.9842 ±0.0088	0.9973 ±0.0058
PRISM	**0.9115** ±0.0375	**0.9434** ±0.0033	0.9914 ±0.0054	**1.0000** ±0.0000
TITAN	0.8868 ±0.0586	0.9257 ±0.0038	**0.9928** ±0.0045	**1.0000** ±0.0000
HYPERION (Ours)	0.9086 ±0.0458	0.9249 ±0.0038	0.9926 ±0.0038	**1.0000** ±0.0000

Biomarker and phenotype status (9 tasks) – 1/2

Method	BCNB			EBRAINS
	ER	HER2	PR	IDH
Prov-GigaPath	0.8218 ±0.0444	0.6875 ±0.0541	0.7831 ±0.0536	0.9307 ±0.0176
CHIEF	0.8883 ±0.0426	0.7332 ±0.0485	0.8314 ±0.0381	0.9622 ±0.0245
PRISM	0.9109 ±0.0188	0.7394 ±0.0563	0.8395 ±0.0449	0.9589 ±0.0175
TITAN	0.9153 ±0.0320	0.7416 ±0.0466	0.8470 ±0.0266	0.9634 ±0.0215
HYPERION (Ours)	**0.9157** ±0.0209	**0.7589** ±0.0473	**0.8539** ±0.0313	**0.9726** ±0.0149

Biomarker and phenotype status (9 tasks) – 2/2

Method	CPTAC BRCA	CPTAC LUAD	CPTAC LSCC		CPTAC COAD
	Immune	Immune	Immune	Grade	Immune
Prov-GigaPath	0.6506 ±0.1438	0.6248 ±0.0902	0.5935 ±0.1340	**0.6910** ±0.2361	0.5153 ±0.1934
CHIEF	0.7260 ±0.0909	**0.7341** ±0.1279	0.6651 ±0.1270	0.6282 ±0.1651	0.5167 ±0.1996
PRISM	0.6040 ±0.1075	0.6377 ±0.1269	0.6169 ±0.1471	0.6030 ±0.2093	0.4861 ±0.1220
TITAN	0.7050 ±0.1093	0.6458 ±0.1292	0.6702 ±0.1019	0.6625 ±0.2177	0.5333 ±0.1308
HYPERION (Ours)	**0.7671** ±0.1323	0.7181 ±0.1060	**0.7111** ±0.0958	0.6818 ±0.2293	**0.5403** ±0.1302

Table A21: **Slide-level prediction by AUC (↑) — mutation group.** Each cell is mean ± standard deviation over multiple folds. CM16 and CM17 denote CAMELYON16 and CAMELYON17. Best and second-best means are bold and underlined respectively, including ties at the displayed precision. HYPERION is highlighted in yellow.

Mutation and MSI prediction (18 tasks) – 1/4

Method	CPTAC GBM		CPTAC LUAD			
	EGFR	TP53	EGFR	KRAS	STK11	TP53
Prov-GigaPath	0.7405 ±0.1798	0.7579 ±0.1297	0.7244 ±0.1685	0.7214 ±0.2018	0.7444 ±0.3035	0.6201 ±0.1270
CHIEF	0.6458 ±0.1957	0.8238 ±0.1421	0.7238 ±0.1534	0.6988 ±0.0983	**0.8556** ±0.1319	0.7007 ±0.1816
PRISM	0.6964 ±0.2562	**0.8841** ±0.1081	0.7464 ±0.1257	**0.7696** ±0.1220	0.7833 ±0.2085	0.7088 ±0.2004
TITAN	**0.7619** ±0.1721	0.8065 ±0.1058	0.8030 ±0.1473	0.6857 ±0.2285	0.8389 ±0.1371	0.7561 ±0.1575
HYPERION (Ours)	0.7262 ±0.1469	0.8500 ±0.1132	**0.8417** ±0.1179	0.7000 ±0.1845	0.8500 ±0.1292	**0.8025** ±0.1068

Mutation and MSI prediction (18 tasks) – 2/4

Method	CPTAC LSCC		CPTAC BRCA	
	ARID1A	KEAP1	PIK3CA	TP53
Prov-GigaPath	0.5578 ±0.3124	0.5089 ±0.2567	0.4976 ±0.1891	0.7050 ±0.1265
CHIEF	0.4989 ±0.2667	**0.7322** ±0.2205	0.5833 ±0.1210	0.7767 ±0.1154
PRISM	0.4522 ±0.2690	0.5822 ±0.1947	0.5917 ±0.1769	0.7950 ±0.1192
TITAN	0.5033 ±0.3196	0.6322 ±0.2635	**0.6155** ±0.1473	0.7967 ±0.0860
HYPERION (Ours)	**0.5656** ±0.2612	0.6378 ±0.2746	0.6083 ±0.1582	**0.8283** ±0.0734

Mutation and MSI prediction (18 tasks) – 3/4

Method	CPTAC COAD			
	APC	KRAS	TP53	PIK3CA
Prov-GigaPath	0.5925 ±0.3111	0.4778 ±0.2360	0.6458 ±0.2870	0.6125 ±0.3328
CHIEF	**0.6025** ±0.2976	**0.6167** ±0.1203	0.6083 ±0.3472	0.5167 ±0.3152
PRISM	0.5100 ±0.2383	0.5500 ±0.2587	0.5354 ±0.2589	0.6458 ±0.2772
TITAN	0.5775 ±0.3249	0.5000 ±0.2194	0.6833 ±0.3261	**0.6500** ±0.2947
HYPERION (Ours)	0.5575 ±0.3650	0.4944 ±0.2502	**0.7250** ±0.3250	0.5458 ±0.3552

Mutation and MSI prediction (18 tasks) – 4/4

Method	CPTAC COAD			
	ARID1A	ACVR2A	SETD1B	MSI-H
Prov-GigaPath	0.7704 ±0.2516	0.6800 ±0.2810	0.6150 ±0.3749	0.7225 ±0.3111
CHIEF	**0.8074** ±0.2619	0.8000 ±0.2037	0.8250 ±0.2250	**0.8925** ±0.1379
PRISM	0.7222 ±0.2740	**0.8600** ±0.1609	0.8400 ±0.2385	0.8800 ±0.1208
TITAN	0.6815 ±0.3729	0.6750 ±0.4000	**0.9600** ±0.1200	0.8550 ±0.1635
HYPERION (Ours)	0.7704 ±0.3283	0.7075 ±0.3705	0.9350 ±0.1343	0.7475 ±0.2807

Table A22: **Slide-level prediction by F1-score (↑) — diagnosis and biomarker groups.** Each cell is mean ± standard deviation over the multiple folds. CM16 and CM17 denote CAMELYON16 and CAMELYON17. Best and second-best means are bold and underlined respectively, including ties at the displayed precision. HYPERION is highlighted in yellow.

