FFM-CP: FEW-SHOT FUSION OF FOUNDATION MODELS FOR COMPUTATIONAL PATHOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Pathology vision-language foundation models vary in performance across diseases and tasks, with no single model consistently performing best. The high cost of expert pathology annotation can also limit the labeled data available for task-specific adaptation. Combining complementary pretrained representations is a potential approach to these limitations, yet learning an effective fusion from few labeled examples remains challenging. We introduce `Few-shot Fusion Foundation Models of Computational Pathology` (FFM-CP), which is a framework that combines multiple pathology vision-language models in the few-shot learning setting. The framework first aligns heterogeneous representations using a closed-form Orthogonal Procrustes transformation estimated from corresponding support images. This alignment preserves within-model feature geometry without training an additional alignment network. Within the aligned space, a unified graph enables information exchange across backbones by jointly refining support-image features as well as visual and textual class prototypes. These refined representations support complementary text-prototype and case-retrieval branches that capture semantic class knowledge and within-class visual variation, respectively. Each branch learns to combine predictions from all ordered backbone pairs, allowing queries encoded by one model to draw on evidence represented by another. We evaluate three backbone combinations on six histopathology datasets at 4, 8, and 16 shots per class. FFM-CP achieves higher mean macro-F1 than the strongest individually adapted member of each fused set in 50 of 54 comparisons. These findings suggest that combining complementary pretrained representations can improve histopathological classification when annotations are limited.

1 INTRODUCTION

Histopathology images are central to cancer diagnosis, grading, and subtyping, but developing computational models for these tasks traditionally requires large collections of images annotated by expert pathologists (Campanella et al., 2019; Wahab et al., 2022). Such annotations are expensive, and they are particularly difficult to obtain for rare entities and institution-specific diagnostic problems. Pathology vision-language foundation models (VLMs) offer a promising alternative. By pretraining visual and text encoders on large image-text collections, these models learn reusable representations that can support zero-shot prediction and adaptation from only a small labeled support set (Ikezogwo et al., 2023; Zhang et al., 2025; Lu et al., 2024; Xiang et al., 2025; Zhou et al., 2026).

The growing number of pathology VLMs, however, creates a new model-selection problem. Their pretraining images, text sources, objectives, architectures, and embedding dimensions differ substantially. Consequently, each model captures a different subset of diagnostically useful morphology and terminology. Recent benchmarks confirm that model rankings vary across tissues and tasks, with no single foundation model consistently dominating the others (Bareja et al., 2026). Selecting one backbone in advance therefore commits the downstream classifier to that model's representational strengths and blind spots while potentially discarding complementary evidence available in other models.

Existing studies address few-shot adaptation and multi-model fusion for computational pathology as separate problems. Few-shot methods tune prompts or lightweight adapters for a single pathology VLM (He et al., 2026; Nguyen et al., 2025; Shi et al., 2024; Han et al., 2025; Guo et al., 2025),

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.

while the other methods combine independent pretrained vision and text encoders (Shalam & Korman; Jiang et al., 2026). Conversely, multi-foundation-model approaches fuse visual features, slide representations, or prediction logits (Dang et al., 2024; Luo et al., 2025; Yang et al., 2025). These methods are generally designed for downstream cohorts containing hundreds or thousands of labeled samples, or they rely on a separate large-scale distillation stage. This leaves an important setting unresolved: how can several intact pathology VLMs be combined when the only downstream datasets consist of K labeled tiles per class?

This setting presents three connected challenges. First, independent pathology-pretrained VLMs produce heterogeneous feature spaces. Their dimensions may differ, and even embeddings of the same dimension need not be directly comparable. Meaningful cross-model interaction therefore requires alignment estimated from a small support set. Second, the models should exchange class-specific evidence during adaptation. Aggregating predictions from independently adapted models does not allow support-image features or class prototypes to benefit from evidence represented by other backbones. Joint refinement would enable visual examples and textual class knowledge from different models to inform one another before prediction. Third, fusion should exploit both semantic class information and the visual variation among individual support examples. Moreover, the usefulness of these complementary sources can vary across backbones and tasks. Together, these challenges motivate a framework that aligns, jointly refines, and combines representations from multiple pathology VLMs in the few-shot learning setting.

We introduce FFM-CP, a support-set-driven framework for cross-backbone fusion of pathology VLMs. FFM-CP first uses a closed-form Orthogonal Procrustes transformation (Schönemann, 1966; Lample et al., 2018), estimated from corresponding support images, to map all model representations into a shared reference space while preserving their internal geometry. It then constructs a unified heterogeneous graph containing support images, visual class prototypes, and text prototypes from every model. Seven relation types propagate visual and semantic evidence within and across backbones. Finally, cross-backbone pair fusion evaluates all M^2 ordered query-evidence combinations in two complementary branches: (1) a text-prototype branch transfers semantic class information, (2) a case-retrieval branch compares each query with its most relevant labeled examples and preserves intra-class variation. The foundation-model encoders remain frozen, and all task-specific components are learned using the few-shot support set alone. Our contributions are threefold:

- We formulate support-set-only fusion of heterogeneous pathology VLMs. To our knowledge, FFM-CP presents the first framework that adapts multiple pathology vision-language backbones in the few-shot learning setting.
- We combine parameter-free cross-backbone alignment, a unified multi-relation knowledge graph, and dual-branch fusion over all ordered model pairs, thereby exploiting complementary visual, textual, and case-based evidence.
- We evaluate five pathology VLMs and three fused model sets on six public datasets spanning four organs under 4-, 8-, and 16-shot protocols. Across the extensive experiments, FFM-CP outperforms the best individual model within the corresponding fusion set in 50 cases, demonstrating the benefits of cross-backbone fusion.

2 Related Work

Pathology vision-language foundation models. Contrastive image-text pretraining (Radford et al., 2021) has been transferred to histopathology by a growing family of foundation models, whose training corpora range from medical Twitter threads (Huang et al., 2023), narration transcribed from educational videos (Ikezogwo et al., 2023) to figure-caption pairs from the open-access literature (Zhang et al., 2025), in-house clinical collections (Lu et al., 2024; Xiang et al., 2025), and released corpora restructured under a disease ontology (Zhou et al., 2026). Benchmarks agree that no member of this family dominates: across 31 models and 41 tasks, the top of the ranking changes with the tissue (Bareja et al., 2026).

Few-shot adaptation of vision-language models for histopathology images. Few-shot adaptation is particularly important in histopathology because assembling labeled whole-slide image datasets requires substantial clinical resources and expert effort. For frozen VLMs, CoOp (Zhou et al., 2022) learns prompt context vectors, while CLIP-Adapter (Gao et al., 2024) adapts fea-

tures through a lightweight bottleneck. TIP-ADAPTER (Zhang et al., 2021) uses a training-free support-set cache, and GRAPH ADAPTER (Li et al., 2023) refines textual representations through a dual knowledge graph.

For WSI analysis, VILA-MIL (Shi et al., 2024) uses LLM-generated prompts at two magnification scales with 16 labeled slides per class. MSCPT (Han et al., 2025) learns multi-scale, context-aware prompts, while FOCUS (Guo et al., 2025) uses language prompts to select diagnostically relevant patches. MGPATH (Nguyen et al., 2025) couples the PROV-GIGAPATH (Xu et al., 2024) vision encoder with the PLIP (Huang et al., 2023) text encoder through multi-granular prompting. PATHPT (He et al., 2026) applies prompt tuning to rare cancer subtyping with 1–10 WSIs per subtype. Related to our alignment stage, FSF (Shalam & Korman) aligns independently pretrained visual and textual encoders using a closed-form orthogonal Procrustes transformation and regularizes adaptation with a flow-matching prior. These methods adapt a single backbone or a predefined vision-text pair. FFM-CP jointly refines visual and textual representations across M frozen VLMs, enabling information exchange during adaptation, and combines predictions from all M^2 ordered backbone pairs.

Combining multiple foundation models in pathology. MFMF (Dang et al., 2024) combines features from multiple patch encoders within a multiple-instance learning network for whole-slide classification. ELF (Luo et al., 2025) integrates five tile-level foundation models through slide-encoder pretraining on 53,699 WSIs from 11 datasets. It combines contrastive alignment with weakly supervised objectives and uses fixed slide representations for downstream linear probing. FuseCPath (Yang et al., 2025) combines patch- and slide-level foundation models across scales. At the prediction level, LogitProd (Huang et al., 2026) learns sample-adaptive weights over frozen experts' logits without aligning their feature spaces. Its training procedure first fits task-specific expert heads, then trains a fusion gate on a separate labeled subset. Meanwhile, FM^2 (Yu et al., 2025) distills general and pathology foundation models into a single student by disentangling their consensus and divergence, followed by zero- and few-shot evaluation.

Our work focuses on learning task-specific fusion directly from a few-shot support set. FFM-CP retains each model's frozen visual and textual encoders and learns cross-model representation refinement and prediction fusion from the $C \cdot K$ labeled support tiles of a C-class, K-shot task. This approach requires no separate large-scale fusion pretraining or distillation stage.

