BEYOND APPEARANCE: PROTEIN-GROUNDED RELATIONAL REASONING FOR STRUCTURAL PATHOLOGICAL EVENT RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern computational pathology has achieved remarkable progress in cancer diagnosis, yet most models still make decisions primarily from histological appearance. This paradigm becomes insufficient for structural pathological events, where visually similar tissue environments can imply opposite diagnoses simply because their constituents are organized differently. We term this ambiguity **histological isomerism**: a model may recognize the right constituents yet still reach the wrong diagnosis if it fails to understand how they constitute a pathological event. We present **FRAME**, which moves computational pathology beyond appearance toward biologically grounded structural reasoning. FRAME learns molecularly grounded H&E representations from limited paired H&E–mIF data, organizes local evidence into anatomical targets and their surrounding constituents, and reasons about how these constituents are configured relative to target boundaries. mIF is required only during training, while inference remains H&E-only. With only hundreds to thousands of task-specific training samples, FRAME outperforms large-scale pathology foundation-model baselines by 2.5–5.1 points on general cancer diagnosis and grading. With only dozens of pathologist-corrected ROIs, it further outperforms the strongest pathology foundation-model baselines by an average of 12.5 points on composition-matched histological isomerism hard subsets.

1 INTRODUCTION

Modern computational pathology has made remarkable progress in extracting diagnostic information from routine hematoxylin and eosin (H&E) images (Chen et al., 2024; Xu et al., 2024; Ding et al., 2025; Vorontsov et al., 2026). Increasingly powerful representations, together with advances in whole-slide and multimodal modeling, have substantially improved the accuracy and data efficiency of cancer diagnosis. Recent methods have further begun to move beyond isolated patches by incorporating broader tissue context and complementary biological knowledge (Zhang et al., 2026; Vorontsov et al., 2026). Yet a common premise remains: diagnostic decisions are ultimately derived from what tissue looks like. **Not all pathological diagnoses, however, are defined by appearance.** A clinically important class of findings, including perineural, vascular, and lymphovascular invasion, is inherently structural: the diagnosis depends not only on which biological constituents are present, but on how they are spatially organized around a specific anatomical target (Seferbekova et al., 2023; Yaniv et al., 2024). For such events, recognizing tumor, nerve, vessel, or stroma is only the beginning of diagnosis; the decisive evidence lies in the configuration they collectively form.

This distinction (Mo et al., 2024; Lázár & Lundeberg, 2026; Ma et al., 2026) exposes a fundamental blind spot of appearance-centric pathology models. Two tissue environments can exhibit similar local morphology and nearly identical constituent composition, yet correspond to opposite diagnoses because their relative geometry differs. We refer to this phenomenon as **histological isomerism**: composition-equivalent tissue environments that encode different pathological events through different spatial configurations. It defines a challenging learning problem in which **what is present is insufficient to determine what event has formed**. A model may correctly recognize every relevant constituent and still reach the wrong diagnosis if it fails to understand how they are organized relative to the anatomical target. More visual context alone does not resolve this ambiguity; what is missing is an explicit representation of the structural relations that define the event.

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

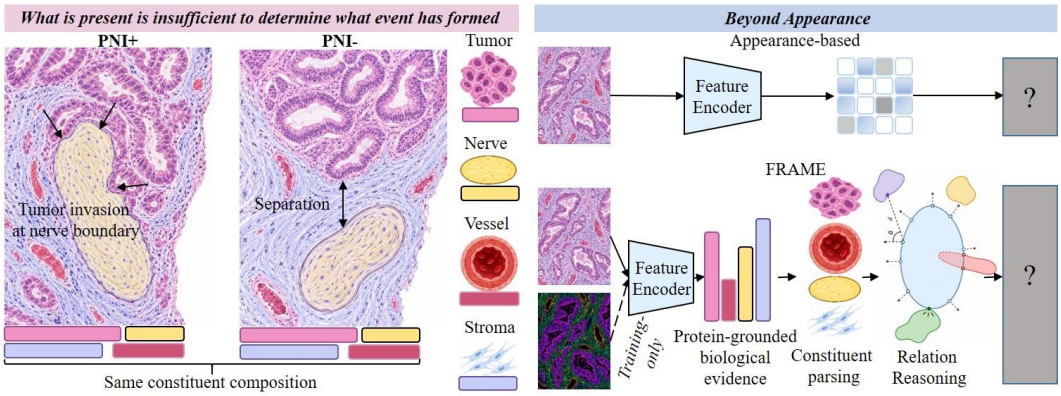


Figure 1: **Motivation of FRAME.** Left, **histological isomerism** shows that similar tissue composition can lead to opposite diagnoses under different target-relative configurations. Right, **FRAME** moves beyond appearance by combining training-time H&E–mIF grounding, target–constituent parsing, and boundary-centric relational reasoning. mIF is required only during training stage.

Learning such relations first requires reliable biological grounding of local morphology. Fine-grained structural tasks conventionally rely on expert annotations of small targets, constituent identities, or lesion boundaries, which are costly to acquire and inevitably subject to annotation variability. Molecular measurements offer a complementary source of information that is dense, spatially resolved, and biologically meaningful. Recent studies have demonstrated that protein and other molecular information can be recovered from, or transferred into, routine H&E representations (Andani et al., 2025; Wu et al., 2025; Farndale et al., 2025; Guo et al., 2026). We take a different step: rather than treating molecular information as the final prediction target or merely as a stronger representation prior, we use it as **privileged biological evidence** for learning how local H&E morphology should be interpreted. This makes it possible to ground tissue constituents biologically without requiring exhaustive manual semantic annotation, while preserving H&E-only inference.

To this end, we introduce **FRAME**, which moves computational pathology beyond appearance toward biologically grounded structural reasoning. FRAME progressively transforms local morphology into increasingly structured diagnostic evidence. It first learns protein-grounded H&E representations from a limited collection of spatially registered same-section H&E–mIF images. Rather than classifying predefined image patches, FRAME then organizes these local cues into anatomical targets and their surrounding constituents, allowing the diagnostic support to adapt to each target. Finally, FRAME uses the **target boundary as the reference frame for diagnosis**. It explicitly models how surrounding constituents are configured relative to this boundary and progressively integrates local relations into a global contour configuration. In this way, diagnosis is no longer based on whether relevant tissue patterns merely appear together, but on whether they form the spatial organization that defines a pathological event. mIF is required only during training, while the entire inference pipeline operates solely on routine H&E images at deployment in real-world settings. Our key contributions are as follows:

- The concept of **histological isomerism** is introduced to formulate structural pathological event recognition beyond appearance, characterizing cases where compositionally similar tissue environments yield different diagnoses through target-relative configurations.
- **FRAME** is developed to transform protein-grounded H&E evidence into anatomical targets and their constituents, followed by boundary-centric relational reasoning from local interactions to global structural configurations, while requiring only H&E at inference.
- Comprehensive experiments demonstrate FRAME's effectiveness and data efficiency. With hundreds to thousands of task-specific samples, FRAME outperforms large-scale pathology foundation models by 2.5–5.1 points on cancer diagnosis and grading. With dozens of pathologist-corrected ROIs, it improves over the strongest foundation-model baselines by an average of 12.5 points on composition-matched histological isomerism subsets.

2 Related Work

2.1 Computational Pathology Representation Learning

Whole-slide modeling has progressively evolved from weakly supervised aggregation of local patch evidence (Lu et al., 2021; Shao et al., 2021) toward hierarchical tissue modeling (Chen et al., 2022) and large-scale pathology foundation models (Chen et al., 2024; Xu et al., 2024; Ding et al., 2025; Vorontsov et al., 2026). These advances have substantially improved the transferability and data efficiency of histopathological representations. More recent work further moves beyond uniform patch representations by adapting the visual support to tissue organization or explicitly encoding global spatial structure (Zhang et al., 2026; Abousamra et al., 2026). Collectively, this line of research has greatly expanded *what* morphological evidence can be represented and *how much* tissue context can be incorporated. FRAME addresses a different question: when a diagnosis is defined by a structural event, richer tissue representations alone do not specify *how* the observed constituents form that event. Our goal is therefore not to construct another general-purpose slide representation, but to transform local histological evidence into an explicit target-relative configuration for diagnosis.

2.2 Molecular Supervision for Histopathology

Molecular measurements provide biological information that is not directly observable from H&E images. Recent studies exploit such information for histopathology representation learning. TANGLE aligns whole-slide morphology with transcriptomic profiles (Jaume et al., 2024), while TriDeNT and MoMKD use non-routine molecular modalities as privileged knowledge during training and retain histology-only inference (Farndale et al., 2025; Guo et al., 2026). CARE further incorporates transcriptomic and proteomic guidance into adaptive tissue representation learning (Zhang et al., 2026). Complementarily, HistoPlexer and ROSIE demonstrate that spatially resolved multiplex protein signals can be inferred directly from H&E morphology (Andani et al., 2025; Wu et al., 2025). These studies establish a connection between histological appearance and underlying molecular information. FRAME assigns molecular supervision a different role: protein information is neither the final prediction target nor merely a representation-level regularizer. Instead, inferred protein cues serve as *intermediate biological evidence* that guides the parsing of anatomical targets and surrounding constituents, which are subsequently used for structural reasoning.

2.3 Structural and Relational Reasoning

Spatial organization is an important source of information in histopathology. Existing methods capture such structure through correlations among instances (Shao et al., 2021), hierarchical tissue context (Chen et al., 2022), learned graph interactions (Li et al., 2024), or slide-level topological organization (Abousamra et al., 2026). These approaches show that tissue-region relationships provide information beyond independent patch appearance. However, their relations are generally induced from patch connectivity, feature similarity, learned interactions, or global tissue topology. Structural pathological events impose a more specific relational structure: the diagnostic meaning of a constituent depends on its geometry with respect to a particular anatomical target. FRAME therefore makes the *target boundary* the reference frame of reasoning. It explicitly represents constituent-to-boundary relations and organizes them hierarchically from local boundary interactions to global contour configurations. In this sense, FRAME differs from generic spatial aggregation or graph reasoning by making the pathology-defined geometry of an event an explicit part of the representation.