Diagnosis and morphology (9 tasks)

Method	EBRAINS		DHMC	DHMC	CM16
	30-cls	12-cls	LUAD	RCC	Tumor
Prov-GigaPath	0.5765 ±0.0285	0.7121 ±0.0235	0.4804 ±0.0810	0.8602 ±0.0507	0.6754 ±0.0611
CHIEF	0.7025 ±0.0347	0.8503 ±0.0220	**0.6938** ±0.0452	0.8742 ±0.0385	0.8083 ±0.0717
PRISM	0.7200 ±0.0294	0.8874 ±0.0224	0.5773 ±0.0933	0.8186 ±0.0528	**0.9673** ±0.0161
TITAN	0.7782 ±0.0221	**0.9331** ±0.0229	0.6438 ±0.0632	0.9276 ±0.0273	0.9496 ±0.0275
HYPERION (Ours)	**0.7841** ±0.0297	0.9327 ±0.0216	0.6896 ±0.1003	**0.9343** ±0.0225	0.9671 ±0.0227

Diagnosis and morphology (9 tasks) (continued)

Method	CM17	PANDA	IMP-CRC	CPTAC
	pN	ISUP	Dysplasia	NSCLC
Prov-GigaPath	0.6061 ±0.0568	0.4909 ±0.0201	0.8813 ±0.0270	0.9061 ±0.0534
CHIEF	0.7424 ±0.0396	0.7141 ±0.0088	0.9206 ±0.0225	0.9464 ±0.0334
PRISM	**0.8552** ±0.0535	**0.7422** ±0.0078	**0.9477** ±0.0291	0.9658 ±0.0441
TITAN	0.8258 ±0.0511	0.6904 ±0.0125	0.9410 ±0.0234	**0.9856** ±0.0308
HYPERION (Ours)	0.8231 ±0.0383	0.6866 ±0.0185	0.9431 ±0.0227	0.9760 ±0.0323

Biomarker and phenotype status (9 tasks)

Method	BCNB			EBRAINS
	ER	HER2	PR	IDH
Prov-GigaPath	0.7723 ±0.0255	0.6718 ±0.0215	0.7580 ±0.0361	0.8666 ±0.0324
CHIEF	0.8368 ±0.0390	0.7287 ±0.0356	0.8062 ±0.0363	0.9219 ±0.0379
PRISM	0.8685 ±0.0252	**0.7442** ±0.0370	0.8193 ±0.0299	0.9008 ±0.0291
TITAN	0.8621 ±0.0303	0.7364 ±0.0295	0.8205 ±0.0226	0.9208 ±0.0294
HYPERION (Ours)	**0.8719** ±0.0382	0.7322 ±0.0346	**0.8383** ±0.0315	**0.9301** ±0.0177

Biomarker and phenotype status (9 tasks) (continued)

Method	CPTAC BRCA	CPTAC LUAD	CPTAC LSCC		CPTAC COAD
	Immune	Immune	Immune	Grade	Immune
Prov-GigaPath	0.4920 ±0.1579	0.4489 ±0.1274	0.4226 ±0.1308	0.5969 ±0.1302	0.3417 ±0.2017
CHIEF	0.4987 ±0.1134	**0.5043** ±0.1362	0.4423 ±0.1567	0.5634 ±0.0992	**0.3969** ±0.2243
PRISM	0.4001 ±0.1262	0.4165 ±0.1725	0.4413 ±0.1854	0.6023 ±0.1378	0.3209 ±0.1311
TITAN	0.5112 ±0.1547	0.4376 ±0.1690	**0.5245** ±0.1292	**0.6192** ±0.1458	0.3158 ±0.1789
HYPERION (Ours)	**0.5517** ±0.1607	0.4988 ±0.1380	0.5093 ±0.1584	0.6169 ±0.1580	0.3082 ±0.2192

Table A23: **Slide-level prediction by F1-score (↑) — mutation group.** Each cell is mean ± standard deviation over the multiple folds. CM16 and CM17 denote CAMELYON16 and CAMELYON17. Best and second-best means are bold and underlined respectively, including ties at the displayed precision. HYPERION is highlighted in yellow.

Mutation and MSI prediction (18 tasks)

Method	CPTAC GBM		CPTAC LUAD			
	EGFR	TP53	EGFR	KRAS	STK11	TP53
Prov-GigaPath	**0.7569** ±0.1241	0.6852 ±0.1138	0.6479 ±0.1318	0.6373 ±0.1086	0.7871 ±0.0923	0.5818 ±0.1282
CHIEF	0.6903 ±0.1356	0.7915 ±0.0701	0.6790 ±0.1590	0.6687 ±0.0951	0.8187 ±0.1378	0.6211 ±0.1656
PRISM	0.7462 ±0.1498	**0.8107** ±0.0986	0.6772 ±0.1130	**0.6915** ±0.0697	0.8193 ±0.1173	0.6582 ±0.1808
TITAN	0.7289 ±0.1492	0.7910 ±0.1121	0.6948 ±0.1300	0.6721 ±0.1346	0.8110 ±0.1152	0.6747 ±0.1360
HYPERION (Ours)	0.7130 ±0.1328	0.7693 ±0.1431	**0.7896** ±0.1176	0.6818 ±0.1404	**0.8678** ±0.0776	**0.7189** ±0.1208

Mutation and MSI prediction (18 tasks) (continued 2/4)

Method	CPTAC LSCC		CPTAC BRCA	
	ARID1A	KEAP1	PIK3CA	TP53
Prov-GigaPath	**0.7984** ±0.0644	0.8048 ±0.0634	0.5396 ±0.1593	0.6480 ±0.1139
CHIEF	0.7900 ±0.0678	**0.8199** ±0.0994	0.5897 ±0.1374	0.6964 ±0.1188
PRISM	0.7591 ±0.0753	0.8032 ±0.0756	0.5619 ±0.1540	**0.7199** ±0.1185
TITAN	0.7505 ±0.0868	0.8192 ±0.0700	**0.6079** ±0.1425	0.6864 ±0.0750
HYPERION (Ours)	0.7847 ±0.0976	0.8057 ±0.1028	0.6042 ±0.0808	0.7071 ±0.0880

Mutation and MSI prediction (18 tasks) (continued 3/4)

Method	CPTAC COAD			
	APC	KRAS	TP53	PIK3CA
Prov-GigaPath	0.6744 ±0.1769	0.4955 ±0.2346	0.6031 ±0.1804	**0.6593** ±0.1887
CHIEF	**0.6863** ±0.1089	**0.5355** ±0.1330	0.6436 ±0.2491	0.6289 ±0.1346
PRISM	0.6015 ±0.1185	0.5240 ±0.2823	0.5471 ±0.1536	0.5854 ±0.2069
TITAN	0.6674 ±0.1356	0.4914 ±0.1630	**0.6930** ±0.2611	0.5796 ±0.2268
HYPERION (Ours)	0.6683 ±0.1438	0.4854 ±0.1698	0.6593 ±0.2456	0.6360 ±0.2130

Mutation and MSI prediction (18 tasks) (continued 4/4)

Method	CPTAC COAD			
	ARID1A	ACVR2A	SETD1B	MSI-H
Prov-GigaPath	0.8070 ±0.1128	0.7081 ±0.1581	0.7593 ±0.1524	0.7475 ±0.1152
CHIEF	0.8066 ±0.1042	0.7765 ±0.1717	0.8491 ±0.1554	**0.8084** ±0.1307
PRISM	0.7725 ±0.1390	**0.7863** ±0.1564	0.7869 ±0.1481	0.8001 ±0.1479
TITAN	**0.8111** ±0.1465	0.7637 ±0.1877	**0.8619** ±0.1427	0.7940 ±0.1157
HYPERION (Ours)	0.7792 ±0.1634	0.7543 ±0.1880	0.8222 ±0.1457	0.7915 ±0.1430

Table A24: **Slide-level prediction by BACC (↑) — diagnosis and biomarker groups.** Each cell is mean ± standard deviation over the multiple folds. CM16 and CM17 denote CAMELYON16 and CAMELYON17. Best and second-best means are bold and underlined respectively, including ties at the displayed precision. HYPERION is highlighted in yellow.