3 METHOD

3.1 PRELIMINARIES

Vision-language foundation models for pathology. Contrastive image-text pretraining has been extended to histopathology (Huang et al., 2023; Ikezogwo et al., 2023; Lu et al., 2024; Zhou et al., 2026; Chen et al., 2024; Xiang et al., 2025; Zhang et al., 2025). These models typically comprise a visual encoder E_v and a text encoder E_t, which map H&E tiles and descriptions to representations z_v and z_t, respectively. Pretraining encourages high cosine similarity between paired image and text representations. In our framework, both encoders remain frozen during downstream adaptation.

Despite this shared structure, their pretraining data, objectives, and architectures differ. Training corpora include medical social media posts (Huang et al., 2023), transcribed educational videos (Ikezogwo et al., 2023), article figure-caption pairs (Zhang et al., 2025), and non-public clinical cohorts (Lu et al., 2024; Xiang et al., 2025). Most models directly align paired tiles and captions. MUSK (Xiang et al., 2025) first applies unified masked modeling to unpaired data comprising 50 million tiles and 10^9 pathology-text tokens, then aligns the modalities. Text encoders range from generic CLIP-style Transformers to biomedical models informed by disease ontology (Zhou et al., 2026). Visual architectures and embedding dimensions also vary. These differences may contribute to task-dependent variation in model performance.

Complementary strengths of pathology foundation models. A recent benchmark of 32 foundation models across 41 computational pathology tasks shows that no single model dominates consistently; instead, performance rankings vary substantially across tissues and tasks (Bareja et al., 2026). Different models perform best on brain and breast tumor classification, bladder and prostate

subtyping, and lung-related tasks, suggesting that their learned representations capture different diagnostically useful signals.

This variation is also consistent with the diversity of morphological cues used across cancer types. Breast carcinoma grading considers tubule formation, nuclear pleomorphism, and mitotic activity (Elston & Ellis, 1991), whereas Gleason/ISUP grading of prostate cancer emphasizes glandular architecture (Epstein et al., 2016). Lung adenocarcinoma subtyping, in contrast, depends largely on the predominant tissue growth pattern (Travis et al., 2011). These differences do not fully explain why model rankings change within a given tissue, but they highlight that pathology tasks rely on distinct visual structures and scales. Together, this empirical and domain-level evidence motivates combining multiple pathology foundation models to exploit their complementary strengths.

Few-shot multi-model setting. We are given M frozen pathology FMs indexed by $m \in \{1, \dots, M\}$, each with its own encoders E_v^m, E_t^m, its own temperature τ_m, and its own embedding dimension d_m. These dimensions can differ across the M models. The support set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ contains K labeled images from each of C classes, giving $N = C \cdot K$ images in total. Here, $\mathbf{x}_i$ denotes the ith support image, and $y_i \in \{1, \dots, C\}$ is its class label. We use $c \in \{1, \dots, C\}$ to index the classes and t_c to denote the text description of class c. For model m, the frozen encoders extract the visual and text features

$$z_{v,i}^m = E_v^m(\mathbf{x}_i) \in \mathbb{R}^{d_m}, \qquad z_{t,c}^m = E_t^m(t_c) \in \mathbb{R}^{d_m}, \qquad \bar{z}_{v,c}^m = \frac{1}{K} \sum_{i:\, y_i = c} z_{v,i}^m, \tag{1}$$

where $z_{v,i}^m$ is the vision features of the ith-sample, $\bar{z}_{v,c}^m$ is the class-level visual prototype of class c, and $z_{t,c}^m$ is the textual features of the class description t_c. We collected the support features, visual prototypes, and text features as row matrices: We collect the feature vectors as rows:

$$Z_v^m = \left[(z_{v,i}^m)^\top\right]_{i=1}^N, \quad \bar{Z}_v^m = \left[(\bar{z}_{v,c}^m)^\top\right]_{c=1}^C, \quad Z_t^m = \left[(z_{t,c}^m)^\top\right]_{c=1}^C. \tag{2}$$

Here, $Z_v^m \in \mathbb{R}^{N \times d_m}$ and $\bar{Z}_v^m, Z_t^m \in \mathbb{R}^{C \times d_m}$. A query tile $\mathbf{x}$ is encoded as $z_v^m(\mathbf{x}) = E_v^m(\mathbf{x})$. A cross-model similarity such as $\mathrm{sim}(z_{v,i}^m, z_{t,c}^n)$ is dimensionally invalid when $d_m \neq d_n$ and is not geometrically meaningful. Even when their dimensions match, independently pretrained models may capture different visual semantic features. The same image or class description can therefore be represented differently across models, so direct cross-model similarities may not reliably reflect semantically corresponding features. This incompatibility is the first challenge addressed by our method.

3.2 The Proposed Method

FFM-CP constructs a single heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with three node types: *(i)* support-image nodes carrying $z_{v,i}^m$, *(ii)* text-prototype nodes carrying $z_{t,c}^m$, and *(iii)* visual-prototype nodes carrying $\bar{z}_{v,c}^m$. The graph therefore contains $M \cdot N$ support-image nodes, $M \cdot C$ text-prototype nodes, and $M \cdot C$ visual-prototype nodes. Its edges may join nodes derived from different models, allowing a class description encoded by one model to be refined using visual evidence extracted by another. The full pipeline is illustrated in Figure 1.

Cross-backbone alignment. As discussed in Section 3.1, pathology FMs may differ in embedding dimension and coordinate system. Their features therefore require alignment before computing cross-backbone similarities in the unified graph. FFM-CP uses Orthogonal Procrustes (OP) to map the features of each model into a shared reference space. The semi-orthogonality constraint preserves norms and inner products, retaining the feature relationships learned during pretraining. OP admits a closed-form solution based on features extracted from the same support tiles. This enables alignment using only the few-shot support set, without additional alignment data or iterative optimization.

Let r denote a reference model with the largest embedding dimension; thus, d_r is the dimension of model r, and let $d = d_r$. For each model $m \neq r$, we obtain a semi-orthogonal map into the reference space by solving an Orthogonal Procrustes problem (Schönemann, 1966):

$$\mathbf{Q}_m = \underset{\mathbf{Q} \in \mathbb{R}^{d \times d_m},\, \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_{d_m}}{\arg\min} \left\| Z_v^r - Z_v^m \mathbf{Q}^\top \right\|_F^2 = \mathbf{U}_m \mathbf{V}_m^\top, \qquad {Z_v^r}^\top Z_v^m = \mathbf{U}_m \mathbf{\Sigma}_m \mathbf{V}_m^\top, \tag{3}$$

where $\mathbf{U}_m \mathbf{\Sigma}_m \mathbf{V}_m^\top$ denotes the thin singular value decomposition (SVD) of the cross-covariance. Here, $\mathbf{U}_m \in \mathbb{R}^{d \times d_m}$ and $\mathbf{\Sigma}_m, \mathbf{V}_m \in \mathbb{R}^{d_m \times d_m}$. We set $\mathbf{Q}_r = \mathbf{I}_d$. For any feature z from model m, its aligned representation is $\tilde{z} = \mathbf{Q}_m z$. Applying this map to the support features and visual and textual class prototypes yields $\tilde{Z}_v^m$, $\tilde{\tilde{Z}}_v^m$, and $\tilde{Z}_t^m$. The map estimated from support images also applies to textual features because both encoders share an embedding space within each model.

The condition $\mathbf{Q}_m^\top \mathbf{Q}_m = \mathbf{I}_{d_m}$ preserves feature norms and within-model cosine similarities. All aligned representations lie in $\mathbb{R}^d$, with cross-model alignment estimated from corresponding support images.

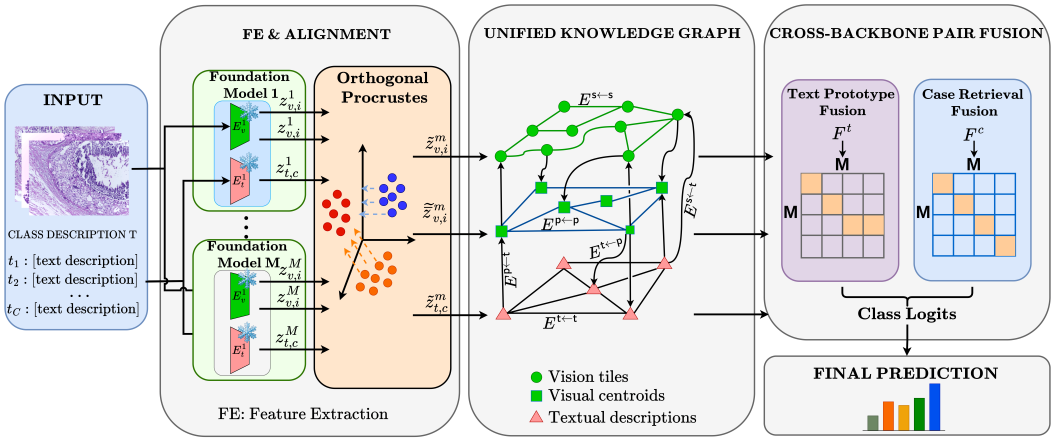

Figure 1: **Overview of FFM-CP**. Frozen pathology foundation models encode support tiles, class descriptions, and the query tile. Cross-backbone alignment projects their heterogeneous representations into a shared reference space, where a unified graph jointly refines support-image, visual-prototype, and text-prototype nodes through seven relation types. Finally, cross-backbone pair fusion aggregates all ordered query–evidence backbone pairs through separate text and case streams.