3 Method

3.1 Problem Formulation

We study *structural pathological event recognition*, where diagnosis depends not only on which tissue constituents are present, but also on how they are spatially organized with respect to a specific anatomical target. This distinguishes structural events from conventional appearance-based recognition: two tissue environments can contain similar tumor, stroma, vessels, and target structures while corresponding to different diagnoses because their relative configurations differ.

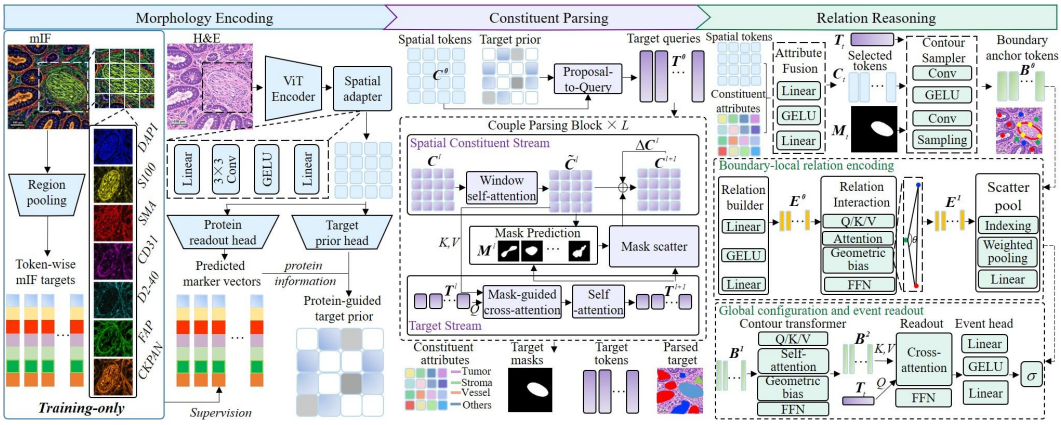


Figure 2: **Overview of FRAME.** FRAME comprises Morphology Encoding, Constituent Parsing, and Relation Reasoning. Training-time H&E–mIF supervision grounds representations with protein cues; parsing recovers targets and surrounding constituents; and boundary-centric reasoning integrates relations into contour configurations for event prediction. Inference requires only H&E.

We refer to this ambiguity as *histological isomerism*. For two tissue environments a and b,

$$\mathrm{Comp}(a) \approx \mathrm{Comp}(b), \qquad \mathrm{Geo}(a) \neq \mathrm{Geo}(b) \quad \Rightarrow \quad y_a \neq y_b, \tag{1}$$

where $\mathrm{Comp}(\cdot)$ denotes constituent composition and $\mathrm{Geo}(\cdot)$ denotes the target-relative configuration. Eq. 1 captures the central challenge illustrated in Fig. 1: recognizing *what is present* is insufficient when the event is defined by *how it is organized* across structurally defined events.

FRAME addresses this problem in three stages (Fig. 2). **Morphology Encoding** grounds H&E with molecular evidence, **Constituent Parsing** organizes evidence into anatomical targets and constituents, and **Relation Reasoning** models target-boundary configurations for event prediction.

3.2 Morphology Encoding via Molecular Grounding

Given an H&E image, ViT encoder produces spatial patch representations. FRAME preserves them rather than collapsing them globally. Lightweight Spatial Adapter, implemented as *Linear–3×3 Conv–GELU–Linear*, projects backbone features into the feature space and enables local information exchange while preserving token locations. Its output F denotes the morphology spatial tokens.

During training, spatially registered same-section mIF images provide token-aligned molecular supervision for each aligned spatial location. A Protein Readout Head predicts protein vectors $\hat{P}$ from F. The predicted molecular cues are then fused with morphology to produce the protein-grounded spatial tokens C^0, while a parallel Target Prior Head produces a soft target prior S. Together, these representations initialize target queries T^0 through proposal-conditioned pooling:

$$\hat{P} = g_{\mathrm{prot}}(F), \qquad C^0 = g_{\mathrm{fuse}}(F, \hat{P}), \qquad S = g_{\mathrm{prior}}(F, \hat{P}), \qquad T^0 = g_{\mathrm{prop}}(C^0, S). \tag{2}$$

Importantly, $\hat{P}$ is treated as *biological evidence*, not a semantic tissue label. Individual protein markers are not assumed to uniquely determine anatomical identity; instead, they complement morphology with molecular cues interpreted jointly with tissue context. Likewise, S indicates likely target locations rather than serving as a final target mask. Consequently, Morphology Encoding provides biologically grounded evidence for parsing without introducing mIF dependency at inference.

3.3 Coupled Constituent Parsing

Protein-grounded spatial evidence does not specify complete anatomical structures. FRAME therefore maintains two coupled representations: spatial constituent tokens C^l, which remain attached to fixed image locations, and target tokens T^l, representing candidate anatomical instances.

Within the l-th Coupled Parsing Block, window self-attention updates C^l to $\tilde{C}^l$ while preserving spatial grid. The target stream uses T^l as queries to read from $\tilde{C}^l$, with the preceding target mask M^{l-1} providing spatial guidance. The updated target representation T^{l+1} and spatial features jointly predict the mask M^l. Finally, Mask Scatter writes target-level context back to the spatial stream:

$$\begin{aligned} \tilde{C}^l &= \mathrm{WSA}(C^l), \\ T^{l+1} &= \mathcal{F}_T\Big(T^l, \tilde{C}^l; M^{l-1}\Big), \\ M^l &= \mathcal{F}_M\Big(\tilde{C}^l, T^{l+1}\Big), \\ \Delta C^l &= \mathcal{F}_S\big(M^l, T^{l+1}\big), \\ C^{l+1} &= \tilde{C}^l + \Delta C^l. \end{aligned} \tag{3}$$

Here, $\mathcal{F}_T$ denotes mask-guided target updating, $\mathcal{F}_M$ Mask Prediction, and $\mathcal{F}_S$ Mask Scatter. First block uses Stage-1 prior S as initial guidance; subsequent blocks use preceding prediction M^{l-1}.

This bidirectional interaction allows spatial evidence to determine the extent of a target while the emerging target instance provides context for interpreting its surrounding constituents. After L blocks, FRAME obtains C^L, M^L, and T^L. An Attribute Head further predicts non-exclusive constituent attributes A from C^L, avoiding the assumption that every location belongs to a single mutually exclusive tissue category. For geometric reasoning, the selected target mask is optionally refined to a higher-resolution mask $\bar{M}_t$ to provide an accurate target contour for subsequent reasoning.

3.4 Boundary-Centric Relation Reasoning

After parsing, the remaining question is no longer which structures are present, but how they are configured. FRAME therefore uses the target boundary, rather than a target center or global target feature, as the geometric reference frame that directly defines the event.

For each valid target t, constituent attributes A are fused with C^L, and target-conditioned selection retains the spatial evidence relevant to that target, yielding C_t. The selected support can vary in size with the target and its surrounding tissue context, avoiding a predefined diagnostic crop. In parallel, the refined target mask $\bar{M}_t$ is converted into a contour and sampled into boundary anchor tokens B_t^0.

FRAME then constructs relations between selected constituents and nearby boundary anchors. Each relation incorporates the constituent and boundary representations together with target-relative geometry, including distance, orientation, inside/outside state, and normalized relative position:

$$E_t^0 = \mathcal{R}\big(C_t, B_t^0; d, \theta, \mathrm{inside/outside}, \rho\big), \tag{4}$$

where $\mathcal{R}$ denotes the Relation Builder and ρ denotes normalized relative position.

The resulting relation tokens E_t^0 are updated by a Geometry-Conditioned Relation Interaction module to obtain E_t^1. Geometry modulates relation interaction rather than serving only as positional metadata, allowing compositionally similar environments with different configurations to remain distinguishable. Relations associated with the same boundary anchor are then aggregated through Scatter Pool into B_t^1, which summarizes local boundary evidence.

Local interactions do not fully characterize a structural event. A Contour Transformer propagates information along the target boundary to obtain B_t^2. Its geometric bias incorporates cues such as arc distance, boundary-normal relationships, curvature, and contour neighborhood. This converts local constituent–boundary evidence into a global configuration of the target contour.

Finally, the target token is used only as a query to read the resulting boundary configuration:

$$z_t = \mathrm{CrossAttn}\left(Q = T_t^L,\ K = V = B_t^2\right), \qquad p_t = g_{\mathrm{event}}(z_t), \tag{5}$$

where z_t is the target-level event representation and p_t is the predicted structural-event probability. Using T_t^L only as the readout query keeps the event decision grounded in the boundary representation rather than directly classifying target appearance.