Diagnosis and morphology (9 tasks)

Method	EBRAINS		DHMC	DHMC	CM16
	30-cls	12-cls	LUAD	RCC	Tumor
Prov-GigaPath	0.4997 ±0.0296	0.5176 ±0.0373	0.2865 ±0.0731	0.7099 ±0.0760	0.6581 ±0.0629
CHIEF	0.6504 ±0.0406	0.7448 ±0.0338	0.5204 ±0.0579	0.7500 ±0.0641	0.7831 ±0.0721
PRISM	0.6794 ±0.0368	0.8314 ±0.0460	0.4502 ±0.0771	0.6710 ±0.1076	**0.9633** ±0.0193
TITAN	0.7464 ±0.0221	**0.8982** ±0.0374	0.4912 ±0.1153	0.8424 ±0.0868	0.9444 ±0.0300
HYPERION (Ours)	**0.7550** ±0.0308	0.8878 ±0.0433	**0.5391** ±0.1634	**0.8523** ±0.0632	0.9602 ±0.0262

Diagnosis and morphology (9 tasks) (continued)

Method	CM17	PANDA	IMP-CRC	CPTAC
	pN	ISUP	Dysplasia	NSCLC
Prov-GigaPath	0.3572 ±0.0604	0.4464 ±0.0195	0.8727 ±0.0296	0.9073 ±0.0534
CHIEF	0.5150 ±0.0467	0.6537 ±0.0109	0.9191 ±0.0241	0.9468 ±0.0334
PRISM	**0.6981** ±0.0967	**0.6813** ±0.0108	**0.9489** ±0.0255	0.9655 ±0.0449
TITAN	0.6415 ±0.0730	0.6336 ±0.0140	0.9427 ±0.0267	**0.9855** ±0.0315
HYPERION (Ours)	0.6308 ±0.0694	0.6299 ±0.0211	0.9460 ±0.0230	0.9759 ±0.0329

Biomarker and phenotype status (9 tasks)

Method	BCNB			EBRAINS
	ER	HER2	PR	IDH
Prov-GigaPath	0.6050 ±0.0363	0.5381 ±0.0225	0.6335 ±0.0426	0.8545 ±0.0349
CHIEF	0.7260 ±0.0596	0.6123 ±0.0397	0.7114 ±0.0507	0.9155 ±0.0394
PRISM	**0.7947** ±0.0544	**0.6495** ±0.0463	0.7457 ±0.0431	0.8976 ±0.0330
TITAN	0.7791 ±0.0505	0.6414 ±0.0374	0.7476 ±0.0332	0.9167 ±0.0328
HYPERION (Ours)	0.7919 ±0.0657	0.6309 ±0.0395	**0.7636** ±0.0515	**0.9250** ±0.0181

Biomarker and phenotype status (9 tasks) (continued)

Method	CPTAC BRCA	CPTAC LUAD	CPTAC LSCC		CPTAC COAD
	Immune	Immune	Immune	Grade	Immune
Prov-GigaPath	0.5250 ±0.1593	0.4667 ±0.1171	0.4472 ±0.1257	0.6058 ±0.1335	0.3500 ±0.2167
CHIEF	0.5111 ±0.1231	**0.5083** ±0.1321	0.4583 ±0.1487	0.5700 ±0.0976	**0.4000** ±0.2603
PRISM	0.4194 ±0.1306	0.4306 ±0.1649	0.4472 ±0.1898	0.6250 ±0.1150	0.3333 ±0.1491
TITAN	0.5333 ±0.1565	0.4472 ±0.1636	**0.5306** ±0.1282	**0.6267** ±0.1476	0.3333 ±0.1972
HYPERION (Ours)	**0.5722** ±0.1611	0.5000 ±0.1411	0.5167 ±0.1492	0.6250 ±0.1417	0.3333 ±0.2236

Table A25: **Slide-level prediction by BACC (↑) — mutation group.** Each cell is mean ± standard deviation over the multiple folds. CM16 and CM17 denote CAMELYON16 and CAMELYON17. Best and second-best means are bold and underlined respectively, including ties at the displayed precision. HYPERION is highlighted in yellow.

Mutation and MSI prediction (18 tasks)

Method	CPTAC GBM		CPTAC LUAD			
	EGFR	TP53	EGFR	KRAS	STK11	TP53
Prov-GigaPath	<u>0.6265</u> ±0.1778	0.6452 ±0.1544	0.6417 ±0.1555	0.5714 ±0.1268	0.5722 ±0.1718	0.5874 ±0.1195
CHIEF	0.5699 ±0.1887	<u>0.7720</u> ±0.1160	0.6744 ±0.1499	0.6071 ±0.1105	0.6611 ±0.2437	0.6199 ±0.1551
PRISM	**0.6330** ±0.2075	**0.8042** ±0.1262	0.6619 ±0.1293	**0.6571** ±0.0943	<u>0.7139</u> ±0.2034	0.6635 ±0.1637
TITAN	0.6176 ±0.2193	0.7613 ±0.1383	<u>0.6821</u> ±0.1399	0.6226 ±0.1447	0.6111 ±0.1972	<u>0.6708</u> ±0.1495
HYPERION (Ours)	0.5988 ±0.1586	0.7280 ±0.1621	**0.7786** ±0.1315	<u>0.6310</u> ±0.1732	**0.7278** ±0.1624	**0.7210** ±0.1231

Mutation and MSI prediction (18 tasks) (continued 2/4)

Method	CPTAC LSCC		CPTAC BRCA	
	ARID1A	KEAP1	PIK3CA	TP53
Prov-GigaPath	0.4728 ±0.0369	0.4794 ±0.0337	0.4994 ±0.1613	0.6467 ±0.1179
CHIEF	0.4639 ±0.0410	0.5689 ±0.2073	0.5589 ±0.1529	0.7050 ±0.1019
PRISM	<u>0.4772</u> ±0.1356	0.5267 ±0.1614	0.5262 ±0.1841	**0.7250** ±0.1194
TITAN	0.4283 ±0.0718	<u>0.5744</u> ±0.1768	**0.5690** ±0.1363	0.6875 ±0.0876
HYPERION (Ours)	**0.4983** ±0.1783	**0.5972** ±0.2348	<u>0.5649</u> ±0.0814	<u>0.7125</u> ±0.0918

Mutation and MSI prediction (18 tasks) (continued 3/4)

Method	CPTAC COAD			
	APC	KRAS	TP53	PIK3CA
Prov-GigaPath	<u>0.5575</u> ±0.2503	0.5083 ±0.2370	0.5667 ±0.1947	**0.5875** ±0.2375
CHIEF	**0.5700** ±0.2002	**0.5500** ±0.1404	0.5958 ±0.2465	0.5500 ±0.1953
PRISM	0.4325 ±0.1204	**0.5500** ±0.2587	0.5250 ±0.1470	0.5333 ±0.2471
TITAN	<u>0.5575</u> ±0.2139	<u>0.5167</u> ±0.1280	**0.6667** ±0.2758	0.4750 ±0.2358
HYPERION (Ours)	0.5300 ±0.2190	0.4917 ±0.1601	<u>0.6375</u> ±0.2197	<u>0.5583</u> ±0.2570

Mutation and MSI prediction (18 tasks) (continued 4/4)

Method	CPTAC COAD			
	ARID1A	ACVR2A	SETD1B	MSI-H
Prov-GigaPath	**0.6352** ±0.2386	0.5975 ±0.2303	0.6350 ±0.2663	0.6150 ±0.2157
CHIEF	0.6019 ±0.2144	0.6800 ±0.2441	<u>0.7375</u> ±0.2649	0.6650 ±0.2346
PRISM	0.5667 ±0.2404	**0.7575** ±0.2139	0.6300 ±0.2492	<u>0.6875</u> ±0.2322
TITAN	**0.6352** ±0.2651	<u>0.6950</u> ±0.2704	**0.7775** ±0.2511	**0.6900** ±0.2134
HYPERION (Ours)	<u>0.6130</u> ±0.2671	0.6575 ±0.2679	0.6900 ±0.2548	0.6500 ±0.2358

Table A26: **Zero-shot slide-level prediction measured by AUC (↑).** Mean values over the same fixed folds as the linear probing tables; no task-specific training is involved. CM16 and CM17 denote CAMELYON16 and CAMELYON17. HYPERION is highlighted in yellow. Best results are bold and second-best results are underlined, including ties at the displayed precision. Prov-GigaPath and CHIEF are omitted: the former is vision-only and the latter exposes only a supervised binary head, so neither can score free-form class prompts.