Unified knowledge graph. Our graph construction builds on GRAPHADAPTER (Li et al., 2023), which refines textual prompt features using task-specific structural knowledge encoded through inter-class relationships. We extend this idea by constructing a unified graph that jointly refines support-image features and visual and textual class prototypes across the M pathology VLMs.

This structural modeling is particularly relevant to histopathology, where diagnostic categories are inherently related rather than independent. Tumor classifications often follow hierarchical structures (Tan et al., 2020), while grading systems are ordinal, such that neighboring grades differ primarily in degree (Elston & Ellis, 1991). Consequently, adjacent or morphologically similar categories can share visual characteristics and are often difficult to distinguish in practice (Brancati et al., 2022).

To explicitly capture these relationships, we construct a unified graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that connects representations both within and across the M backbones. Its node set contains three disjoint types, $\mathcal{V} = \mathcal{V}^{\mathsf{t}} \cup \mathcal{V}^{\mathsf{p}} \cup \mathcal{V}^{\mathsf{s}}$, corresponding to text prototypes, visual prototypes, and support images, respectively. We stack their aligned features model by model:

$$\mathbf{X}^{\mathsf{t}} = \left[\tilde{Z}_t^1; \ldots; \tilde{Z}_t^M\right], \quad \mathbf{X}^{\mathsf{p}} = \left[\tilde{\tilde{Z}}_v^1; \ldots; \tilde{\tilde{Z}}_v^M\right], \quad \mathbf{X}^{\mathsf{s}} = \left[\tilde{Z}_v^1; \ldots; \tilde{Z}_v^M\right]. \tag{4}$$

Here, $\mathbf{X}^{\mathsf{t}}, \mathbf{X}^{\mathsf{p}} \in \mathbb{R}^{MC \times d}$ and $\mathbf{X}^{\mathsf{s}} \in \mathbb{R}^{MN \times d}$. Within each model, image-text similarity provides class probabilities. For a tile $\mathbf{x}$ with representation $z_v^m = E_v^m(\mathbf{x})$, model m computes

$$P_m(y = c \mid \mathbf{x}) = \frac{\exp\left(\tau_m \, \mathrm{sim}(z_v^m, z_{t,c}^m)\right)}{\sum_{c'=1}^{C} \exp\left(\tau_m \, \mathrm{sim}(z_v^m, z_{t,c'}^m)\right)}. \tag{5}$$

The OP maps align the models using corresponding support images, allowing similarity comparisons across backbones. We use these aligned representations to construct the graph edges.

Let $a, b \in \{\mathsf{t}, \mathsf{p}, \mathsf{s}\}$ denote node types. The edge set $\mathcal{E}^{a\leftarrow b} \subseteq \mathcal{E}$ contains edges from nodes of type b to nodes of type a. The graph contains six relations derived from feature similarities:

$$\underbrace{\mathcal{E}^{\mathsf{s}\leftarrow\mathsf{t}},\ \mathcal{E}^{\mathsf{s}\leftarrow\mathsf{s}}}_{\text{into support images}} \qquad \underbrace{\mathcal{E}^{\mathsf{p}\leftarrow\mathsf{t}},\ \mathcal{E}^{\mathsf{p}\leftarrow\mathsf{p}}}_{\text{into visual prototypes}} \qquad \underbrace{\mathcal{E}^{\mathsf{t}\leftarrow\mathsf{p}},\ \mathcal{E}^{\mathsf{t}\leftarrow\mathsf{t}}}_{\text{into text prototypes}}.$$

For each node type a, let $n_a = |\mathcal{V}^a|$ and let $\mathbf{X}_i^a$ denote the ith row of $\mathbf{X}^a$. Its mean feature is $\boldsymbol{\mu}^a = \frac{1}{n_a}\sum_{i=1}^{n_a} \mathbf{X}_i^a$. The weighted adjacency matrix $\mathbf{A}^{a\leftarrow b} \in \mathbb{R}^{n_a \times n_b}$ has receiving nodes in its rows and sending nodes in its columns. For the six feature-derived relations, we compute

$$\left[\mathbf{A}^{a\leftarrow b}\right]_{ij} = \mathrm{sim}\big(\mathbf{X}_i^a - \boldsymbol{\mu}^a,\ \mathbf{X}_j^b - \boldsymbol{\mu}^b\big). \tag{6}$$

For same-type relations ($a = b$), we set the diagonal entries to zero to remove self-loops.

The seventh relation, $\mathcal{E}^{\mathsf{p}\leftarrow\mathsf{s}}$, uses the support labels. Each visual prototype links to the K support-image nodes of the same class and backbone. Its binary adjacency matrix is

$$\left[\mathbf{A}^{\mathsf{p}\leftarrow\mathsf{s}}\right]_{(m,c),(n,i)} = \mathbf{1}[m = n]\,\mathbf{1}[y_i = c]. \tag{7}$$

Here, (m, c) indexes the visual prototype of class c from model m, and (n, i) indexes support image i represented by model n.

Graph learning across backbones. We sequentially refine support-image features, visual prototypes, and text prototypes. Each stage uses the latest available source features, and a superscript $*$ denotes refined features:

$$\begin{aligned}
\mathbf{X}^{\mathsf{s},*} &= \mathbf{X}^{\mathsf{s}} + w_1\, g_{\mathsf{s}}\big(\mathbf{X}^{\mathsf{s}}, \mathbf{X}^{\mathsf{t}}, \mathbf{A}^{\mathsf{s}\leftarrow\mathsf{t}}\big) + w_2\, g_{\mathsf{s}}\big(\mathbf{X}^{\mathsf{s}}, \mathbf{X}^{\mathsf{s}}, \mathbf{A}^{\mathsf{s}\leftarrow\mathsf{s}}\big),\\
\mathbf{X}^{\mathsf{p},*} &= \mathbf{X}^{\mathsf{p}} + w_3\, g_{\mathsf{p}}\big(\mathbf{X}^{\mathsf{p}}, \mathbf{X}^{\mathsf{s},*}, \mathbf{A}^{\mathsf{p}\leftarrow\mathsf{s}}\big) + w_4\, g_{\mathsf{p}}\big(\mathbf{X}^{\mathsf{p}}, \mathbf{X}^{\mathsf{t}}, \mathbf{A}^{\mathsf{p}\leftarrow\mathsf{t}}\big) + w_5\, g_{\mathsf{p}}\big(\mathbf{X}^{\mathsf{p}}, \mathbf{X}^{\mathsf{p}}, \mathbf{A}^{\mathsf{p}\leftarrow\mathsf{p}}\big),\\
\mathbf{X}^{\mathsf{t},*} &= \mathbf{X}^{\mathsf{t}} + w_6\, g_{\mathsf{t}}\big(\mathbf{X}^{\mathsf{t}}, \mathbf{X}^{\mathsf{p},*}, \mathbf{A}^{\mathsf{t}\leftarrow\mathsf{p},*}\big) + w_7\, g_{\mathsf{t}}\big(\mathbf{X}^{\mathsf{t}}, \mathbf{X}^{\mathsf{t}}, \mathbf{A}^{\mathsf{t}\leftarrow\mathsf{t}}\big).
\end{aligned} \tag{8}$$

Here, $\{w_i\}_{i=1}^7$ are learned relation scalars, and g_{s}, g_{p}, and g_{t} are destination-specific message-passing operators. Support-image nodes first receive messages from text prototypes and other support-image nodes. Visual prototypes then incorporate the refined support features, and text prototypes incorporate the refined visual prototypes. The feature-derived relations exchange information within and across backbones.

The adjacency $\mathbf{A}^{\mathsf{t}\leftarrow\mathsf{p},*}$ is recomputed using Equation 6 with $\mathbf{X}^{\mathsf{t}}$ and $\mathbf{X}^{\mathsf{p},*}$, including the updated mean of the visual prototypes. The remaining feature-derived adjacency matrices are computed from the initial aligned features and kept fixed. The membership adjacency $\mathbf{A}^{\mathsf{p}\leftarrow\mathsf{s}}$ is fixed by the support labels.