3.5 Training Objectives and H&E-only Inference

FRAME is supervised at three levels corresponding to its representation hierarchy. Molecular supervision trains the Protein Readout Head to recover token-wise protein evidence; parsing supervision

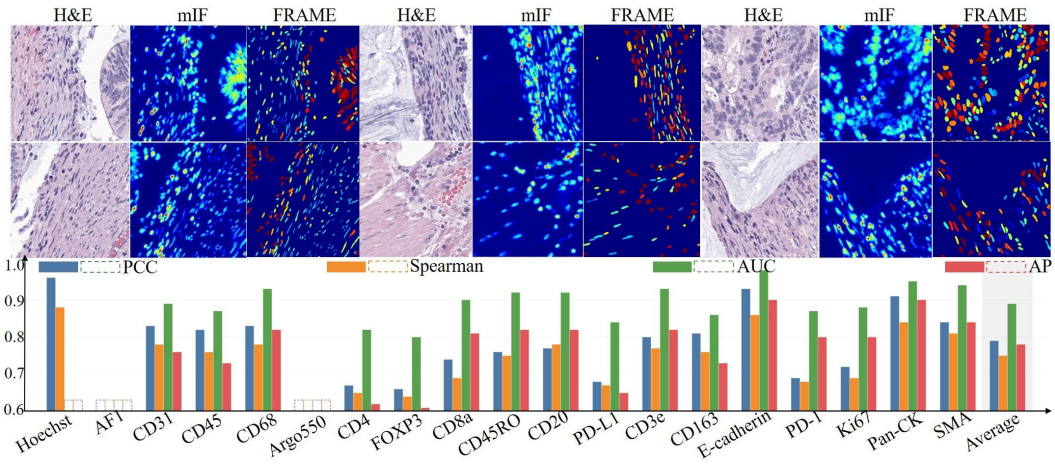


Figure 3: **Protein-grounded morphology learning from H&E images.** Upper: FRAME-predicted spatial protein maps compared with registered mIF measurements. Lower: quantitative evaluation across multiplex channels using PCC, Spearman, AUROC, and AP. **Dashed empty bars denote unavailable metrics for non-protein channels or missing binary evaluation**.

constrains the target prior, target masks, and constituent attributes; and target-level labels supervise the final structural-event prediction during joint training. The overall objective is

$$\mathcal{L} = \lambda_{\text{prot}}\mathcal{L}_{\text{prot}} + \lambda_{\text{parse}}\mathcal{L}_{\text{parse}} + \lambda_{\text{event}}\mathcal{L}_{\text{event}}, \tag{6}$$

where the three terms correspond to molecular grounding, constituent–target parsing, and event recognition, respectively. The realization of each loss is provided in the implementation details.

Paired mIF is used only for training-time molecular supervision. At inference, mIF images, Region Pooling, and molecular targets are removed, while the Protein Readout Head predicts $\hat{P}$ directly from the H&E morphology tokens F. The resulting C^0, S, and T^0 are passed through the same constituent parsing and boundary-centric reasoning pipeline to obtain p_t. Thus, FRAME retains molecularly grounded structural reasoning while requiring only routine H&E images at deployment.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate FRAME along three tracks. General cancer classification is assessed on TCGA-CRC (The Cancer Genome Atlas Research Network et al., 2013), PANDA (Bulten et al., 2022), and CPTAC-CCRCC (Clark et al., 2019). Structural-event and fine-grained lesion recognition uses PAIP 2021 (Choi et al., 2021), CAMELYON16 (Ehteshami Bejnordi et al., 2017), TLSD, and PIND. ORION-CRC (Lin et al., 2023a) evaluates molecular grounding by comparing H&E-predicted protein signals with registered fluorescence measurements. Dataset statistics and task definitions are detailed in Appendix A.1.

Paired H&E–mIF data. Molecular grounding is trained using spatially registered same-section H&E–mIF pairs from internal tumor-specific multiplex panels. mIF is acquired before H&E restaining on the same physical section, enabling direct spatial alignment between molecular measurements and histological morphology. mIF is used only during training, whereas all downstream inference remains H&E-only. Acquisition protocols and marker panels are provided in Appendix A.2.

Baselines. For cancer classification, we compare FRAME with representative GNN-based, MIL-based, and pathology foundation models. For structural-event recognition, we include CNN-, Transformer-, and SAM-based segmentation methods. All methods use identical partitions and supervision budgets when applicable. Detailed baseline configurations are provided in Appendix A.4.

Table 1: AUC comparison with pathology foundation models on PANDA, CPTAC-CCRCC, and TCGA-CRC for slide-level diagnosis and grading. **Bold** denotes the best result. **Detailed comparisons with GNN-based and MIL-based baselines** are provided in Appendix Table 6.

Architecture	Method	Pretraining Samples	PANDA		CPTAC-CCRCC		TCGA-CRC	
			Diagnosis	Grading	Diagnosis	Grading	Diagnosis	Grading
PFM-based	HIPT	10K+	0.858	0.825	0.733	0.697	0.858	0.849
	CTrans	35K+	0.930	0.927	0.772	0.742	0.845	0.838
	UNI	100,000+	0.948	0.944	0.824	0.779	0.870	0.866
	CHIEF	60K+	<u>0.953</u>	0.941	0.839	0.801	<u>0.896</u>	<u>0.878</u>
	CONCH	117,000+	0.949	0.934	<u>0.842</u>	<u>0.813</u>	0.873	0.862
Ours	FRAME	–	**0.981**	**0.973**	**0.867**	**0.858**	**0.937**	**0.929**
	Δ vs PFM		+2.8%	+2.9%	+2.5%	+4.5%	+4.1%	+5.1%

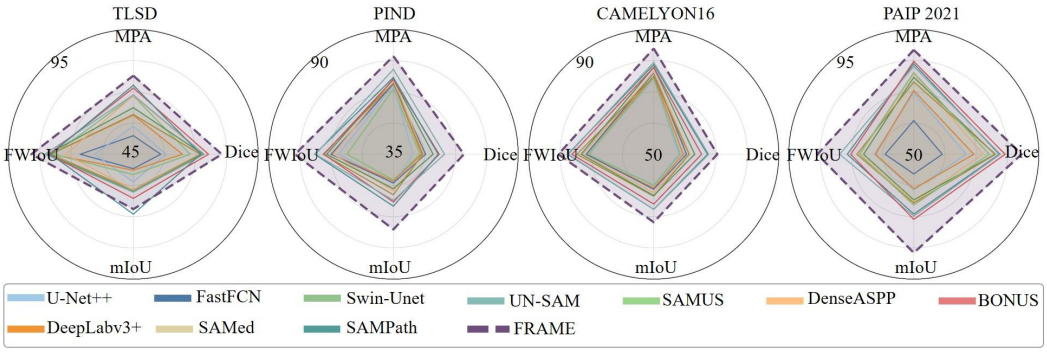


Figure 4: **Performance comparison of Coupled Constituent Parsing.** Radar charts summarize MPA, Dice, mIoU, and FWIoU on TLSD, PIND, CAMELYON16, and PAIP 2021 datasets.

4.2 Result Analysis

4.2.1 Morphology Encoding recovers spatial protein information from H&E

We first evaluate whether Morphology Encoding captures biologically meaningful information beyond visual appearance by comparing FRAME-predicted protein signals with spatially registered mIF measurements on ORION-CRC. As shown in Fig. 3, FRAME recovers spatial protein patterns that closely follow the corresponding mIF distributions across epithelial, stromal, vascular, and immune-associated markers. This agreement indicates that H&E morphology can provide spatially resolved molecular evidence for subsequent constituent parsing and structural reasoning.

Quantitatively, FRAME achieves an average PCC of 0.79 and Spearman correlation of 0.75 across the evaluated protein channels, together with an average AUROC of 0.89 and AP of 0.78 for marker-positive region identification. Strong recovery is observed for representative epithelial and structural markers, including E-cadherin (PCC=0.93, AUROC=0.98), Pan-CK (PCC=0.91, AUROC=0.95), and SMA (PCC=0.84, AUROC=0.94). Detailed channel-wise results are in Appendix Table 4.

4.2.2 Coupled Constituent Parsing improves target and constituent parsing

We evaluate Coupled Constituent Parsing against CNN-, Transformer-, and SAM-based segmentation methods on TLSD, PIND, CAMELYON16, and PAIP. As shown in Fig. 4, FRAME achieves the strongest performance, demonstrating accurate target delineation and constituent localization.

FRAME reaches Dice scores of 80.45%, 82.45%, 70.36%, and 75.44% on TLSD, PIND, CAMELYON16, and PAIP, outperforming the strongest baselines by 5.31, 7.26, 2.75, and 6.97 points, respectively. On PIND, it achieves 78.42% mIoU, indicating robust parsing under complex tissue

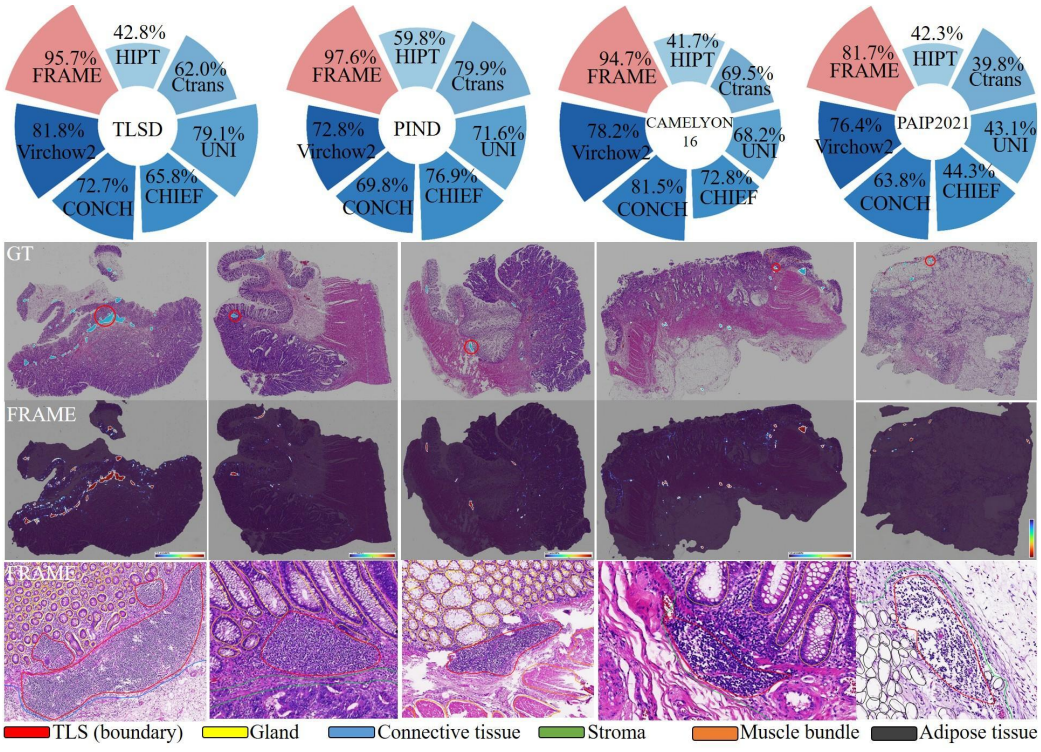


Figure 5: **Boundary-centric reasoning improves structural diagnosis beyond tissue appearance.** Upper: performance on composition-matched histological isomerism subsets across four structural-event datasets. Lower: representative examples of FRAME identifying tertiary lymphoid structures through constituent organization and boundary-centric structural reasoning.

organization. These results support the effectiveness of coupled target–constituent modeling across diverse lesion scales and pathological contexts. Detailed results are in Appendix Table 5.