Diagnosis and morphology (9 tasks)

Method	EBRAINS		DHMC	DHMC	CM16	CM17	PANDA	IMP-CRC	CPTAC
	30-cls	12-cls	LUAD	RCC	Tumor	pN	ISUP	Dysplasia	NSCLC
PRISM	**0.4520**	**0.6189**	0.8171	0.5475	**0.9825**	**0.9211**	**0.8877**	**0.9597**	**0.9973**
TITAN	0.4008	0.5451	**0.8433**	**0.9684**	0.9317	0.6241	0.7509	0.9464	0.9945
HYPERION (Ours)	0.4408	0.5620	0.8155	0.9432	0.9374	0.7625	0.8311	0.9489	**0.9973**

Biomarker and phenotype status (9 tasks)

Method	BCNB			EBRAINS	CPTAC BRCA	CPTAC LUAD	CPTAC LSCC		CPTAC COAD
	ER	HER2	PR	IDH	Immune	Immune	Immune	Grade	Immune
PRISM	0.3054	0.7089	0.3482	0.2055	**0.5534**	**0.5591**	0.4196	0.5170	0.4903
TITAN	0.6200	**0.7507**	0.4638	**0.9585**	0.5056	0.5024	**0.5859**	**0.5720**	0.5625
HYPERION (Ours)	**0.7276**	0.7063	**0.5849**	0.9562	0.4714	0.5429	0.5758	0.5452	**0.6250**

Mutation and MSI prediction (18 tasks)

Method	CPTAC GBM		CPTAC LUAD				CPTAC LSCC		CPTAC BRCA	
	EGFR	TP53	EGFR	KRAS	STK11	TP53	ARID1A	KEAP1	PIK3CA	TP53
PRISM	0.3107	0.6010	0.5417	**0.4530**	0.2833	**0.6710**	0.6133	**0.7144**	**0.5173**	0.6650
TITAN	**0.4926**	0.6212	**0.6482**	0.4012	0.3833	0.5243	0.5344	0.4278	0.4327	**0.8375**
HYPERION (Ours)	0.4488	**0.6702**	0.5702	0.2821	**0.4222**	0.6300	**0.6433**	0.4444	0.4304	0.6158

Method	CPTAC COAD							
	APC	KRAS	TP53	PIK3CA	ARID1A	ACVR2A	SETD1B	MSI-H
PRISM	0.5275	0.5222	0.4583	0.4500	0.6259	0.7225	**0.6000**	0.5400
TITAN	**0.5325**	0.5056	**0.6375**	0.3417	**0.6444**	0.3850	0.3650	0.6650
HYPERION (Ours)	0.5250	**0.5611**	0.4125	**0.5792**	0.6037	**0.7900**	0.4450	**0.6800**

Table A27: **Zero-shot slide-level prediction measured by F1-score (↑).** Mean values over the same fixed folds as the linear probing tables; no task-specific training is involved. CM16 and CM17 denote CAMELYON16 and CAMELYON17. HYPERION is highlighted in yellow. Best results are bold and second-best results are underlined, including ties at the displayed precision. Prov-GigaPath and CHIEF are omitted: the former is vision-only and the latter exposes only a supervised binary head, so neither can score free-form class prompts.

Diagnosis and morphology (9 tasks)

Method	EBRAINS		DHMC	DHMC	CM16	CM17	PANDA	IMP-CRC	CPTAC
	30-cls	12-cls	LUAD	RCC	Tumor	pN	ISUP	Dysplasia	NSCLC
PRISM	0.0159	0.2556	0.1558	0.2339	**0.9391**	0.6154	**0.5666**	0.7985	0.9515
TITAN	**0.0250**	0.1665	**0.6600**	**0.8404**	0.2554	0.1944	0.1820	0.5220	**0.9660**
HYPERION (Ours)	0.0144	**0.2871**	0.3007	0.7728	0.9123	**0.6694**	0.3855	**0.8582**	0.9564

Biomarker and phenotype status (9 tasks)

Method	BCNB			EBRAINS	CPTAC BRCA	CPTAC LUAD	CPTAC LSCC		CPTAC COAD
	ER	HER2	PR	IDH	Immune	Immune	Immune	Grade	Immune
PRISM	0.6805	0.3718	0.6351	0.2373	0.1666	0.2924	0.1560	0.3179	**0.3289**
TITAN	0.3370	**0.7000**	0.3536	**0.8738**	**0.2706**	0.1955	**0.4348**	**0.3447**	0.1784
HYPERION (Ours)	**0.7551**	0.4003	**0.6565**	0.8469	0.1766	**0.3184**	0.2379	0.2584	0.2186

Mutation and MSI prediction (18 tasks)

Method	CPTAC GBM		CPTAC LUAD				CPTAC LSCC		CPTAC BRCA	
	EGFR	TP53	EGFR	KRAS	STK11	TP53	ARID1A	KEAP1	PIK3CA	TP53
PRISM	0.2606	0.3404	**0.4352**	**0.1921**	0.1386	**0.4871**	0.0280	**0.4911**	0.2218	0.2397
TITAN	0.1300	0.1985	0.3055	0.1734	0.0484	0.4594	**0.5411**	0.0592	0.2795	**0.6968**
HYPERION (Ours)	**0.4304**	**0.6609**	0.3527	0.1470	**0.2099**	0.4864	0.2054	0.2547	**0.5059**	0.5336

Method	CPTAC COAD							
	APC	KRAS	TP53	PIK3CA	ARID1A	ACVR2A	SETD1B	MSI-H
PRISM	**0.7237**	0.3031	0.5089	0.4600	0.2688	0.2505	0.2186	0.5694
TITAN	0.6671	0.5114	**0.6124**	0.4939	0.0463	0.1100	0.0571	0.5878
HYPERION (Ours)	0.5235	**0.5221**	0.2300	**0.6183**	**0.6745**	**0.7362**	**0.6939**	**0.6493**