For destination and source feature matrices $\mathbf{X} \in \mathbb{R}^{n_X \times d}$ and $\mathbf{Y} \in \mathbb{R}^{n_Y \times d}$, respectively, let $\mathbf{A} \in \mathbb{R}^{n_X \times n_Y}$ be their weighted adjacency. For destination type $a \in \{\mathsf{s}, \mathsf{p}, \mathsf{t}\}$, we define

$$g_a(\mathbf{X}, \mathbf{Y}, \mathbf{A}) = \sigma\Big(\mathbf{D}_X^{-1}\mathbf{X}\mathbf{W}_a^{\mathrm{self}} + \widehat{\mathbf{A}}\mathbf{Y}\mathbf{W}_a^{\mathrm{msg}} + \mathbf{b}_a\Big). \tag{9}$$

The matrices $\mathbf{W}_a^{\mathrm{self}}, \mathbf{W}_a^{\mathrm{msg}} \in \mathbb{R}^{d\times d}$ and bias $\mathbf{b}_a \in \mathbb{R}^d$ are learned separately for each destination type. They are shared across relations with the same destination type. The bias is added to each row, and σ denotes an element-wise activation.

The normalized adjacency is $\widehat{\mathbf{A}} = \mathbf{D}_X^{-1/2}\mathbf{A}\mathbf{D}_Y^{-1/2}$. The diagonal degree matrices are defined for each supplied adjacency by

$$\mathbf{D}_X = \mathrm{diag}\left(1 + \sum_{j=1}^{n_Y} |A_{ij}|\right), \qquad \mathbf{D}_Y = \mathrm{diag}\left(1 + \sum_{i=1}^{n_X} |A_{ij}|\right).$$

Absolute edge weights accommodate signed similarities, while the added one ensures positive degrees.

Training objective. For support examples $(\mathbf{x}_i, y_i)$, we optimize $\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{fuse}} + \eta\mathcal{L}_{\text{diag}}$ where

$$\mathcal{L}_{\text{fuse}} = \frac{1}{N}\sum_{i=1}^{N}\Big[\text{CE}\big(\boldsymbol{\ell}^{\text{t}}(\mathbf{x}_i), y_i\big) + \lambda\,\text{CE}\big(\boldsymbol{\ell}^{\text{c}}(\mathbf{x}_i), y_i\big)\Big],$$
$$\mathcal{L}_{\text{diag}} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{m=1}^{M}\Big[\text{CE}\big(\boldsymbol{Z}^{\text{t}}_{m,m}(\mathbf{x}_i), y_i\big) + \lambda\,\text{CE}\big(\boldsymbol{Z}^{\text{c}}_{m,m}(\mathbf{x}_i), y_i\big)\Big]. \tag{10}$$

Here, CE denotes cross-entropy loss. The coefficient λ weights the case-stream loss in both terms. The coefficient η controls the contribution of the auxiliary diagonal loss. The diagonal term preserves each same-model classifier. We optimize the parameters of the message-passing operators g_s, g_p, and g_t, the relation weights $\{w_j\}_{j=1}^{7}$, and the pair-fusion matrices $\mathbf{F}^{\text{t}}$ and $\mathbf{F}^{\text{c}}$. All visual and textual encoders and Procrustes maps remain fixed. At inference, we apply softmax to each stream's fused logits and average the resulting class probabilities:

$$P(y = c \mid \mathbf{x}) = \frac{1}{2}\Big[\text{softmax}\big(\boldsymbol{\ell}^{\text{t}}(\mathbf{x})\big)_c + \text{softmax}\big(\boldsymbol{\ell}^{\text{c}}(\mathbf{x})\big)_c\Big]. \tag{11}$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluated the proposed method on six public histopathology datasets: **LungHist700** (Diosdado et al., 2024) ($C = 7$ classes), **HeidelbergSkin** (Kriegsmann et al., 2022) ($C = 16$ classes), **Kather2016** (Kather et al., 2016) ($C = 8$ classes), **LubLung** (Rączkowska et al., 2022) ($C = 9$ classes), **BRACS** (Brancati et al., 2022) ($C = 7$ classes), **BACH** (Aresta et al., 2019) ($C = 4$ classes). The datasets span four organs, including lung, skin, colorectum, and breast. Moreover, they cover tissue-type recognition (Kather2016, LubLung, HeidelbergSkin) and diagnostic subtyping (LungHist700, BRACS, BACH), with sizes ranging from 400 images to 129,364 tiles and class counts from 4 to 16. Detailed descriptions are included in Appendix A.

Every dataset was divided into disjoint training, validation, and test proportions. Given a shot count K, we draw K examples per class from the training partition, and $\min(K, 4)$ examples per class from the validation partition, following standard practice for few-shot adaptation of vision-language models (Li et al., 2023). For BRACS, we used the officially released test set as the test partition without further modification. Each experiment was repeated five times, and we provided the mean and standard deviation of these runs to ensure robustness and reproducibility.

Implementation details. In this study, we use the released pretrained weights of all foundation models and keep their encoders frozen throughout training. For each category, we average the text embeddings of $[P]$ prompt templates constructed from its diagnostic description t_c (Appendix B). The reference model r is the member with the largest embedding dimension: KEEP (Zhou et al., 2026) ($d = 768$) for KQP and KPB, and MUSK (Xiang et al., 2025) ($d = 2048$) for MQP. The Procrustes maps in Equation 3 are computed once from the $C \cdot K$ support tiles. The loss weights are set to $\lambda = 0.5$ and $\eta = 0.5$ arcoss the datasets.

Baseline methods. Following the criterion in Sec. 3.1, we evaluate each fused set against its strongest member. Each candidate model: KEEP (Zhou et al., 2026), MUSK (Xiang et al., 2025), QuiltNet (Ikezogwo et al., 2023), PLIP (Huang et al., 2023), and BioMedCLIP (Zhang et al., 2025), is adapted independently using the proposed pipeline with $M = 1$. For each dataset and shot count, the fusion gain Δ is measured relative to the best-performing member of the corresponding fused set. We additionally evaluate a late-fusion ensemble in which each model compares query-tile features only with its own text prototypes and support-tile features. This ensemble is reported as variant (a) of the ablation study in Sec. 4.3.

4.2 Results

Tables 1, 2, and 3 report macro-F1 across six histopathology datasets at 16, 8, and 4 shots, respectively. Results are presented as the mean and standard deviation over five random seeds. We evaluate three fused sets: KEEP+QuiltNet+PLIP (KQP), KEEP+PLIP+BiomedCLIP (KPB), and MUSK+QuiltNet+PLIP (MQP). All gains are expressed in percentage points relative to the strongest individual member of the corresponding fused set on each dataset.

Table 1: Macro F1 (%) at $K = 16$ shots on all six histopathology datasets. Mean $\pm$ standard deviation over 5 random seeds. **Bold** = best, underlined = second best in each column. Δ indicates the gain over the strongest individual member of that same fused set, selected per column. Accuracy and AUC are in Appendix C.

Method	LungHist700	HeidelbergSkin	Kather2016	LubLung	BRACS	BACH	Avg.
Vision–language foundation models (GCN adapter)							
KEEP	78.68 $\pm$2.14	90.64 $\pm$0.99	87.28 $\pm$1.10	81.44 $\pm$1.28	52.11 $\pm$2.74	89.32 $\pm$1.37	79.91
MUSK	70.75 $\pm$2.68	83.27 $\pm$0.40	87.66 $\pm$0.21	75.73 $\pm$2.58	50.55 $\pm$3.78	84.00 $\pm$1.12	75.33
QuiltNet	66.76 $\pm$2.90	75.50 $\pm$0.28	86.64 $\pm$0.51	74.64 $\pm$1.16	45.78 $\pm$3.12	72.93 $\pm$4.48	70.37
PLIP	67.19 $\pm$2.26	75.51 $\pm$0.61	86.34 $\pm$0.84	74.53 $\pm$2.85	43.03 $\pm$2.54	72.60 $\pm$3.71	69.87
BiomedCLIP	61.72 $\pm$1.05	69.66 $\pm$1.36	81.40 $\pm$2.10	68.96 $\pm$2.62	49.42 $\pm$2.60	68.49 $\pm$4.30	66.61
Fused VLMs (ours: OP alignment + combined graph + GCN)							
KEEP+QuiltNet+PLIP	84.11 $\pm$1.63	**92.68** $\pm$0.32	**91.66** $\pm$1.56	**86.86** $\pm$0.99	52.77 $\pm$1.96	92.18 $\pm$2.20	83.37
Δ	+5.43	+2.04	+4.38	+5.42	+0.66	+2.86	+3.47
KEEP+PLIP+BiomedCLIP	**84.25** $\pm$1.38	92.58 $\pm$0.65	91.63 $\pm$1.52	86.42 $\pm$0.88	**54.53** $\pm$3.04	**92.38** $\pm$1.86	**83.63**
Δ	+5.57	+1.94	+4.35	+4.98	+2.42	+3.06	+3.72
MUSK+QuiltNet+PLIP	78.67 $\pm$2.10	87.25 $\pm$0.70	91.19 $\pm$0.76	81.73 $\pm$1.00	52.00 $\pm$2.68	85.19 $\pm$0.74	79.34
Δ	+7.92	+3.98	+3.53	+6.00	+1.45	+1.19	+4.01

Table 2: Macro F1 (%) at $K = 8$ shots on all six histopathology datasets. Mean $\pm$ standard deviation over 5 random seeds. **Bold** = best, underlined = second best in each column. Accuracy and AUC are in Appendix C.