4.2.3 Boundary-centric reasoning improves diagnosis beyond appearance

We first compare FRAME with pathology foundation models on general cancer diagnosis and grading. As shown in Table 1, FRAME achieves the best performance across PANDA, CPTAC-CCRCC, and TCGA-CRC, with AUCs of 0.981/0.973, 0.867/0.858, and 0.937/0.929 for diagnosis/grading, respectively. These consistent gains across different cancer types correspond to improvements of 2.5–5.1 points over the strongest foundation-model baselines despite substantially smaller task-specific training sets. Comparisons with GNN- and MIL-based methods are in Appendix Table 6.

We further evaluate histological isomerism using composition-matched hard subsets. For each lesion ROI, we compute a constituent composition vector $r_t = [p_{\text{tumor}}, p_{\text{target}}, p_{\text{stroma}}, \ldots, p_{\text{vessel}}]$ and select oppositely labeled pairs satisfying $\|r_i - r_j\| < \epsilon$. As shown in Fig. 5, FRAME achieves 95.7%, 97.6%, 94.7%, and 81.7% accuracy on TLSD, PIND, CAMELYON16, and PAIP2021, respectively, outperforming the strongest foundation-model baselines by an average of 12.5 points. The larger gains on composition-matched cases indicate that FRAME benefits from modeling target-relative structural relations rather than appearance alone.

4.2.4 Ablation and annotation-efficiency analysis

We conduct systematic ablation studies to validate the contribution of each component in FRAME and evaluate its annotation efficiency under progressively reduced expert supervision. Detailed quantitative results for all experimental variants are provided in Appendix Tables 7, 8, and 9.

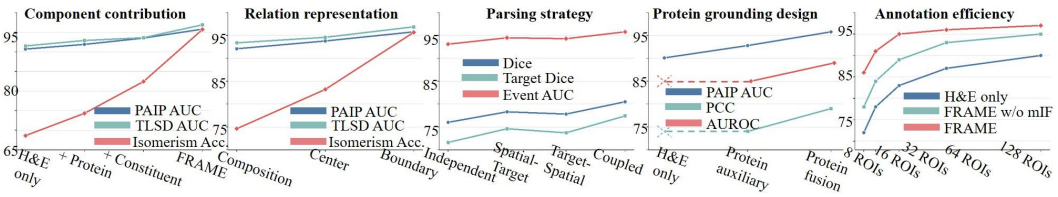


Figure 6: **Ablation and annotation-efficiency analysis of FRAME.** Ablation studies evaluate the contribution of major components and design choices, while annotation-efficiency analysis measures performance under varying numbers of physician-corrected ROIs.

Component contribution. We evaluate the progressive contribution of three major components in FRAME. Starting from an H&E-only appearance model, introducing molecular grounding, constituent parsing, and boundary-centric reasoning leads to consistent improvements across both standard and hard structural cases. The complete FRAME improves the histological isomerism recognition accuracy from 68.6% to 95.7%, while increasing PAIP AUC from 0.907 to 0.958 and TLSD AUC from 0.915 to 0.969. These results demonstrate that each stage provides complementary information, and that explicit structural reasoning is essential beyond appearance-based representation.

Boundary-centric relation representation. We investigate whether target-relative geometry is necessary for structural diagnosis across compositionally similar tissue environments. Replacing boundary-centric relations with composition-only or center-based geometry reduces isomerism recognition accuracy from 95.7% to 74.6% and 83.2%, respectively. The consistent improvement of boundary-centric reasoning across PAIP and TLSD confirms pathological events are determined by constituent configurations relative to anatomical boundaries rather than constituent presence alone.

Coupled constituent parsing. We further examine the interaction strategy between constituent and target representations. Compared with independent parsing or one-way interactions, coupled parsing achieves the best constituent localization and event recognition performance, improving Dice to 80.5%, Target Dice to 77.4%, and Event AUC to 0.958. These results support the necessity of jointly refining target identity and surrounding constituent interpretation during structural parsing.

Protein grounding design. We compare direct H&E modeling, protein prediction as an auxiliary objective, and protein-guided fusion. Protein-guided fusion improves protein recovery (PCC: 0.79, AUROC: 0.89) and downstream diagnosis (PAIP AUC: 0.958), outperforming auxiliary protein prediction alone (PCC: 0.74, AUROC: 0.85, PAIP AUC: 0.928). This indicates that molecular supervision provides effective biological grounding when incorporated into the representation rather than used only as an auxiliary constraint during optimization and downstream reasoning.

Annotation efficiency. Finally, we evaluate the dependence on expert annotations by varying the number of physician-corrected ROIs across progressively larger supervision budgets. FRAME consistently maintains superior performance under limited annotation budgets, achieving 0.86 AUC with only 8 corrected ROIs and 0.95 AUC with 32 ROIs, compared with 0.72 and 0.83 for the H&E-only baseline. These results demonstrate that molecular grounding and structural reasoning substantially reduce reliance on extensive manual annotation while preserving strong diagnostic performance under limited supervision across challenging structural-event recognition settings.

5 Conclusion

Structural pathological events reveal limitation of appearance-centric computational pathology: similar tissue compositions can imply different diagnoses when their spatial configurations differ. We formulate this challenge as **histological isomerism** and introduce **FRAME**, which combines molecularly grounded H&E representations, coupled constituent parsing, and boundary-centric reasoning to model pathological events. Experiments across cancer diagnosis, fine-grained lesion recognition, and composition-matched hard cases validate the effectiveness and data efficiency of FRAME. By shifting the focus from *what tissue looks like* to *how tissue constituents are organized*, FRAME provides a general direction for structure-aware computational pathology in real-world diagnosis.

AI USE STATEMENT

ChatGPT was used for language polishing, manuscript organization, and assistance with LaTeX formatting. The scientific ideas, methodological design, experimental protocols, code, figures, results, and conclusions were proposed, validated, and finalized by the authors. All AI-assisted text was reviewed and revised by the authors to ensure scientific accuracy and consistency with the actual experiments. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work presents a computational pathology framework aimed at improving the accuracy and interpretability of pathological image analysis. All experiments were conducted using either publicly available datasets or private data collected under approval from the relevant institutional review boards and in accordance with applicable ethical guidelines. By improving the recognition of compositionally similar yet structurally distinct pathological events, as formulated through **histological isomerism**, and by providing a more transparent structural reasoning process, this work has the potential to support pathologists and reduce variability in diagnostic assessment. While such advances may contribute to more consistent and standardized diagnostic workflows, we recognize that the integration of AI systems into clinical practice requires careful validation, engagement with relevant clinical stakeholders, and continued evaluation of their reliability and effectiveness in real-world.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of FRAME's modeling principles and architectural components in the Method section. The Appendix further documents the practical implementation of the framework, including dataset and task definitions, preprocessing procedures, paired H&E–mIF data acquisition and alignment, experimental settings, baseline configurations, and evaluation protocols. Detailed quantitative results and ablation studies are also provided to facilitate verification of the reported findings and reproduction of the experimental pipeline.

REFERENCES

Shahira Abousamra, Asmita Sood, and Sylvia Plevritis. Toposlide: Topologically-informed histopathology whole slide image representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, June 2026.

Sonali Andani, Boqi Chen, Ficek-Pascual, et al. Histopathology-based protein multiplex generation using deep learning. *Nature Machine Intelligence*, 7(8):1292–1307, 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01074-y. URL https://doi.org/10.1038/s42256-025-01074-y.

Wouter Bulten, Kimmo Kartasalo, Peter H. C. Chen, Petter Ström, Hans Pinckaers, Kunal Nagpal, Yukun Cai, David F. Steiner, Henk van Boven, Raoul Vink, Robert Hägerling, Geert Litjens, Jeroen A. W. M. van der Laak, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28:154–163, 2022. doi: 10.1038/s41591-021-01620-2. URL https://doi.org/10.1038/s41591-021-01620-2.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision Workshops(ECCVW)*, 2022.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, pp. 833–851, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01233-5. doi: 10.1007/978-3-030-01234-2_49. URL https://doi.org/10.1007/978-3-030-01234-2_49.

Richard J. Chen, Ming Y. Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware

survival prediction using patch-based graph convolutional networks. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 339–349, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87237-3.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16144–16155, June 2022.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL https://doi.org/10.1038/s41591-024-02857-3.

Zhen Chen, Qing Xu, Xinyu Liu, and Yixuan Yuan. Un-sam: Domain-adaptive self-prompt segmentation for universal nuclei images. *Medical Image Analysis*, pp. 103607, 2025.

Jinwook Choi, Kyoungbun Lee, Won-Ki Jeong, and Se Young Chun. PAIP2021: Perineural Invasion in Multiple Organ Cancer (Colon, Prostate, and Pancreatobiliary tract). International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2021. URL https://doi.org/10.5281/zenodo.4575424.