E Additional Qualitative Results

E.1 Zero-shot Segmentation

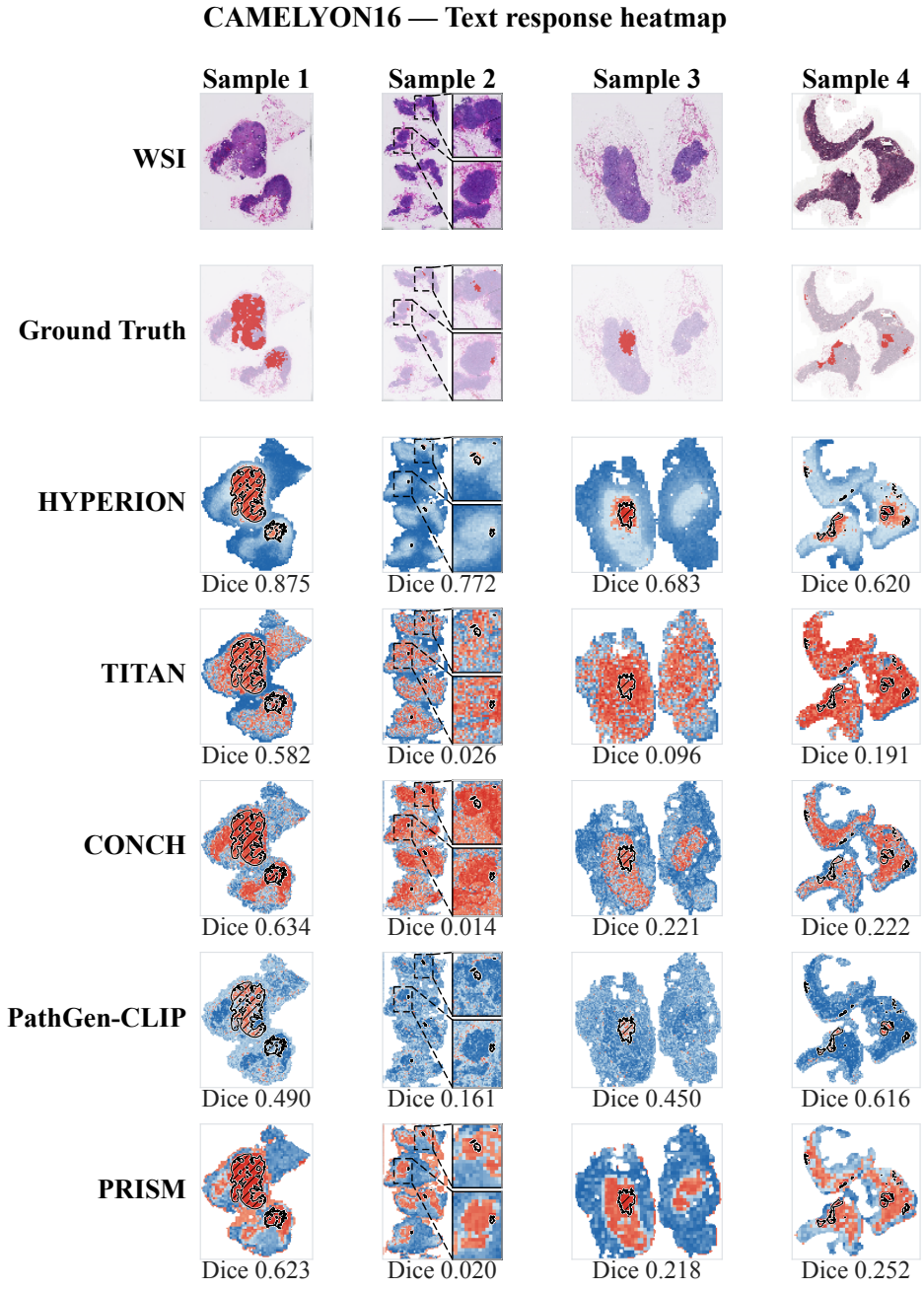

Figure A3: **Zero-shot tumor segmentation on CAMELYON16.** Columns show four representative slides, while rows show the H&E image, pixel-level tumor annotation, and signed-percentile text-response heatmaps from each method. Red and blue indicate positive tumor-side and negative normal-side responses, respectively, and ground-truth tumor regions are hatched with a black outline. Slide-level Dice scores are shown below each model panel.

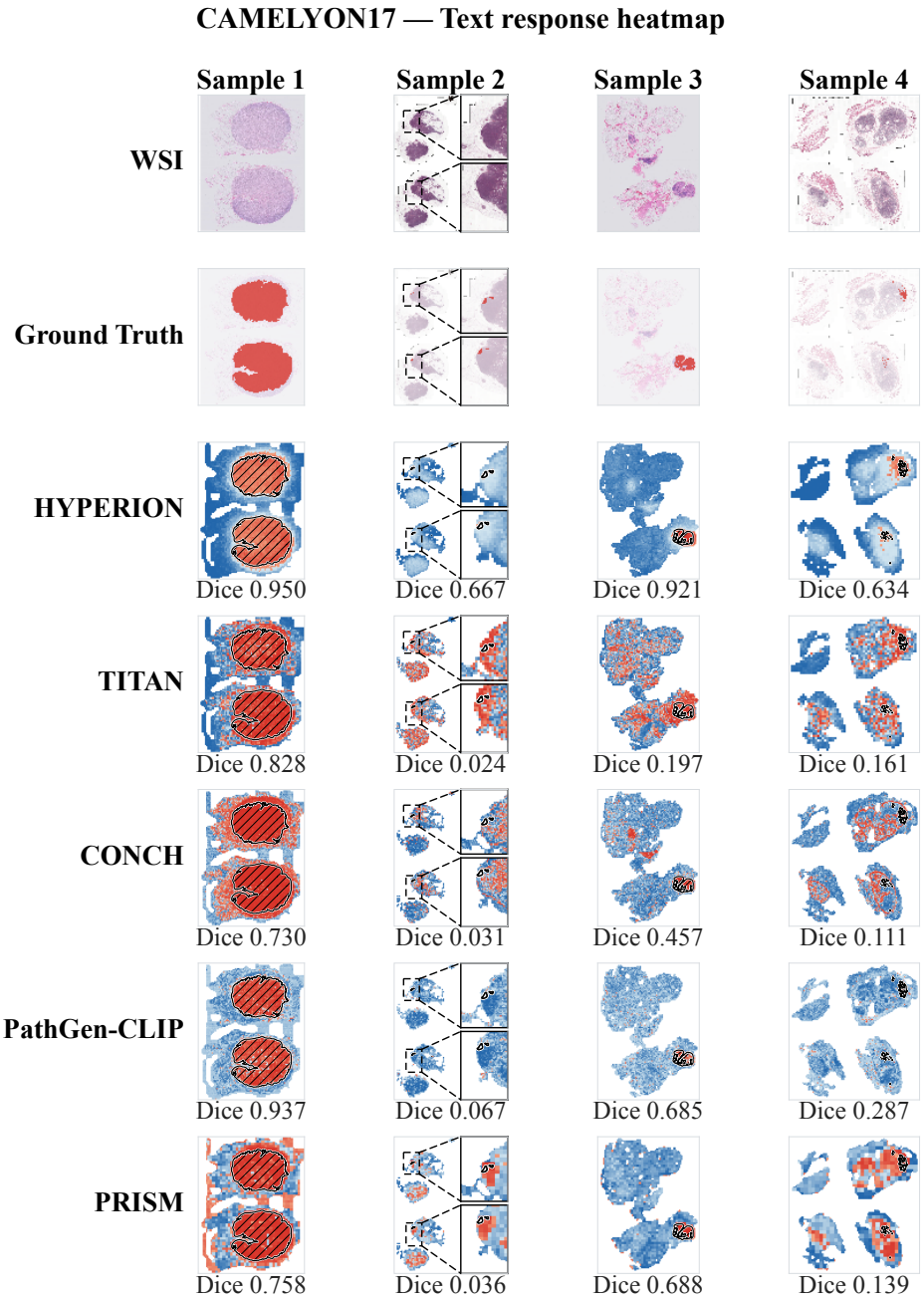

Figure A4: **Zero-shot tumor segmentation on CAMELYON17.** Layout as in Fig. A3.