Method	LungHist700	HeidelbergSkin	Kather2016	LubLung	BRACS	BACH	Avg.
Vision-language foundation models (GCN adapter)							
KEEP	75.62 $\pm$6.87	87.90 $\pm$0.61	86.83 $\pm$1.08	77.53 $\pm$1.20	50.96 $\pm$2.34	89.35 $\pm$3.14	78.03
MUSK	67.60 $\pm$3.87	79.86 $\pm$3.38	85.17 $\pm$2.03	72.00 $\pm$2.19	44.99 $\pm$0.49	81.47 $\pm$4.08	71.85
QuiltNet	57.93 $\pm$9.33	70.89 $\pm$1.12	83.56 $\pm$1.41	71.76 $\pm$0.55	42.31 $\pm$2.24	67.82 $\pm$4.94	65.71
PLIP	60.28 $\pm$4.97	71.52 $\pm$1.32	82.66 $\pm$1.74	70.84 $\pm$1.19	40.61 $\pm$1.85	66.91 $\pm$9.57	65.47
BiomedCLIP	56.28 $\pm$3.78	64.68 $\pm$1.38	77.77 $\pm$2.36	63.30 $\pm$0.84	43.44 $\pm$2.33	70.52 $\pm$1.70	62.66
Fused VLMs (ours: OP alignment + combined graph + GCN)							
KEEP+QuiltNet+PLIP	**77.40** $\pm$2.95	90.74 $\pm$1.12	**90.87** $\pm$1.11	**83.48** $\pm$0.83	**53.23** $\pm$3.31	89.53 $\pm$2.29	**80.88**
Δ	+1.78	+2.84	+4.04	+5.95	+2.27	+0.18	+2.84
KEEP+PLIP+BiomedCLIP	76.98 $\pm$2.67	**91.48** $\pm$1.23	90.53 $\pm$1.02	82.36 $\pm$1.30	52.30 $\pm$2.19	**89.78** $\pm$2.77	80.57
Δ	+1.36	+3.58	+3.7	+4.83	+1.34	+0.43	+2.54
MUSK+QuiltNet+PLIP	72.21 $\pm$2.87	84.25 $\pm$0.34	88.82 $\pm$0.97	78.59 $\pm$1.14	48.47 $\pm$1.16	82.60 $\pm$4.35	75.82
Δ	+4.61	+4.39	+3.65	+6.59	+3.48	+1.13	+3.98

Fusion improves performance consistently. Across six datasets, three shot settings, and three fused sets, FFM-CP outperforms the strongest constituent in 50 of 54 comparisons, including all cases at 8 and 16 shots. Average gains increase from 2.35 points at 4 shots to 3.73 points at 16 shots, suggesting that cross-backbone fusion remains effective across few-shot regimes and benefits from a more informative support set.

Fusion helps even when constituent models differ substantially in strength. FFM-CP can improve upon the strongest backbone even when the remaining models perform considerably worse on their own. For example, on LungHist700 at 16 shots, fusing KEEP, PLIP, and BiomedCLIP improves macro-F1 from 78.68% for KEEP alone to 84.25%. This suggests that lower-performing models can still provide complementary evidence that strengthens the fused prediction.

4.3 Ablation Studies

We examine the contributions of cross-backbone prediction pairs and feature alignment using $K = 16$ support examples per class and $M = 3$ backbones: KEEP, QuiltNet, and PLIP. Table 4 compares the full model with three variants. The diagonal variant restricts both prediction streams to same-backbone pairs ($m = n$) while retaining the full graph. The other two variants replace Procrustes alignment with zero-padding or a two-layer MLP for each model. The MLPs are trained jointly with the graph.

Table 3: Macro F1 (%) at $K = 4$ shots on all six histopathology datasets. Mean ± standard deviation over 5 random seeds. **Bold** = best, underlined = second best in each column. Accuracy and AUC are in Appendix C.

Method	LungHist700	HeidelbergSkin	Kather2016	LubLung	BRACS	BACH	Avg.
Vision–language foundation models (GCN adapter)							
KEEP	65.33 ±3.02	86.26 ±0.83	76.40 ±1.66	72.10 ±4.99	48.64 ±4.91	**86.26** ±9.94	72.50
MUSK	60.52 ±4.39	74.49 ±2.75	79.44 ±3.30	61.51 ±3.18	38.67 ±2.09	75.66 ±6.89	65.05
QuiltNet	53.15 ±1.89	63.89 ±1.27	75.42 ±3.93	62.44 ±1.68	34.83 ±2.90	52.17 ±6.43	56.98
PLIP	50.94 ±4.08	64.76 ±0.57	76.34 ±2.79	61.50 ±1.50	32.35 ±2.63	55.72 ±7.00	56.94
BiomedCLIP	52.62 ±7.02	58.00 ±3.83	70.72 ±5.14	55.96 ±0.75	37.65 ±2.81	62.04 ±8.48	56.16
Fused VLMs (ours: OP alignment + combined graph + GCN)							
KEEP+QuiltNet+PLIP	**66.76** ±6.63	88.30 ±1.71	82.47 ±3.62	**76.08** ±4.62	48.33 ±2.32	84.61 ±3.57	74.43
Δ	+1.43	+2.04	+6.07	+3.98	−0.31	−1.65	+1.93
KEEP+PLIP+BiomedCLIP	66.19 ±6.87	**88.38** ±2.23	82.69 ±3.48	76.02 ±3.67	**49.53** ±2.69	85.94 ±2.78	**74.79**
Δ	+0.86	+2.12	+6.29	+3.92	+0.89	−0.32	+2.29
MUSK+QuiltNet+PLIP	62.08 ±4.37	77.87 ±1.24	**82.88** ±4.36	68.98 ±2.28	42.22 ±3.48	74.23 ±5.80	68.04
Δ	+1.56	+3.38	+3.44	+6.54	+3.55	−1.43	+2.84

Table 4: Ablation of FFM-CP at $K = 16$ shots using KEEP, QuiltNet, and PLIP. (a) Restricts prediction fusion to diagonal pairs ($m = n$) in both streams while retaining the full graph. (b) Replaces the Procrustes map in Equation 3 with zero-padding to $\mathbb{R}^d$. (c) Replaces Procrustes with a two-layer MLP per model, trained jointly with the graph. Results are macro-F1 (%) scores.

Configure	LungHist700	HeidelbergSkin	Kather2016	LubLung	BRACS	BACH
FFM-CP(FULL)	**84.11±1.63**	92.68±0.32	**91.66±1.56**	**86.86±0.99**	52.77±1.96	**92.18±2.20**
(a) FFM-CP(DIAGONAL)	82.24±1.76	**92.71±0.31**	90.61±1.84	86.77±1.01	**52.79**±2.15	91.58±1.93
(b) ZERO-PADDING	80.94±1.16	91.63±0.20	90.09±1.14	85.22±1.01	51.79±2.59	90.22±2.35
(c) MLP	79.73±2.53	90.60±0.79	90.34±0.95	85.30±0.58	51.03±0.85	90.54±1.82

Contribution of cross-backbone prediction pairs. Including cross-backbone query–evidence pairs yields a modest average gain of 0.59 macro-F1 points, with improvements on four of six datasets. The effect is therefore smaller and more task-dependent than that of representation alignment, suggesting that cross-model prediction transfer is helpful when backbones provide complementary evidence, but is not uniformly beneficial across datasets.

Contribution of Procrustes alignment. Procrustes alignment consistently outperforms zero-padding and learned MLP alignment, improving mean macro-F1 by 1.73 and 2.12 points on average, respectively. While zero-padding only matches feature dimensions, Procrustes additionally aligns the backbones' coordinate systems while preserving their internal geometry. This suggests that, in the few-shot regime, preserving pretrained structure while establishing cross-model correspondence is more effective than learning a flexible alignment from limited supervision.

5 Conclusion

We introduced FFM-CP to study few-shot fusion of heterogeneous pathology foundation models. Our results show that even weaker individual backbones can contribute complementary diagnostic evidence when their representations are aligned and jointly refined. Ablations further suggest that preserving pretrained geometry through simple Procrustes alignment is more effective than learning a flexible alignment under limited supervision. Overall, heterogeneity across pathology foundation models can be an asset: combining complementary frozen representations can yield stronger few-shot predictions than relying on any single model.