D. J. Clark, S. M. Dhanasekaran, F. Petralia, J. Pan, X. Song, Y. Hu, F. da Veiga Leprevost, B. Reva, T. M. Lih, H. Y. Chang, W. Ma, C. Huang, C. J. Ricketts, L. Chen, A. Krek, Y. Li, D. Rykunov, Q. K. Li, L. S. Chen, U. Ozbek, and Clinical Proteomic Tumor Analysis Consortium. Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell*, 179(4): 964–983.e31, 2019. doi: 10.1016/j.cell.2019.10.007. URL https://doi.org/10.1016/j.cell.2019.10.007.

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*, pp. 1–13, 2025.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. doi: 10.1001/jama.2017.14585.

Lucas Farndale, Robert Insall, and Ke Yuan. Trident : Triple deep network training for privileged knowledge distillation in histopathology. *Medical Image Analysis*, 102:103479, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103479. URL https://www.sciencedirect.com/science/article/pii/S1361841525000271.

Yongxin Guo, Hao Lu, Onur C. Koyun, Zhengjie Zhu, Muhammet F. Demir, and Metin N. Gurcan. Momentum memory for knowledge distillation in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6889–6899, June 2026.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Guillaume Jaume, Lukas Oldenburg, Anurag Vaidya, Richard J. Chen, Drew F.K. Williamson, Thomas Peeters, Andrew H. Song, and Faisal Mahmood. Transcriptomics-guided slide representation learning in computational pathology. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9632–9644, 2024. doi: 10.1109/CVPR52733.2024.00920.

Yongju Lee, Jeong Hwan Park, Sohee Oh, Kyoungseob Shin, Jiyu Sun, Minsun Jung, Cheol Lee, Hyojin Kim, Jin-Haeng Chung, Kyung Chul Moon, and Sunghoon Kwon. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nature Biomedical Engineering*, 6(12):1452–1466, 2022. ISSN 2157-846X. doi: 10.1038/s41551-022-00923-0. URL https://doi.org/10.1038/s41551-022-00923-0.

Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11323–11332, 2024. doi: 10.1109/CVPR52733.2024.01076.

Jia-Ren Lin, Yu-An Chen, Daniel Campton, Jeremy Cooper, Shannon Coy, Clarence Yapp, Juliann B. Tefft, Erin McCarty, Keith L. Ligon, Scott J. Rodig, Steven Reese, Tad George, Sandro Santagata, and Peter K. Sorger. High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. *Nature Cancer*, 4(7):1036–1052, 2023a. ISSN 2662-1347. doi: 10.1038/s43018-023-00576-1. URL https://doi.org/10.1038/s43018-023-00576-1.

Xian Lin, Yangyang Xiang, Li Zhang, Xin Yang, Zengqiang Yan, and Li Yu. Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation, 2023b.

Yi Lin, Zhiyong Qu, Hao Chen, Zhongke Gao, Yuexiang Li, Lili Xia, Kai Ma, Yefeng Zheng, and Kwang-Ting Cheng. Nuclei segmentation with point annotations from pathology images via self-supervised learning and co-training. *Medical Image Analysis*, 89:102933, 2023c. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102933. URL https://www.sciencedirect.com/science/article/pii/S1361841523001937.

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):555–570, 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w. URL https://doi.org/10.1038/s41551-020-00682-w.

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*, 30:863–874, 2024.

Enikő Lázár and Joakim Lundeberg. Spatial architecture of development and disease. *Nature Reviews Genetics*, 27(2):118–136, 2026. ISSN 1471-0064. doi: 10.1038/s41576-025-00892-5. URL https://doi.org/10.1038/s41576-025-00892-5.

Lichun Ma, Barbara Xiong, Meng Liu, and Kai Tan. Cellular neighborhoods in cancer. *Nature Cancer*, 7(1):16–28, 2026. ISSN 2662-1347. doi: 10.1038/s43018-025-01107-w. URL https://doi.org/10.1038/s43018-025-01107-w.

Chia-Kuei Mo, Jingxian Liu, Siqi Chen, et al. Tumour evolution and microenvironment interactions in 2D and 3D space. *Nature*, 634(8036):1178–1186, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-08087-4. URL https://doi.org/10.1038/s41586-024-08087-4.

Zaira Seferbekova, Artem Lomakin, Lucy R. Yates, and Moritz Gerstung. Spatial biology of cancer evolution. *Nature Reviews Genetics*, 24(5):295–313, 2023. ISSN 1471-0064. doi: 10.1038/s41576-022-00553-x. URL https://doi.org/10.1038/s41576-022-00553-x.

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. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

Liang Tang, Shuang Diao, Chao Li, Ming He, Kai Ru, and Wei Qin. Global contextual representation via graph-transformer fusion for hepatocellular carcinoma prognosis in whole-slide images. *Computerized Medical Imaging and Graphics*, 115:102378, 2024. doi: 10.1016/j.compmedimag.2024.102378. Epub 2024 Apr 16.

Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew F. K. Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 699–715, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19803-8.

The Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. M. 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. doi: 10.1038/ng.2764.

Eugene Vorontsov, George Shaikovski, Adam Casson, Julian Viret, Eric Zimmermann, Neil Tenenholtz, Yi Kan Wang, Jan H. Bernhard, Ran A. Godrich, Juan A. Retamero, Jinru Shia, Mithat Gonen, Martin R. Weiser, David S. Klimstra, Razik Yousfi, Nicolò Fusi, Thomas J. Fuchs, Kristen Severson, and Siqi Liu. End-to-end multimodal pathology foundation model with clinical dialogue. *Nature Medicine*, 32(9):3203–3213, 2026. ISSN 1546-170X. doi: 10.1038/s41591-026-04521-4. URL https://doi.org/10.1038/s41591-026-04521-4.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 2022.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, Fang Wang, Yulong Peng, Junyou Zhu, Jing Zhang, Christopher R. Jackson, Jun Zhang, Deborah Dillon, Nancy U. Lin, Lynette Sholl, Thomas Denize, David Meredith, Keith L. Ligon, Sabina Signoretti, Shuji Ogino, Jeffrey A. Golden, MacLean P. Nasrallah, Xiao Han, Sen Yang, and Kun-Hsing Yu. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, October 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07894-z. URL https://doi.org/10.1038/s41586-024-07894-z.

Eric Wu, Matthew Bieniosek, Zhenqin Wu, Nitya Thakkar, Gregory W. Charville, Ahmad Makky, Christian M. Schürch, Jeroen R. Huyghe, Ulrike Peters, Christopher I. Li, Li Li, Hannah Giba, Vivek Behera, Arjun Raman, Alexandro E. Trevino, Aaron T. Mayer, and James Zou. ROSIE: AI generation of multiplex immunofluorescence staining from histopathology images. *Nature Communications*, 16(1):7633, 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-62346-0. URL https://doi.org/10.1038/s41467-025-62346-0.

Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yu Yizhou. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. In *arXiv preprint arXiv:1903.11816*, 2019.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024.

Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Dan Yaniv, Brandi Mattson, Sebastien Talbot, Frederico O. Gleber-Netto, and Moran Amit. Targeting the peripheral neural-tumour microenvironment for cancer therapy. *Nature Reviews Drug Discovery*, 23(10):780–796, 2024. ISSN 1474-1784. doi: 10.1038/s41573-024-01017-z. URL https://doi.org/10.1038/s41573-024-01017-z.

Di Zhang, Zhangpeng Gong, Xiaobo Pang, Jiashuai Liu, Junbo Lu, Hao Cui, Jiusong Ge, Zhi Zeng, Kai Yi, Yinghua Li, Si Liu, Tingsong Yu, Haoran Wang, Mireia Crispin-Ortuzar, Weimiao Yu, Chen Li, and Zeyu Gao. Care: A molecular-guided foundation model with adaptive region modeling for whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21078–21088, June 2026.

Jingwei Zhang, Ke Ma, Saarthak Kapse, Joel Saltz, Maria Vakalopoulou, Prateek Prasanna, and Dimitris Samaras. Sam-path: A segment anything model for semantic segmentation in digital pathology. *arXiv preprint arXiv:2307.09570*, 2023.

Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2020. doi: 10.1109/TMI.2019.2959609.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

A Appendix

A.1 Benchmark Datasets

We organize the evaluation datasets according to the scientific claim they are intended to test. General cancer classification evaluates the transferability and data efficiency of FRAME representations. Structural-event and fine-grained lesion datasets evaluate target localization, constituent parsing, and event recognition. ORION-CRC is reserved for direct evaluation of molecular grounding quality.

Table 2: **Benchmark datasets used in this study.** Private cohorts are marked as *Private*. PAIP 2021 follows the challenge data-use procedure. ORION-CRC is used only for molecular grounding evaluation.

Dataset	Access	Scale	Primary role
General cancer classification			
TCGA-CRC (The Cancer Genome Atlas Research Network et al., 2013)	Public	551 WSIs	Colorectal cancer classification
PANDA (Bulten et al., 2022)	Public	10,616 WSIs	Prostate cancer diagnosis and grading
CPTAC-CCRCC (Clark et al., 2019)	Public	262 WSIs	Renal cancer classification
Structural and fine-grained recognition			
PAIP 2021 (Choi et al., 2021)	Controlled	240 WSIs	Cross-organ perineural invasion
CAMELYON16 (Ehteshami Bejnordi et al., 2017)	Public	400 WSIs	Metastatic lesion localization
TLSD	Private	356 WSIs	Structural-event recognition
PIND	Private	224 WSIs	Structural-event recognition
Molecular grounding			
ORION-CRC (Lin et al., 2023a)	Public	40 cases	H&E-to-protein recovery

PAIP 2021 contains H&E WSIs from colon, prostate, and pancreatobiliary cancers with expert annotations of nerve regions and perineural-invasion junctions. CAMELYON16 contains sentinel lymph-node WSIs with pixel-level metastatic-region annotations and is used to evaluate fine-grained lesion localization. Patient-level separation is maintained for the private TLSD, and PIND cohorts to avoid information leakage across training, validation, and test sets.