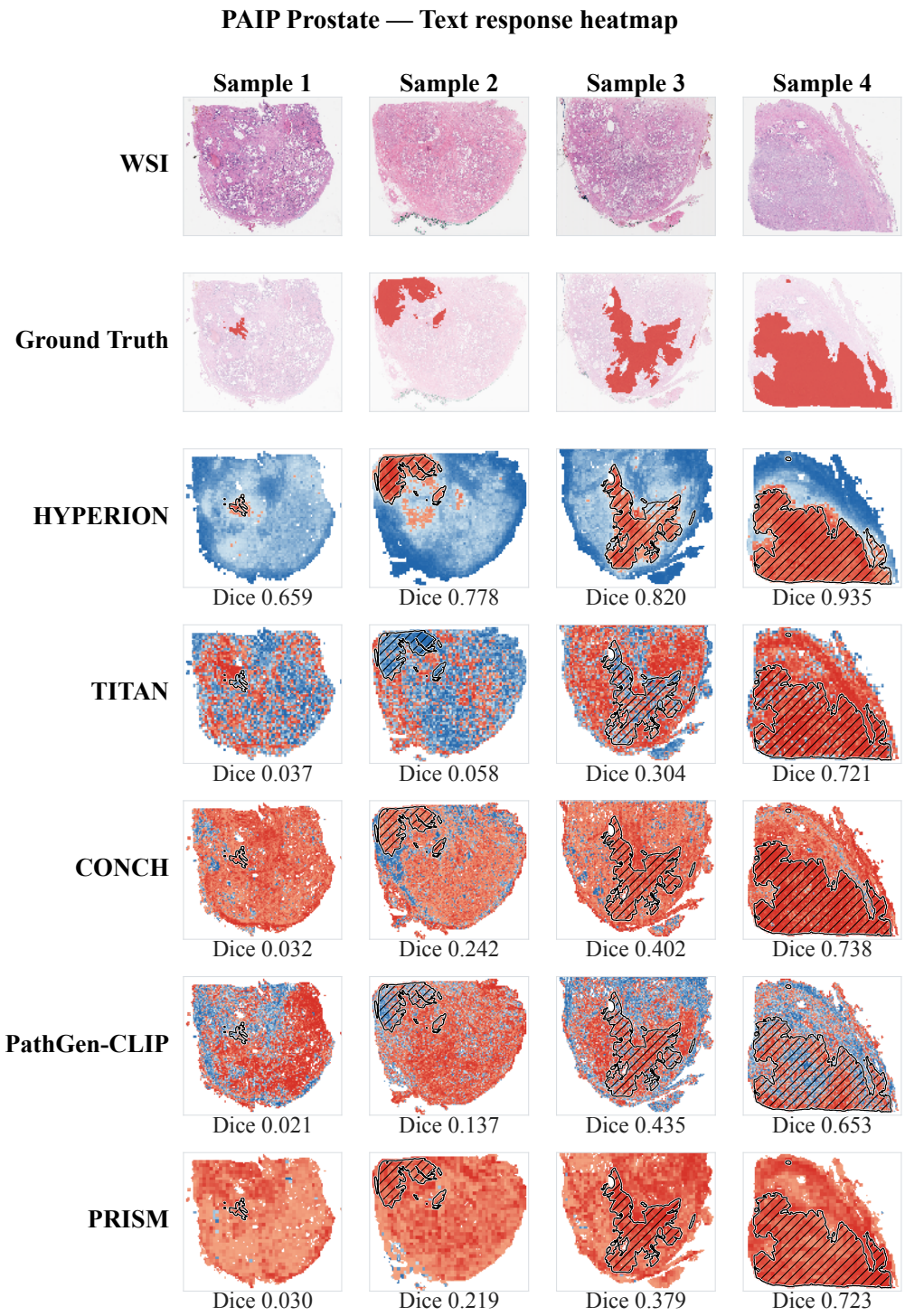

Figure A5: **Zero-shot tumor segmentation on PAIP Prostate.** Layout as in Fig. A3.

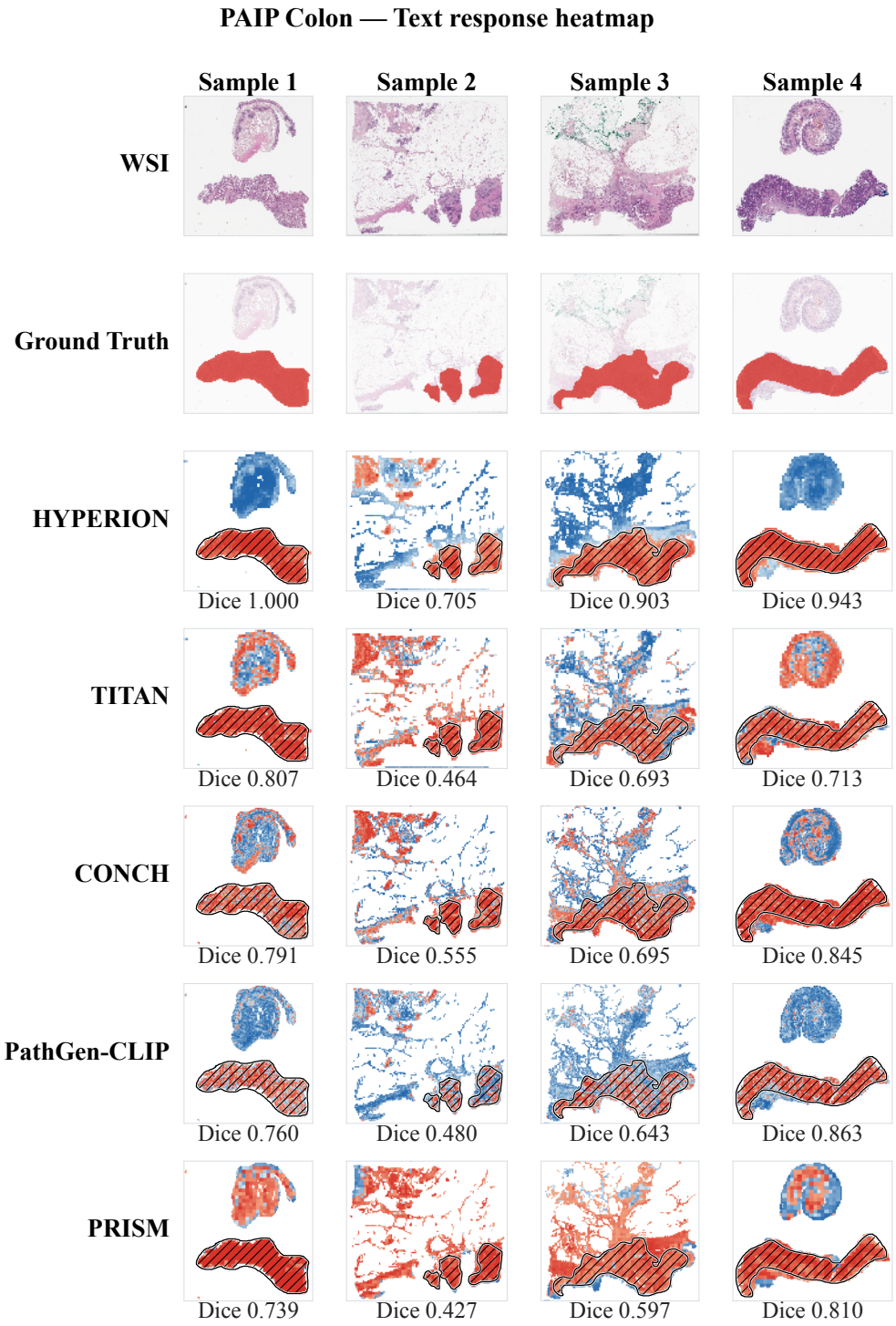

Figure A6: **Zero-shot tumor segmentation on PAIP Colon.** Layout as in Fig. A3.