Limitations. Our evaluation is limited to six public classification datasets and a fixed set of pathology VLMs; broader validation on additional backbones, whole-slide tasks, and cross-institution cohorts remains future work. Considering all M^2 ordered backbone pairs also increases fusion cost as the model set grows. Moreover, Procrustes assumes a global linear orthogonal relation between feature spaces, which may not capture more complex cross-model discrepancies. Finally, because alignment and graph construction rely on the few-shot support set, performance may depend on how representative those examples are, especially in extremely low-shot settings.

ETHICS STATEMENT

This study conducts a secondary analysis of six publicly available histopathology datasets containing images derived from human tissue. No new participants were recruited, no additional tissue was collected, and no clinical interventions were performed. We used the released images and labels without accessing personally identifiable patient information. The datasets and pretrained models are credited through their original publications. Potential risks include biases inherited from the source datasets and pretrained models, as well as diagnostic errors when they are applied outside the evaluated settings. The proposed method is intended for methodological research, and clinical use would require independent validation and assessment of performance across institutions and patient groups.

REPRODUCIBILITY STATEMENT

We support reproducibility through the methodological description in Section 3 and the experimental protocol in Section 4.1. These sections specify the alignment, graph construction, prediction branches, training objective, few-shot sampling procedure, and single-model baselines. Appendix A describes the six public datasets, and Appendix B provides the prompt templates. Experiments use the released pretrained model weights with all encoders frozen. Results are reported as means and standard deviations over five random seeds in Sub-section 4.2 and Appendix C. Sub-section 4.3 evaluates the contribution of individual framework components.

AI USE STATEMENT

The authors used ChatGPT (OpenAI) during manuscript preparation to assist with drafting and revising selected passages, language editing, and improving the clarity and organization of methodological descriptions. The authors reviewed and revised the AI-assisted material and checked its consistency with the study methods and experimental results. The authors take full responsibility for the accuracy, originality, and scientific integrity of the final manuscript.

REFERENCES

Guilherme Aresta, Teresa Araújo, Scotty Kwok, et al. BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019. doi: 10.1016/j.media.2019.05.010. (Cited on pages 7, 14, and 15.)

Rohan Bareja, Francisco Carrillo-Perez, Yuanning Zheng, Marija Pizurica, Tarak Nath Nandi, Lu Tian, Jeanne Shen, Ravi Madduri, and Olivier Gevaert. A benchmark study of vision and pathology foundation models for computational pathology. *Nature Communications*, 2026. (Cited on pages 1, 2, and 3.)

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. BRACS: A dataset for BReAst carcinoma subtyping in H&E histology images. *Database*, 2022, 2022. doi: 10.1093/database/baac093. (Cited on pages 5, 7, 14, and 15.)

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*, 25(8):1301–1309, 2019. (Cited on page 1.)

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

Thao M Dang, Yuzhi Guo, Hehuan Ma, Qifeng Zhou, Saiyang Na, Jean Gao, and Junzhou Huang. Mfmf: multiple foundation model fusion networks for whole slide image classification. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–8, 2024. (Cited on pages 2 and 3.)

Jorge Diosdado, Pere Gilabert, Santi Seguí, and Henar Borrego. Lunghist700: A dataset of histological images for deep learning in pulmonary pathology. *Scientific Data*, 11(1):1088, 2024. (Cited on pages 7, 14, and 15.)

Christopher W Elston and Ian O Ellis. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991. (Cited on pages 4 and 5.)

Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, Peter A Humphrey, Grading Committee, et al. The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. *The American journal of surgical pathology*, 40(2):244–252, 2016. (Cited on page 4.)

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024. (Cited on page 2.)

Zhengrui Guo, Conghao Xiong, Jiabo Ma, Qichen Sun, Lishuang Feng, Jinzhuo Wang, and Hao Chen. Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15590–15600. IEEE, 2025. (Cited on pages 1 and 3.)

Minghao Han, Linhao Qu, Dingkang Yang, Xukun Zhang, Xiaoying Wang, and Lihua Zhang. Mscpt: Few-shot whole slide image classification with multi-scale and context-focused prompt tuning. *IEEE Transactions on Medical Imaging*, 44(9):3756–3769, 2025. (Cited on pages 1 and 3.)

Dexuan He, Xiao Zhou, Wenbin Guan, Liyuan Zhang, Xiaoman Zhang, Sinuo Xu, Ge Wang, Lifeng Wang, Xiaojun Yuan, Jing Ma, et al. Boosting pathology foundation models via few-shot prompt-tuning for rare cancer subtyping. *Nature Communications*, 2026. (Cited on pages 1 and 3.)

Gexin Huang, Anqi Li, Yusheng Tan, Beidi Zhao, Gang Wang, Zu-Hua Gao, and Xiaoxiao Li. Plug-and-play logit fusion for heterogeneous pathology foundation models. *arXiv preprint arXiv:2604.07779*, 2026. (Cited on page 3.)

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):2307–2316, 2023. (Cited on pages 2, 3, and 7.)

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. *Advances in neural information processing systems*, 36:37995–38017, 2023. (Cited on pages 1, 2, 3, and 7.)

Ziqi Jiang, Yanghao Wang, and Long Chen. Exploring cross-modal flows for few-shot learning. In *International Conference on Learning Representations*, volume 2026, pp. 17074–17090, 2026. (Cited on page 2.)

Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6(1):27988, 2016. (Cited on pages 7, 14, and 15.)

Katharina Kriegsmann, Frithjof Lobers, Christiane Zgorzelski, Joerg Kriegsmann, Charlotte Janssen, Rolf Rüdinger Meliß, Thomas Muley, Ulrich Sack, Georg Steinbuss, and Mark Kriegsmann. Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology*, 12:1022967, 2022. (Cited on pages 7, 14, and 15.)

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International conference on learning representations*, 2018. (Cited on page 2.)

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36:13448–13466, 2023. (Cited on pages 3, 5, and 7.)

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024. (Cited on pages 1, 2, and 3.)

Xiangde Luo, Xiyue Wang, Feyisope Eweje, Xiaoming Zhang, Sen Yang, Ryan Quinton, Jinxi Xiang, Yuchen Li, Yuanfeng Ji, Zhe Li, et al. Ensemble learning of foundation models for precision oncology. *arXiv preprint arXiv:2508.16085*, 2025. (Cited on pages 2 and 3.)

Anh-Tien Nguyen, Duy Minh Ho Nguyen, Nghiem Tuong Diep, Trung Quoc Nguyen, Nhat Ho, Jacqueline Michelle Metsch, Miriam Cindy Maurer, Daniel Sonntag, Hanibal Bohnenberger, and Anne-Christin Hauschild. Mgpath: A vision-language model with multi-granular prompt learning for few-shot whole slide pathology classification. *Transactions on Machine Learning Research*, 2025. (Cited on pages 1 and 3.)

Alicja Rączkowska, Iwona Paśnik, Michał Kukiełka, Marcin Nicoś, Magdalena A Budzinska, Tomasz Kucharczyk, Justyna Szumiło, Paweł Krawczyk, Nicola Crosetto, and Ewa Szczurek. Deep learning-based tumor microenvironment segmentation is predictive of tumor mutations and patient survival in non-small-cell lung cancer. *BMC cancer*, 22(1):1001, 2022. (Cited on pages 7, 14, and 15.)

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021. (Cited on page 2.)

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. (Cited on pages 2 and 4.)

Daniel Shalam and Simon Korman. Flow-based alignment of uni-modal vision and text encoders for few-shot image classification. (Cited on pages 2 and 3.)

Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. Vila-mil: Dual-scale vision-language multiple instance learning for whole slide image classification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11248–11258. IEEE, 2024. (Cited on pages 1 and 3.)

PH Tan, I Ellis, K Allison, E Brogi, SB Fox, S Lakhani, AJ Lazar, and EA Morris. The world health organization classification of tumours of the breast. *Histopathology*, 77(2):181–5, 2020. (Cited on page 5.)

William D Travis, Elisabeth Brambilla, Masayuki Noguchi, Andrew G Nicholson, Kim R Geisinger, Yasushi Yatabe, David G Beer, Charles A Powell, Gregory J Riely, Paul E Van Schil, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *Journal of thoracic oncology*, 6(2):244–285, 2011. (Cited on page 4.)

Noorul Wahab, Islam M Miligy, Katherine Dodd, Harvir Sahota, Michael Toss, Wenqi Lu, Mostafa Jahanifar, Mohsin Bilal, Simon Graham, Young Park, et al. Semantic annotation for computational pathology: multidisciplinary experience and best practice recommendations. *The Journal of Pathology: Clinical Research*, 8(2):116–128, 2022. (Cited on page 1.)

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025. (Cited on pages 1, 2, 3, and 7.)

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. (Cited on page 3.)

Zhidong Yang, Xiuhui Shi, Wei Ba, Zhigang Song, Haijing Luan, Taiyuan Hu, Senlin Lin, Jiguang Wang, Shaohua Kevin Zhou, and Rui Yan. Fusion of multi-scale heterogeneous pathology foundation models for whole slide image analysis. *arXiv preprint arXiv:2510.27237*, 2025. (Cited on pages 2 and 3.)