Table 3: **Representative multiplex fluorescence acquisition settings.** Each staining round uses a 40-min primary-antibody incubation followed by a 20-min secondary-antibody incubation.

Marker	Channel (nm)	Dilution
SMA	520	1:2000
S100	480	1:30000
CD31	570	1:800
FAP	690	1:500
D2-40	620	1:800
CKP	780	1:500

For CPTAC-CCRCC, we use the 262 slides retained by our preprocessing pipeline. When multiple tissue slides are available for the same subject, the slide-selection criterion is fixed before model training and applied consistently to all compared methods.

A.2 Paired H&E–mIF Data

Same-section acquisition. All paired data used for molecular grounding are acquired from the same physical tissue section. Adjacent-section material available for part of the collection is excluded from the experiments reported in this work, avoiding section-to-section morphological variation and enabling direct spatial association between molecular measurements and H&E morphology.

For internal samples, multiplex immunofluorescence is performed before H&E staining. mIF is acquired through sequential staining rounds, with a 40-min primary-antibody incubation followed by a 20-min secondary-antibody incubation in each round. After completing all fluorescence channels spanning the 480–780 nm detection range, the entire section is scanned by whole-slide fluorescence imaging.

The same slide is subsequently destained and cleaned, restained with standard hematoxylin and eosin, and scanned again in bright-field mode. The mIF and H&E WSIs are then registered to establish dense spatial correspondence between molecular measurements and morphology. mIF is acquired first because multiplex fluorescence staining is more sensitive to antigen preservation, whereas H&E restaining can be performed after fluorescence imaging. Using the same physical section also avoids the geometric ambiguity introduced by adjacent-section registration.

Internal multiplex panels. The internal paired collection contains five tumor-specific same-section panels: **Rectum19mIF**, **Kidney7mIF**, **Prostate9mIF**, **Liver9mIF**, and **Gastric15mIF**, corresponding to 19-, 7-, 9-, 9-, and 15-plex molecular measurements, respectively. These panels cover epithelial, stromal, vascular, immune, proliferation, and tumor-associated signals, providing molecular supervision for both anatomical constituents and their surrounding microenvironment.

Representative tumor-associated markers include P53, Ki-67, HepPar-1, Arginase-1, AFP, CK7, CK19, GPC3, and HBs in the liver panel; CKpan, P53, Ki-67, P504s, PSA, ERG, AR, P63, and S100 in the prostate panel; and CD138, MPO, D2-40, Vimentin, CD34, CKpan, P53, Ki-67, HER2, CEA, CDX2, MUC5AC, MUC6, Claudin18.2, and Syn in the gastric panel. Protein channels used for each experiment are fixed according to the corresponding acquisition panel before model training.

A.3 Data Registration

Although H&E and mIF images were acquired from the same physical tissue section, differences in slide orientation, image scale, scanning coordinates, and staining-induced local deformation can introduce spatial misalignment between the two modalities. We therefore developed a coarse-to-fine registration pipeline that uses the DAPI channel as the structural reference of the mIF image and aligns it to the corresponding H&E image in three stages: orientation normalization, thumbnail-level global registration, and patch-level local refinement. The complete procedure is summarized in Algorithm 1, with representative registration results shown in Fig. 7.

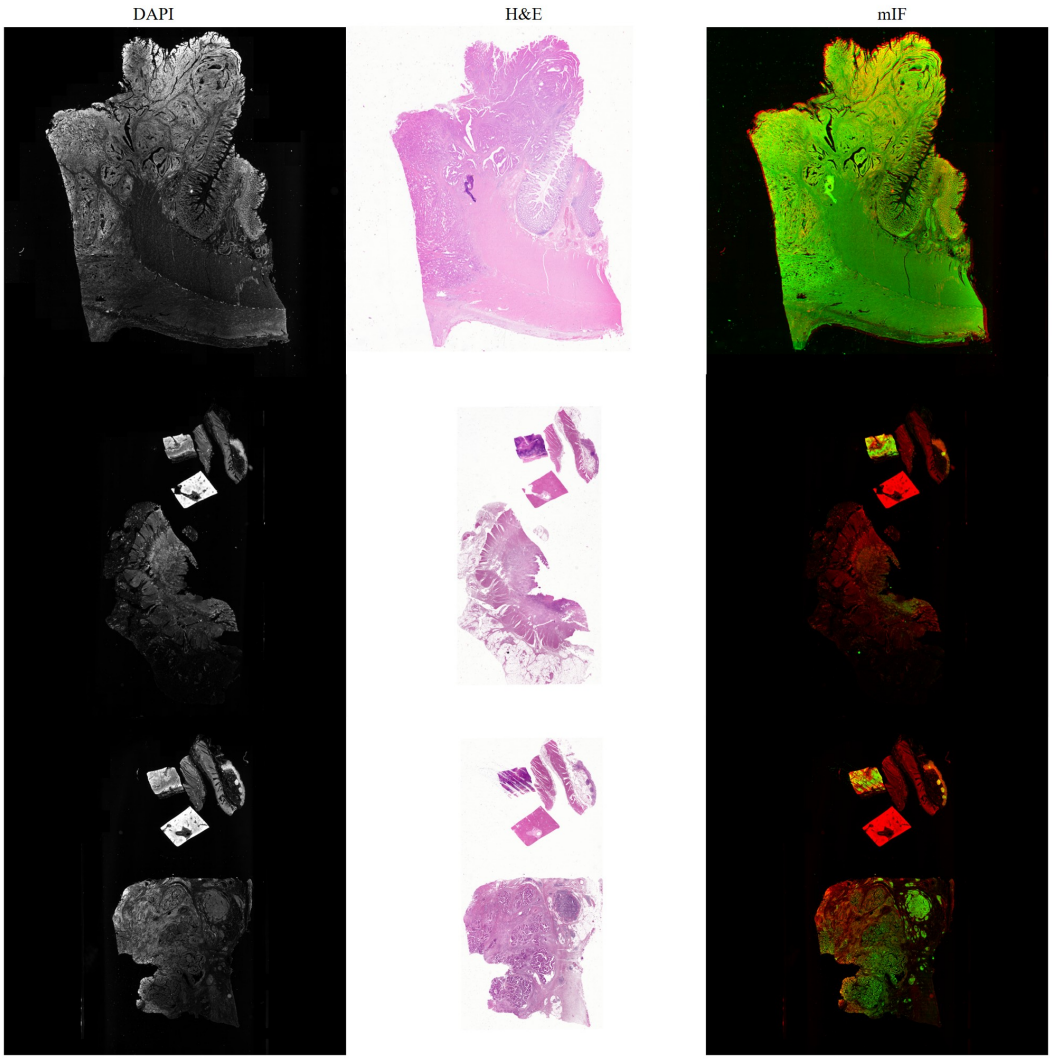


Figure 7: **Coarse-to-fine registration of paired H&E and mIF images.** DAPI serves as the structural reference of the mIF image. Orientation normalization and thumbnail-level global registration establish coarse correspondence, followed by hematoxylin–DAPI local refinement to correct residual patch-level misalignment. Representative overlays illustrate the spatial correspondence before and after registration.

Orientation normalization and global registration. When the tissue orientations of the two modalities differed, the H&E slide was first rotated to match the major orientation of the DAPI image while preserving the complete tissue field. H&E and DAPI images were then independently downsampled with their aspect ratios preserved such that the longest side did not exceed 1024 pixels. Rather than directly resizing the two modalities to an identical shape, we center-padded the thumbnails to a common canvas to avoid geometric distortion. DAPI was treated as the moving image and H&E as the fixed reference. After grayscale normalization and Gaussian smoothing ($\sigma = 1.5$), a 2D affine transformation was optimized using Mattes mutual information with 50 histogram bins and regular-step gradient descent. The affine registration was further refined at the thumbnail level using an 80-iteration Demons deformation to compensate for residual local distortions. The resulting transformation parameters, original image dimensions, thumbnail dimensions, and padding offsets were retained for subsequent coordinate mapping.

Algorithm 1 Coarse-to-Fine H&E–mIF Registration

Require: H&E image I_H, DAPI image I_D, mIF channels $\{I_M^c\}_{c=1}^C$, patch size $P = 512$, stride $S = 512$
Ensure: Spatially aligned patch pairs $\mathcal{D} = \{(H_i, \{M_i^c\}_{c=1}^C)\}$

Stage 1: Orientation normalization
Rotate I_H when necessary to match the major tissue orientation of I_D
Preserve the complete field of view and fill newly introduced background regions
Stage 2: Thumbnail-level global registration
Generate aspect-ratio-preserving thumbnails $\tilde{I}_H$ and $\tilde{I}_D$ with maximum side length 1024
Center-pad $\tilde{I}_H$ and $\tilde{I}_D$ to a common canvas
Convert both thumbnails to grayscale and apply Gaussian smoothing with $\sigma = 1.5$
Initialize a 2D affine transformation T_A using image-center alignment
Optimize T_A using Mattes mutual information with 50 histogram bins and regular-step gradient descent
Refine the registered thumbnail using an 80-iteration Demons transformation T_D
Store T_A, T_D, original image sizes, thumbnail sizes, and padding offsets
Stage 3: Full-resolution coordinate mapping
Initialize aligned patch set $\mathcal{D} \leftarrow \emptyset$
for each DAPI-grid location (x_i, y_i) with stride S **do**
  Define DAPI patch $D_i = I_D[x_i : x_i + P,\ y_i : y_i + P]$
  Map the four patch corners from the original DAPI space to the padded thumbnail space
  Apply T_A^{-1} and remove H&E padding offsets
  Rescale the transformed coordinates to the original H&E resolution
  Obtain the corresponding coarse H&E region H_i^{coarse}
  if H_i^{coarse} is background-dominated or D_i contains insufficient nuclear signal **then**
    continue
  end if
  Stage 4: Patch-level local refinement
  Expand the H&E search region by 100 pixels around the coarse location
  Extract the hematoxylin channel using HED decomposition and apply Otsu thresholding
  Convert D_i to grayscale and apply Otsu thresholding
  Initialize $s^* \leftarrow -\infty$ and $\Delta^* \leftarrow (0, 0)$
  for each candidate offset $\Delta = (\Delta x, \Delta y)$ within ± 100 pixels with step 5 **do**
    Crop the candidate H&E mask and resize it to $P \times P$
    Compute $s_\Delta = \dfrac{2|H_\Delta \cap D_i|}{|H_\Delta| + |D_i| + \epsilon}$
    if $s_\Delta > s^*$ **then**
      $s^* \leftarrow s_\Delta$, $\Delta^* \leftarrow \Delta$
    end if
  end for
  if $s^* < 0.1$ **then**
    continue
  end if
  Crop the final H&E patch H_i using Δ^*
  Extract all mIF channels $\{M_i^c\}_{c=1}^C$ at (x_i, y_i)
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(H_i, \{M_i^c\}_{c=1}^C)\}$
end for
return $\mathcal{D}$