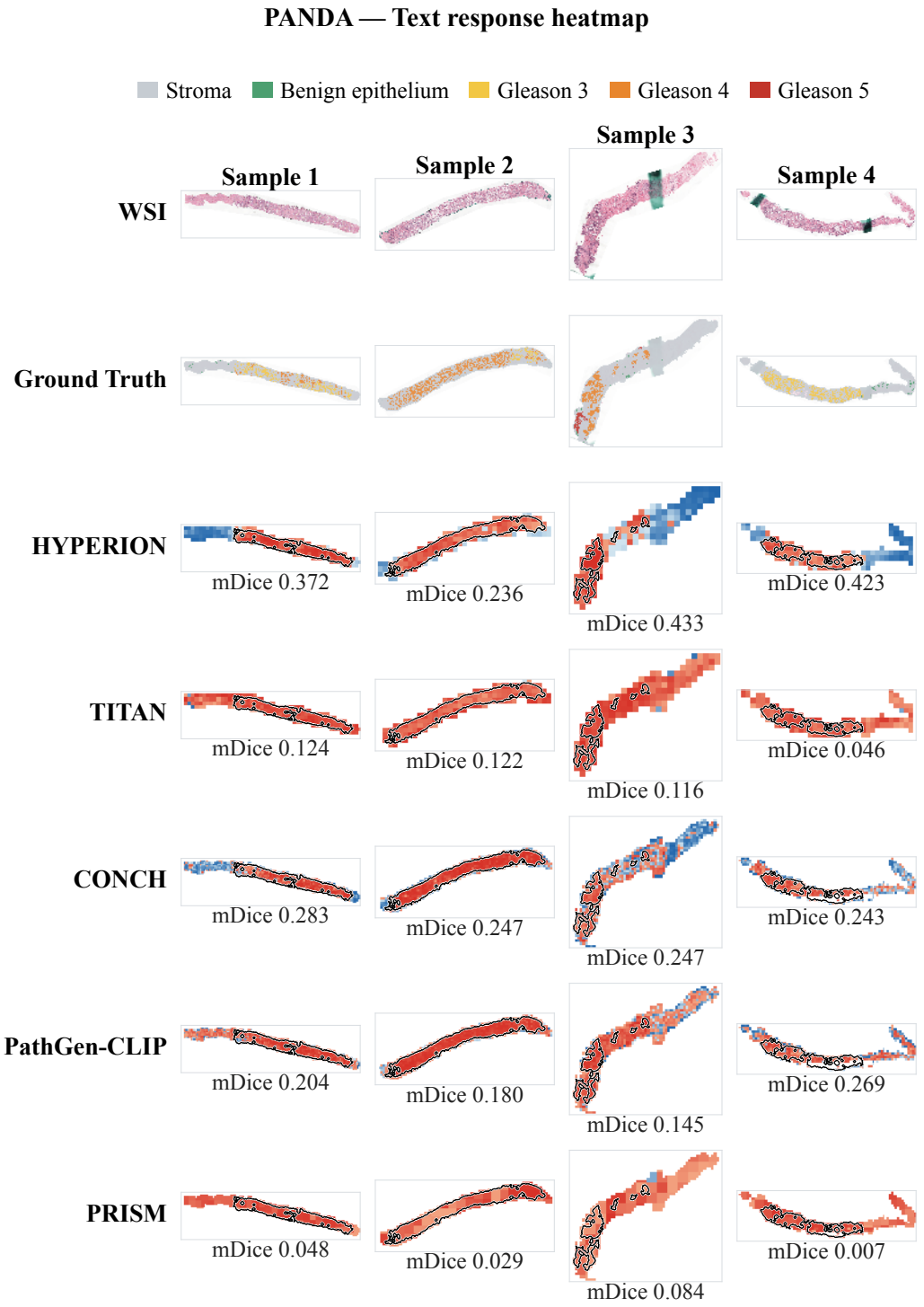

Figure A7: **Zero-shot tumor segmentation on PANDA.** Columns show four selected slides, while rows show the H&E image, five-class pixel annotation, and signed-percentile text-response heatmaps from each method. Model responses are computed as $\max(\mathrm{G3}, \mathrm{G4}, \mathrm{G5}) - \max(\mathrm{stroma}, \mathrm{benign})$; red and blue denote cancer-side and non-cancer-side responses, respectively. A black outline marks a shared ground-truth cancer region derived from the Gleason 3–5 regions of the pixel annotation and simplified for display, while five-class mDice is computed on each model's native grid and reported below each model panel.