Ziqi Yu, Shengjie Zhang, Nidan Qiao, Yao Zhao, Lequan Yu, Tingying Peng, and Xiao-Yong Zhang. Fm2: Fusing multiple foundation models for pathology image analysis via disentangled consensus-divergence representation. *Information Fusion*, pp. 103840, 2025. (Cited on page 3.)

R Zhang, R Fang, W Zhang, P Gao, K Li, J Dai, Y Qiao, and H Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. arxiv. *arXiv preprint arXiv:2111.03930*, 2021. (Cited on page 3.)

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm Ai*, 2(1):AIoa2400640, 2025. (Cited on pages 1, 2, 3, and 7.)

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16795–16804. IEEE, 2022. (Cited on page 2.)

Xiao Zhou, Luoyi Sun, Dexuan He, Wenbin Guan, Ge Wang, Ruifen Wang, Lifeng Wang, Xiaojun Yuan, Xin Sun, Ya Zhang, et al. Knowledge-enhanced pretraining for vision-language pathology foundation model on cancer diagnosis. *Cancer Cell*, 44(4):777–791, 2026. (Cited on pages 1, 2, 3, and 7.)

APPENDIX

A DATASETS

LungHist700. (Diosdado et al., 2024) contains 691 high-resolution 1200×1600 lung images from 45 patients at Hospital Clínico Universitario de Valladolid in 2023 during routine diagnostics, captured at both $20\times$ and $40\times$ magnification. The tissues are categorized into three superclasses, namely adenocarcinoma (`aca`), squamous cell carcinoma (`scc`), and normal lung (`nor`). Each carcinoma is further graded as well, moderately or poorly differentiated.

HeidelbergSkin. (Kriegsmann et al., 2022). The dataset comprises 129,364 tiles of $\approx 395 \times 395$ px annotated in QuPath from 386 cases. The dataset was drawn from the archives of the Institute of Pathology, Heidelberg University, the MVZ for Histology, Cytology and Molecular Diagnostics Trier, and the Institute for Dermatopathology Hannover. Diagnoses followed the WHO Classification of Skin Tumors. The patches were annotated for 16 categories: chondral tissue, dermis, elastosis, epidermis, hair follicle, skeletal muscle, necrosis, nerves, sebaceous glands, subcutis, eccrine glands, vessels, BCC, SqCC, naevi and melanoma.

Kather2016. (Kather et al., 2016). The dataset is a colorectal cancer texture dataset collected from the pathology archive of the Institute of Pathology, University Medical Center Mannheim, Heidelberg UniversitEight tissue types were selected, namely tumor epithelium, simple stroma, complex stroma, immune cell conglomerates, debris, normal mucosal glands, adipose tissue, and background, with 625 tiles each, giving 5,000 images of 150×150 px.

LubLung. (Rączkowska et al., 2022). The dataset comprises 23,199 tiles of 172×172 px derived from the formalin-fixed, paraffin-embedded (FFPE) tissue samples from 55 primary tumors of lung cancer at Medical University of Lublin, Poland. The tiles were annotated by an expert pathologist with nine labels: tumor, stroma, mixed, immune, vessel, bronchi, necrosis, lung, and background.

BRACS. (Brancati et al., 2022) provides 4,539 regions of interest extracted from 547 whole-slide images from 189 patients between 2019 and 2020 at the Department of Pathology at the National Cancer Institute - Scientific Institute for Research, Hospitalization and Healthcare (IRCCS) 'Fondazione G. Pascale'. Each WSI and its ROIs were annotated by the consensus of three board-certified pathologists into three lesion types: benign, atypical, and malignant. Then, the tiles are subtyped into seven categories: Normal, Pathological Benign, Usual Ductal Hyperplasia, Flat Epithelial Atypia, Atypical Ductal Hyperplasia, Ductal Carcinoma In Situ, and Invasive Carcinoma. Evaluation uses the released test partition of 570 ROIs unchanged. In this study, few-shot sampling draws the K-shot support from the 3,657 training ROIs and a validation subset from the 312 validation ROIs. We used the released test set proportion for evaluating the performance of the training process.

BACH. (Aresta et al., 2019) is the ICIAR 2018 breast histology challenge set: 400 microscopy images of 2048×1536 px at 0.42 µm/px, balanced across normal, benign, *in situ* carcinoma and invasive carcinoma at 100 images each. Its four balanced classes make it the coarsest task we evaluate.

The six public datasets used in this study cover a diverse range of organs, tissue types, image resolutions, and dataset sizes, providing a comprehensive benchmark for evaluating the generalizability of the proposed method. Table 5 summarizes the key characteristics of these datasets, including their organ of origin, number of classes, and total number of samples.

Table 5: The description of the six datasets used in the study.

Dataset	Organ	Classes	Samples
LungHist700 (Diosdado et al., 2024)	Lung	7	691
HeidelbergSkin (Kriegsmann et al., 2022)	Skin	16	129,364
Kather2016 (Kather et al., 2016)	Colorectum	8	5,000
LubLung (Rączkowska et al., 2022)	Lung	9	23,199
BRACS (Brancati et al., 2022)	Breast	7	4,539
BACH (Aresta et al., 2019)	Breast	4	400

B Additional Implementation Details

In this paper, we used the prompts listed in Table 6 for training and testing.

Table 6: The number of classes and the prompt templates used for each dataset. The five prompt embeddings of a class are averaged into its textual feature $z_{t,c}^{m}$

Datasets	# Classes	Prompt Template
LungHist700 (Diosdado et al., 2024)	7	"an H&E stained lung histopathology image showing [class]."†
HeidelbergSkin (Kriegsmann et al., 2022)	15	"an H&E stained skin histopathology image showing [class]."†
Kather2016 (Kather et al., 2016)	8	"an H&E stained colorectal histopathology image showing [class]."†
LubLung (Rączkowska et al., 2022)	9	"an H&E stained lung histopathology image showing [class]."†
BRACS (Brancati et al., 2022)	7	"an H&E stained breast histopathology image showing [class]."†
BACH (Aresta et al., 2019)	4	"an H&E stained breast histopathology image showing [class]."†

†Each dataset uses an ensemble of five templates: the one listed and "a [organ] histopathology image showing [class].", "a microscopic image of [organ] tissue showing [class].", "a pathology slide of [organ] showing [class]." and "[class] in [organ] histopathology, H&E stain.", with [organ] the organ word of the listed template (lung, skin, colorectal or breast).

C Vision-Language Backbones and Their Fusions

Tables 7 and 8 complement macro-F1 results with accuracy and AUC at $K = 16$ on `LungHist700`, `HeidelbergSkin`, `Kather2016`, `LubLung`, `BRACS`, and `BACH` in Table 1. For each dataset and metric, Δ is calculated relative to the strongest constituent of the corresponding fused set.

`LungHist700`. KPB achieves the highest mean macro-F1 (84.25%) and AUC (98.15%), improving upon its strongest constituent by 5.57 and 1.35 percentage points, respectively. KQP yields similar scores (84.11% and 98.09%), and both configurations achieve 85.76% accuracy. MQP produces the largest gains over its own constituents: 7.61, 7.92, and 2.31 percentage points in accuracy, macro-F1, and AUC, respectively.

`HeidelbergSkin`. On `HeidelbergSkin`, the strongest constituent AUC is already 99.09-99.48%, leaving limited room for absolute improvement. KQP achieves the highest mean accuracy (93.28%), macro-F1 (92.68%), and AUC (99.79%), with gains of 2.26, 2.04, and 0.31 percentage points over KEEP. KPB follows closely, differing by at most 0.10 percentage points across the three metrics. MQP improves macro-F1 from MUSK's 83.27% to 87.25%, although this remains below KEEP alone (90.64%). Across the three fused sets, AUC increases by 0.30-0.40 percentage points from already high constituent scores of 99.09-99.48%. These small absolute AUC gains accompany bigger improvements in accuracy (2.22-3.29 percentage points) and macro-F1 (1.94-3.98 percentage points).

`Kather2016`. All three fused sets exceed 91% in both accuracy and macro-F1, with gains of 3.44-4.37 and 3.53-4.38 percentage points, respectively. KQP achieves the highest mean accuracy (91.65%), macro-F1 (91.66%), and AUC (99.27%). The fused configurations are closely grouped, with differences of at most 0.51, 0.47, and 0.10 percentage points across the three metrics. AUC improves by 0.41-0.69 percentage points over the strongest constituent for each fused set. Although

MUSK is the strongest individual backbone on this dataset across all three metrics, KQP and KPB attain slightly higher fused scores than MQP. The ranking of individual backbones therefore does not directly determine the ranking of their fusions.

Table 7: Accuracy, macro F1, and AUC (%) at $K = 16$ shots on `LungHist700`, `HeidelbergSkin`, `Kather2016`, for the vision-language backbones and the fused sets whose members are all vision-language models. Mean ± standard deviation over 5 random seeds. **Bold** = best, underlined = second best in each column.