Full-resolution coordinate mapping and local refinement. For full-resolution paired patch extraction, candidate mIF regions were generated on the DAPI coordinate system using 512×512 patches. Each patch was first projected to the H&E coordinate system by accounting for thumbnail downsampling and center-padding offsets and applying the inverse affine transformation to its four corners. The transformed coordinates were then mapped back to the original H&E resolution. The thumbnail-derived displacement field was not directly applied at this stage, thereby avoiding propagation of low-resolution deformation estimates to high-resolution image regions. Instead, residual misalignment was corrected independently for each patch.

Table 4: **Protein recovery on ORION-CRC.** PCC and Spearman measure expression consistency; AUROC and AP evaluate marker-positive region discrimination.

Channel	PCC ↑	Spearman ↑	AUROC ↑	AP ↑
Hoechst	0.96	0.88	–	–
AF1	–	–	–	–
CD31	0.83	0.78	0.89	0.76
CD45	0.82	0.76	0.87	0.73
CD68	0.83	0.78	0.93	0.82
Argo550	–	–	–	–
CD4	0.67	0.65	0.82	0.62
FOXP3	0.66	0.64	0.80	0.61
CD8a	0.74	0.69	0.90	0.81
CD45RO	0.76	0.75	0.92	0.82
CD20	0.77	0.78	0.92	0.82
PD-L1	0.68	0.67	0.84	0.65
CD3e	0.80	0.77	0.93	0.82
CD163	0.81	0.76	0.86	0.73
E-cadherin	0.93	0.86	0.98	0.90
PD-1	0.69	0.68	0.87	0.80
Ki67	0.72	0.69	0.88	0.80
Pan-CK	0.91	0.84	0.95	0.90
SMA	0.84	0.81	0.94	0.84
Average	**0.79**	**0.75**	**0.89**	**0.78**

Specifically, an enlarged H&E region of ± 100 pixels around the globally mapped position was extracted. Its hematoxylin component was obtained by HED color deconvolution, while the corresponding DAPI patch was converted to grayscale. Both signals were binarized using Otsu thresholding. Candidate H&E locations were searched with a 5-pixel step, resized to the DAPI patch resolution, and evaluated using $\mathrm{Dice}(H, D) = 2|H \cap D|/(|H| + |D| + \epsilon)$. The offset yielding the highest Dice score was selected as the final local alignment. Before refinement, regions dominated by blank or dark H&E background, low image variation, or insufficient DAPI signal were discarded. Patches with a final Dice score below 0.1 were also excluded. The resulting H&E patch and all mIF channels at the corresponding DAPI coordinates were retained as spatially paired observations for molecular grounding.

A.4 Baseline Methods

General cancer classification. We compare FRAME with representative methods from three major computational pathology families. GNN-based methods include Tea-Graph (Lee et al., 2022), PatchGCN (Chen et al., 2021), and TransGNN (Tang et al., 2024). MIL-based methods include ZoomMIL (Thandiackal et al., 2022), CLAM (Lu et al., 2021), TransMIL (Shao et al., 2021), and ABMIL (Ilse et al., 2018). Pathology foundation models include CTrans (Wang et al., 2022), HIPT (Chen et al., 2022), UNI (Chen et al., 2024), CHIEF (Wang et al., 2024), CONCH (Lu et al., 2024), and Virchow2 (Zimmermann et al., 2024).

Structural-event and fine-grained lesion recognition. For structural-event evaluation, CNN-based segmentation methods include FastFCN (Wu et al., 2019), U-Net++ (Zhou et al., 2020), DeepLabv3+ (Chen et al., 2018), and DenseASPP (Yang et al., 2018). Swin-Unet (Cao et al., 2022) represents Transformer-based segmentation, while SAM-based methods include SAMUS (Lin et al., 2023b), SAMed (Zhang & Liu, 2023), SAMPath (Zhang et al., 2023), UN-SAM (Chen et al., 2025), and BONUS (Lin et al., 2023c).

All competing methods use identical dataset partitions and supervision budgets whenever applicable. Official pretrained checkpoints and recommended preprocessing pipelines are used when available. No competing method receives additional task-specific supervision beyond that provided to FRAME under the corresponding experimental setting.

Table 5: **Quantitative comparison of constituent parsing performance.** Results are reported using MPA, Dice, mIoU, and FWIoU on four structural-event benchmarks. Best results are shown in bold.

(a) TLSD

Method	MPA	Dice	mIoU	FWIoU
CNN-based				
FastFCN	52.44	55.98	50.76	66.32
U-Net++	56.29	58.17	56.08	60.53
DeepLabv3+	60.72	64.69	51.36	78.15
DenseASPP	68.49	73.04	59.29	72.21
Transformer-based				
Swin-Unet	63.69	72.89	59.80	79.45
SAM-based				
SAMUS	68.41	67.79	53.18	76.97
SAMed	60.93	71.89	59.17	78.16
SAMPath	72.67	72.01	**68.99**	78.14
UN-SAM	68.77	73.40	60.18	80.48
BONUS	71.56	75.14	62.67	78.83
FRAME	**76.48**	**80.45**	66.98	**85.89**

(b) PIND

Method	MPA	Dice	mIoU	FWIoU
CNN-based				
FastFCN	49.85	47.56	43.67	47.61
U-Net++	62.14	58.09	50.72	52.36
DeepLabv3+	63.12	61.47	50.21	51.88
DenseASPP	70.63	65.41	56.18	58.62
Transformer-based				
Swin-Unet	68.94	70.75	55.12	57.09
SAM-based				
SAMUS	71.03	68.21	57.36	58.94
SAMed	66.85	70.96	56.47	60.05
SAMPath	74.92	72.63	61.35	64.28
UN-SAM	73.76	73.11	62.24	67.61
BONUS	75.98	75.19	63.71	62.73
FRAME	**80.99**	**82.45**	**78.42**	**77.41**

(c) CAMELYON16

Method	MPA	Dice	mIoU	FWIoU
CNN-based				
FastFCN	76.13	60.51	61.22	71.60
U-Net++	74.07	58.04	59.78	70.77
DeepLabv3+	74.99	59.44	60.83	71.62
DenseASPP	76.43	61.49	62.30	72.72
Transformer-based				
Swin-Unet	77.90	63.14	63.26	73.22
SAM-based				
SAMUS	74.47	58.50	60.01	70.89
SAMed	74.89	61.35	63.43	74.57
SAMPath	78.73	67.22	67.66	77.55
UN-SAM	79.48	67.61	67.68	77.33
BONUS	77.87	65.32	65.97	76.15
FRAME	**83.89**	**70.36**	**71.78**	**80.49**

(d) PAIP 2021

Method	MPA	Dice	mIoU	FWIoU
CNN-based				
FastFCN	75.42	61.63	60.35	72.48
U-Net++	74.86	57.12	60.54	69.93
DeepLabv3+	75.71	60.22	59.66	72.37
DenseASPP	77.25	60.87	63.14	71.68
Transformer-based				
Swin-Unet	77.08	64.36	62.41	74.09
SAM-based				
SAMUS	73.52	59.61	59.28	66.77
SAMed	75.63	60.54	64.62	75.34
SAMPath	77.69	66.18	68.72	78.41
UN-SAM	80.62	68.47	66.59	78.14
BONUS	76.94	66.53	67.21	75.08
FRAME	**85.36**	**75.44**	**77.08**	**85.49**

A.5 Quantitative Evaluation of Protein Recovery

We evaluate the protein recovery capability of FRAME using spatially registered H&E–mIF pairs from the ORION-CRC cohort. For each molecular channel, predicted protein maps are compared with experimentally measured fluorescence signals at corresponding spatial locations.

Two types of evaluation metrics are considered. PCC and Spearman correlation measure the agreement between predicted and measured continuous expression patterns, where PCC evaluates linear consistency and Spearman evaluates rank consistency. In addition, each protein channel is converted into a binary marker-positive map according to the fluorescence threshold determined from the corresponding mIF measurement. AUROC and AP are then used to evaluate the ability of FRAME to distinguish positive and negative molecular regions.

As summarized in Table 4, FRAME achieves consistently high agreement across diverse molecular markers. Structural and lineage-associated markers show particularly strong recovery, including E-cadherin, Pan-CK, SMA, CD31, and CD68. Immune-associated markers exhibit relatively lower but still reliable performance due to their sparse and heterogeneous spatial distributions. These results

Table 6: Full performance comparison on PANDA, CPTAC-CCRCC and TCGA-CRC. We report AUC for slide-level diagnosis and grading. **Bold** indicates the best performance and underline indicates the second best in each column. † marks the best result among non-PFM (Pathology Foundation Model) baselines, including GNN-based and MIL-based methods.