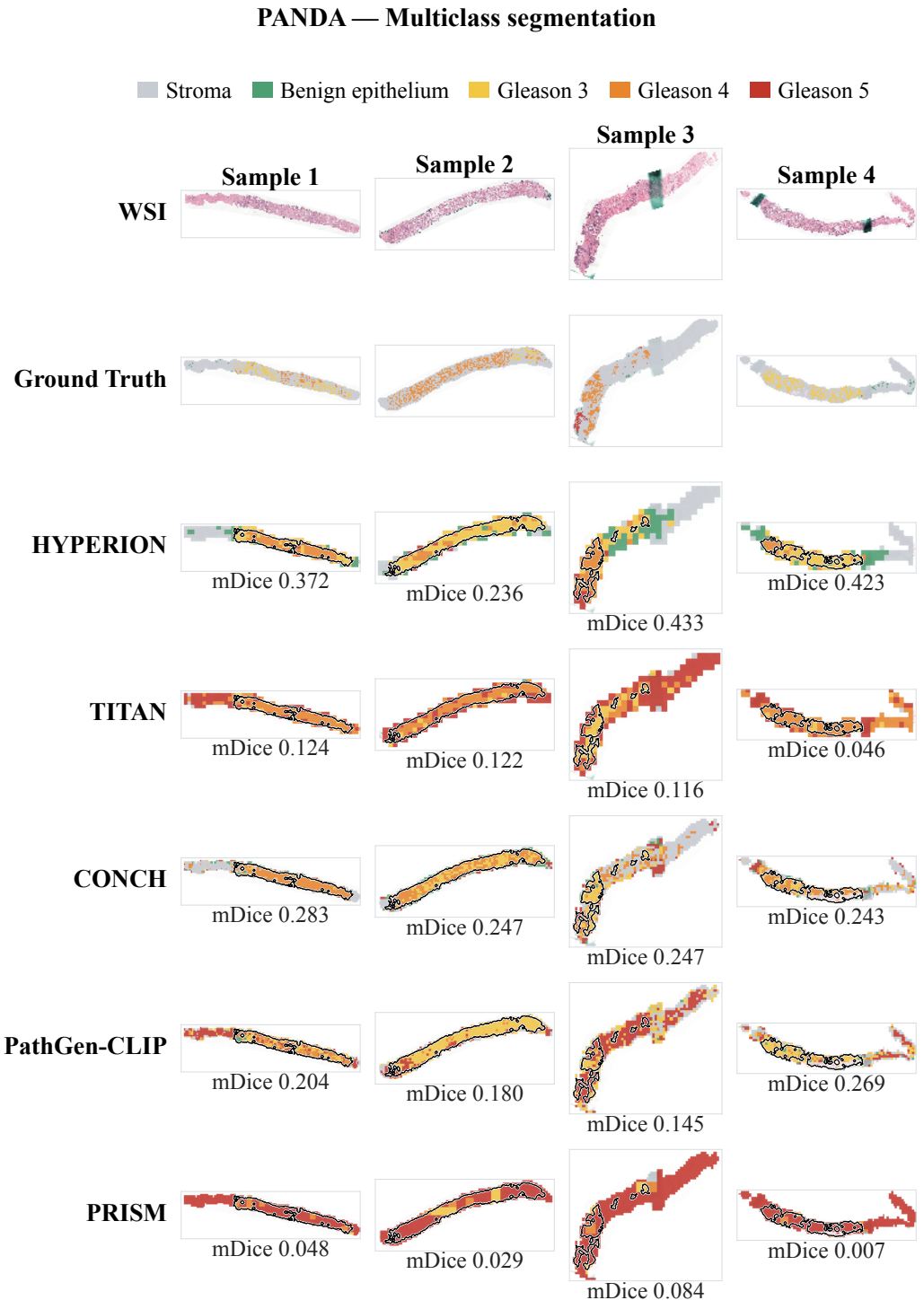

Figure A8: **Zero-shot five-class segmentation on PANDA.** Using the same slides and layout as Fig. A7, signed-percentile response heatmaps are replaced by native-grid five-class argmax predictions. Colors denote stroma, benign epithelium, and Gleason patterns 3, 4, and 5; per-slide five-class mDice is shown below each prediction.

E.2 Additional MLLM Grounding Examples

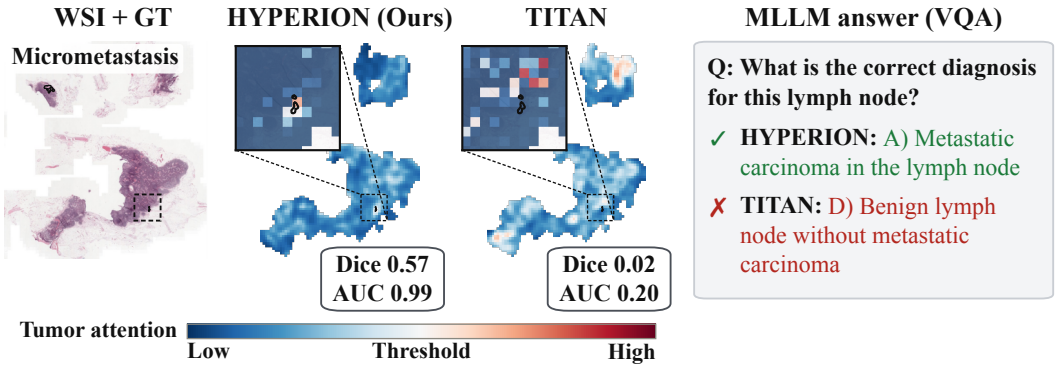

Figure A9: **Visual grounding inside the MLLM, micrometastasis.** Layout as in Fig. 5; the inset is 2.7 mm wide. Right: the diagnosis question of CM16-Bench with both answers (correct, incorrect).

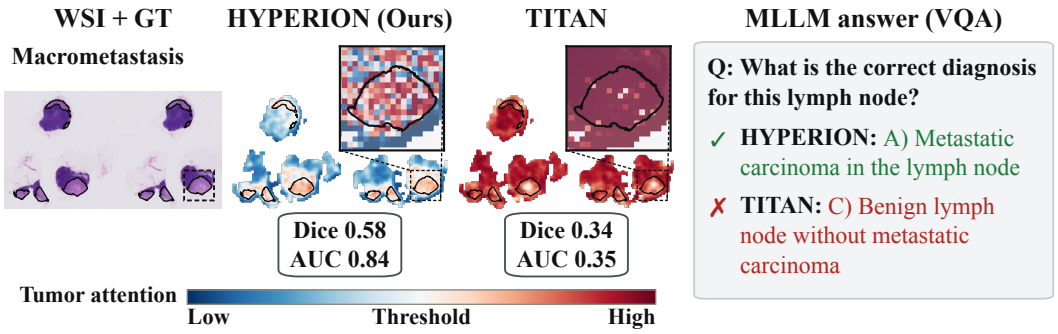

Figure A10: **Visual grounding inside the MLLM, multiple macrometastases.** Layout as in Fig. 5; metastases cover 14% of the tissue, and the inset is 5.0 mm wide. Right: the diagnosis question of CM16-Bench with both answers (correct, incorrect).

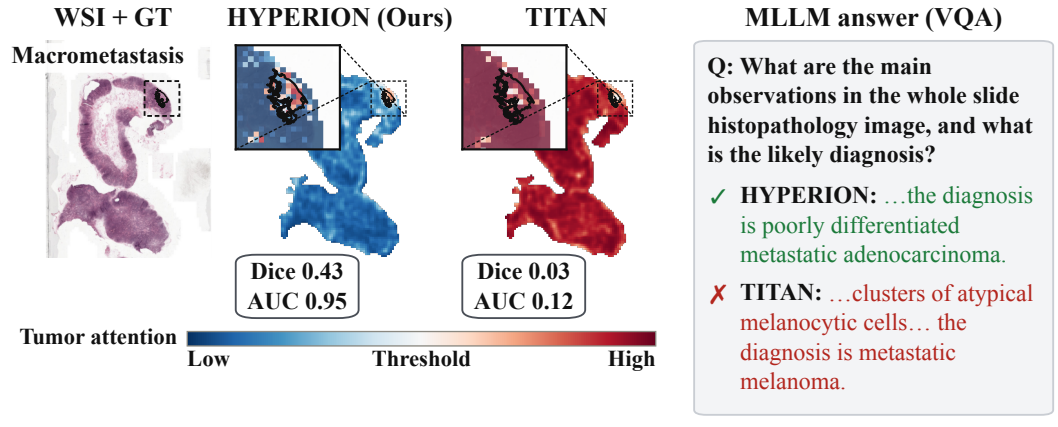

Figure A11: **Visual grounding inside the MLLM, single macrometastasis.** Layout as in Fig. 5; the inset is 5.0 mm wide. Right: the report question of CM16-Bench with excerpts of both reports (correct, incorrect diagnosis).

F LIMITATIONS

Although HYPERION substantially improves fine-grained visual grounding, stronger grounding does not necessarily translate proportionally into improvements in downstream language understanding or slide-level prediction. In particular, an MLLM may benefit not only from more accurately localized pathological evidence but also from access to a broader and more diverse set of visual features that capture global tissue context, morphology, and other complementary cues. This may explain why the gains in grounding are larger than those observed in report generation and prediction. Nevertheless, our results show that improving the grounding quality of pretrained slide representations can consistently preserve or enhance their downstream utility across segmentation, prediction, and language-generation tasks. We therefore view grounding as a necessary component of general-purpose pathological representation learning, and expect future work to combine stronger grounding with richer and more diverse feature representations for further improvements in MLLM reasoning and downstream prediction.