Architecture	Methods	Pretraining Samples	PANDA		CPTAC-CCRCC		TCGA-CRC	
			Diagnosis	Grading	Diagnosis	Grading	Diagnosis	Grading
GNN-based	Tea-Graph	–	0.868†	0.836†	0.747†	0.723†	0.860†	0.853†
	Patch-GCN	–	0.824	0.818	0.726	0.689	0.823	0.818
	TransGNN	–	0.848	0.817	0.612	0.588	0.844	0.838
MIL-based	CLAM	–	0.836	0.816	0.579	0.440	0.844	0.838
	ABMIL	–	0.847	0.794	0.604	0.578	0.829	0.806
	TransMIL	–	0.847	0.834	0.706	0.681	0.829	0.824
	ZoomMIL	–	0.860	0.833	0.646	0.622	0.838	0.831
PFM-based	HIPT	10K+	0.858	0.825	0.733	0.697	0.858	0.849
	CTrans	35K+	0.930	0.927	0.772	0.742	0.845	0.838
	UNI	100,000+	0.948	0.944	0.824	0.779	0.870	0.866
	CHIEF	60K+	0.953	0.941	0.839	0.801	0.896	0.878
	CONCH	117,000+	0.949	0.934	0.842	0.813	0.873	0.862
Ours	FRAME	–	**0.981**	**0.973**	**0.867**	**0.858**	**0.937**	**0.929**
	Δ vs PFM	✓	+2.8%	+2.9%	+2.5%	+4.5%	+4.1%	+5.1%
	Δ vs non-PFM	✗	+11.3%	+13.7%	+12.0%	+13.5%	+7.7%	+7.6%

demonstrate that Morphology Encoding can recover biologically meaningful spatial molecular cues from routine H&E morphology.

A.6 Detailed Constituent Parsing Evaluation

To comprehensively evaluate constituent parsing, we compare FRAME with representative segmentation methods from CNN, Transformer, and SAM-based families. Four metrics are reported: MPA, Dice, mIoU, and FWIoU. Dice and mIoU are emphasized as primary overlap-based metrics because structural-event recognition requires accurate localization of both large tissue regions and small anatomical constituents.

FRAME consistently outperforms all competing methods across the four benchmarks. On TLSD, FRAME achieves the highest Dice (80.45%) and FWIoU (85.89%), indicating improved recovery of both target structures and overall tissue organization. On PIND, FRAME obtains substantial gains in Dice and mIoU, suggesting that the coupled target–constituent interaction is particularly effective for complex structural relationships. On CAMELYON16, FRAME improves fine-grained lesion localization despite the small size and heterogeneous appearance of metastatic regions. Finally, on PAIP, FRAME achieves 85.36% MPA and 75.44% Dice, demonstrating robust target parsing across organ-specific pathological contexts.

Compared with conventional segmentation architectures, FRAME benefits from protein-grounded initialization and coupled target–constituent reasoning, which allow local morphological evidence to be interpreted in the context of complete anatomical structures rather than isolated pixels.

A.7 Detailed Ablation and Annotation-efficiency Analysis

We provide detailed ablation experiments to investigate the contribution of each component and design choice in FRAME. All experiments are conducted under the same training and evaluation settings as the main experiments.

Table 7: **Ablation of FRAME components and relation representations.**

Variant	PAIP AUC	TLSD AUC	Isomerism Acc. (%)
Component contribution			
H&E only	0.907	0.915	68.6
+ Protein grounding	0.919	0.929	74.3
+ Constituent parsing	0.935	0.936	82.4
FRAME	**0.958**	**0.969**	**95.7**
Relation representation			
Composition only	0.921	0.934	74.6
Center-based geometry	0.938	0.946	83.2
Boundary-centric geometry	**0.958**	**0.969**	**95.7**

A.7.1 Component Contribution and Relation Representation

Component contribution. We progressively introduce the three major components of FRAME, including molecular grounding, constituent parsing, and boundary-centric reasoning. Starting from an H&E-only model, protein grounding improves PAIP AUC from 0.907 to 0.919 and histological isomerism accuracy from 68.6% to 74.3%. Introducing constituent parsing further improves isomerism accuracy to 82.4%, demonstrating the benefit of explicitly organizing local morphology into anatomical targets and surrounding constituents. Finally, the complete FRAME achieves 0.958 PAIP AUC, 0.969 TLSD AUC, and 95.7% isomerism accuracy, confirming the necessity of integrating all three components.

Relation representation. We compare three strategies for representing diagnostic relations: composition-only features, center-based geometry, and boundary-centric geometry. Composition-only modeling achieves 74.6% accuracy on histological isomerism cases, while center-based geometry improves this value to 83.2%. However, boundary-centric reasoning achieves 95.7% accuracy, together with the highest PAIP and TLSD AUCs. These results demonstrate that anatomical boundaries provide a more informative reference frame than constituent presence or global spatial proximity alone.

Table 8: Ablation study of coupled constituent parsing strategy.

Variant	Dice	Target Dice	Event AUC
Independent	76.0	71.6	0.931
Spatial $\rightarrow$ Target	78.3	74.6	0.945
Target $\rightarrow$ Spatial	77.8	73.7	0.943
Coupled Parsing	**80.5**	**77.4**	**0.958**

Table 9: Ablation study of protein grounding design.

Variant	PCC	AUROC	Event AUC
H&E only	–	–	0.901
Protein auxiliary	0.74	0.85	0.928
Protein-guided fusion	**0.79**	**0.89**	**0.958**

A.7.2 Coupled Parsing and Protein Grounding

Parsing strategy. We compare independent parsing with one-way and bidirectional interactions between constituent and target representations. Independent parsing achieves 76.0% Dice and 0.931 Event AUC. Spatial-to-target and target-to-spatial interactions provide moderate improvements, while the proposed coupled parsing achieves the best performance with 80.5% Dice, 77.4% Target Dice, and 0.958 Event AUC. This confirms that target localization and constituent interpretation benefit from mutual refinement.

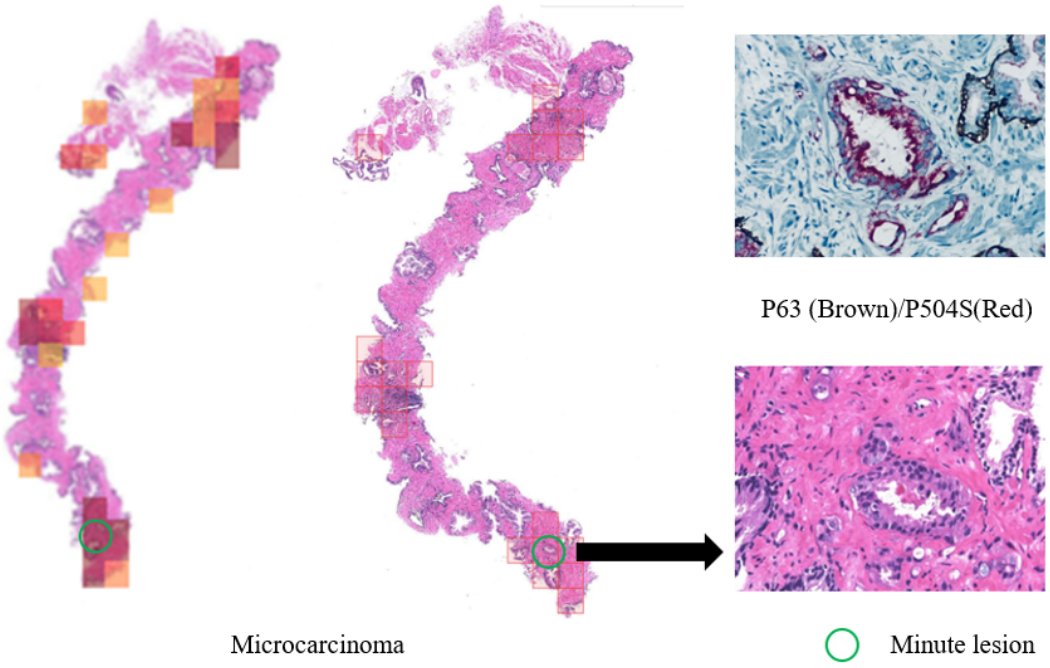


Figure 8: **Representative physician-missed cases identified by FRAME during institutional deployment.** Examples show clinically challenging lesions that were missed during routine pathological review but detected by FRAME, illustrating its potential to assist pathologists in recognizing subtle structural abnormalities and reducing diagnostic oversight in real-world practice.

Protein grounding design. We further evaluate how molecular supervision should be incorporated. Compared with H&E-only learning, protein prediction as an auxiliary objective improves downstream diagnosis from 0.901 to 0.928 AUC. Protein-guided fusion provides additional improvement, achieving 0.958 AUC, demonstrating that molecular information is more effective when integrated as an intermediate biological representation rather than used only as an auxiliary optimization constraint.

A.7.3 Annotation Efficiency Analysis

Annotation efficiency. We evaluate the dependence of FRAME on expert annotations by varying the number of physician-corrected ROIs. As summarized in Table 10, FRAME consistently achieves higher performance under all annotation budgets. With only 8 corrected ROIs, FRAME reaches 0.86 AUC, compared with 0.72 for the H&E-only baseline and 0.78 for FRAME without mIF supervision. Increasing the annotation budget further improves performance, but FRAME reaches 0.95 AUC with only 32 corrected ROIs, demonstrating that molecular grounding and structural reasoning substantially reduce the need for extensive expert annotation.

Table 10: Annotation efficiency evaluation with different numbers of physician-corrected ROIs.

ROI annotations	H&E only	FRAME w/o mIF	FRAME
8	0.72	0.78	0.86
16	0.78	0.84	0.91
32	0.83	0.89	0.95
64	0.87	0.93	0.96
128	0.90	0.95	0.97