Method	LungHist700			HeidelbergSkin			Kather2016		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Vision–language foundation models (GCN adapter)									
KEEP	80.39 ±1.80	78.68 ±2.14	96.80 ±0.51	91.02 ±1.12	90.64 ±0.99	99.48 ±0.10	87.28 ±1.13	87.28 ±1.10	98.54 ±0.21
MUSK	72.98 ±2.59	70.75 ±2.68	95.04 ±0.77	85.64 ±0.53	83.27 ±0.40	99.09 ±0.07	87.70 ±0.21	87.66 ±0.21	98.76 ±0.10
QuiltNet	69.56 ±2.01	66.76 ±2.90	91.43 ±1.21	79.42 ±0.78	75.50 ±0.28	98.01 ±0.14	86.66 ±0.48	86.64 ±0.51	97.82 ±0.17
PLIP	69.95 ±2.71	67.19 ±2.26	91.55 ±1.38	79.90 ±0.66	75.51 ±0.61	98.23 ±0.08	86.42 ±0.78	86.34 ±0.84	98.58 ±0.13
BiomedCLIP	65.27 ±1.56	61.72 ±1.05	90.74 ±1.20	73.96 ±1.38	69.66 ±1.36	96.60 ±0.21	81.29 ±2.03	81.40 ±2.10	97.59 ±0.33
Fused sets (ours: OP alignment + combined graph + GCN)									
KEEP+QuiltNet+PLIP	85.76 ±1.46	84.11 ±1.63	98.09 ±0.47	**93.28** ±0.46	**92.68** ±0.32	**99.79** ±0.02	**91.65** ±1.54	**91.66** ±1.56	**99.27** ±0.23
Δ	+5.37	+5.43	+1.29	+2.26	+2.04	+0.31	+4.37	+4.38	+0.69
KEEP+PLIP+BiomedCLIP	**85.76** ±1.32	**84.25** ±1.38	**98.15** ±0.34	93.24 ±0.68	92.58 ±0.65	99.78 ±0.03	91.61 ±1.51	91.63 ±1.52	99.25 ±0.15
Δ	+5.37	+5.57	+1.35	+2.22	+1.94	+0.30	+4.33	+4.35	+0.67
MUSK+QuiltNet+PLIP	80.59 ±1.80	78.67 ±2.10	97.35 ±0.39	88.93 ±1.00	87.25 ±0.70	99.49 ±0.06	91.14 ±0.74	91.19 ±0.76	99.17 ±0.11
Δ	+7.61	+7.92	+2.31	+3.29	+3.98	+0.4	+3.44	+3.53	+0.41

`LubLung`. KQP achieves the highest mean accuracy (90.00%), macro-F1 (86.86%), and AUC (98.92%), improving upon KEEP by 4.38, 5.42, and 1.10 percentage points, respectively. KPB follows closely, with differences of 0.25, 0.44, and 0.05 percentage points across the three metrics. MQP achieves the largest macro-F1 gain over its constituents (6.00 percentage points). All three fused sets also show lower standard deviations than their strongest constituent across all three metrics, indicating less variation across the five random seeds.

`BRACS`. KPB achieves the highest mean accuracy (56.25%), macro-F1 (54.53%), and AUC (88.13%), with gains of 3.79, 2.42, and 2.57 percentage points over KEEP. Its macro-F1 exceeds KQP by 1.76 percentage points, the largest difference between these two configurations across the six datasets. KQP yields a smaller macro-F1 improvement of 0.66 percentage points, alongside a 2.12-percentage-point increase in AUC. MQP improves all three metrics over MUSK, but its macro-F1 (52.00%) remains slightly below KEEP alone (52.11%). Thus, all fused sets improve upon their own constituents, while KPB provides the largest gains across all three metrics on this dataset.

Table 8: Accuracy, macro F1 and AUC (%) at $K = 16$ shots on `LubLung`, `BRACS`, `BACH`, for the vision-language backbones and the fused sets whose members are all vision-language models. Mean ± standard deviation over 5 random seeds. **Bold** = best, underlined = second best in each column.

Method	LubLung			BRACS			BACH		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Vision–language foundation models (GCN adapter)									
KEEP	85.62 ±1.54	81.44 ±1.28	97.82 ±0.34	52.46 ±2.08	52.11 ±2.74	85.56 ±0.85	89.33 ±1.43	89.32 ±1.37	98.35 ±0.42
MUSK	80.16 ±2.83	75.73 ±2.58	96.66 ±0.36	50.49 ±4.20	50.55 ±3.78	84.53 ±1.80	83.83 ±1.08	84.00 ±1.12	95.62 ±0.26
QuiltNet	79.67 ±1.25	74.64 ±1.16	96.22 ±0.32	46.25 ±3.37	45.78 ±3.12	79.71 ±1.64	72.50 ±4.59	72.93 ±4.48	88.64 ±3.58
PLIP	79.31 ±2.10	74.53 ±2.85	96.11 ±0.48	43.16 ±3.08	43.03 ±2.54	78.91 ±1.13	72.33 ±3.75	72.60 ±3.71	88.30 ±1.06
BiomedCLIP	74.52 ±3.21	68.96 ±2.62	94.43 ±0.76	49.51 ±2.50	49.42 ±2.60	82.17 ±1.19	68.67 ±4.21	68.49 ±4.30	87.13 ±1.77
Fused sets (ours: OP alignment + combined graph + GCN)									
KEEP+QuiltNet+PLIP	**90.00** ±0.76	**86.86** ±0.99	**98.92** ±0.10	54.07 ±1.41	52.77 ±1.96	87.68 ±0.63	92.17 ±2.19	92.18 ±2.20	99.33 ±0.14
Δ	+4.38	+5.42	+1.1	+1.61	+0.66	+2.12	+2.84	+2.86	+0.98
KEEP+PLIP+BiomedCLIP	89.75 ±0.70	86.42 ±0.88	98.87 ±0.08	**56.25** ±2.65	**54.53** ±3.04	**88.13** ±0.69	**92.33** ±1.85	**92.38** ±1.86	**99.37** ±0.15
Δ	+4.13	+4.98	+1.05	+3.79	+2.42	+2.57	+3.00	+3.06	+1.02
MUSK+QuiltNet+PLIP	85.65 ±0.84	81.73 ±1.00	98.20 ±0.17	53.02 ±2.70	52.00 ±2.68	86.46 ±0.70	85.17 ±0.68	85.19 ±0.74	97.02 ±1.10
Δ	+5.49	+6.00	+1.54	+2.53	+1.45	+1.93	+1.34	+1.19	+1.40

BACH. KPB achieves the highest mean accuracy (92.33%), macro-F1 (92.38%), and AUC (99.37%), improving upon KEEP by 3.00, 3.06, and 1.02 percentage points, respectively. KQP performs similarly, trailing by only 0.16, 0.20, and 0.04 percentage points. MQP yields smaller accuracy and macro-F1 gains (1.34 and 1.19 percentage points), despite achieving the largest AUC gain (1.40 percentage points).

Tables 9 and 10 report accuracy, macro-F1, and AUC at $K = 8$ on `LungHist700`, `HeidelbergSkin`, and `Kather2016`, `LubLung`, `BRACS`, and `BACH` in the setting of $K = 8$ shots.

LungHist700. KQP achieves the highest mean accuracy (79.63%) and macro-F1 (77.40%), improving upon KEEP by 1.09 and 1.78 percentage points, respectively. KPB attains a marginally higher AUC (96.61% versus 96.58%), while its macro-F1 is 76.98%. For these two fusions, macro-F1 gains over KEEP are 1.36-1.78 percentage points, smaller than the 5.43-5.57 gains at $K = 16$. At eight shots, both also reduce the macro-F1 standard deviation from 6.87 for KEEP to 2.95 for KQP and 2.67 for KPB. MQP achieves the largest macro-F1 gain over its constituents (4.61 percentage points), but its final score (72.21%) remains below KEEP alone (75.62%).

HeidelbergSkin. KPB achieves the highest mean accuracy (92.02%), macro-F1 (91.48%), and AUC (99.69%), with gains of 2.69, 3.58, and 0.51 percentage points over KEEP. KQP follows with 90.74% macro-F1 and 99.67% AUC. The difference between these two configurations is therefore more pronounced in macro-F1 (0.74 percentage points) than in AUC (0.02 percentage points). MQP produces the largest macro-F1 gain over its constituents (4.39 percentage points), although its final score (84.25%) remains below the two KEEP-based fusions. Despite these differences in final macro-F1, all three fused sets achieve similar AUC gains of 0.49-0.52 percentage points.

Kather2016. KQP achieves the highest mean accuracy (90.84%), macro-F1 (90.87%), and AUC (99.10%), improving upon KEEP by 4.04, 4.04, and 0.72 percentage points, respectively. KPB follows closely with 90.53% macro-F1, corresponding to a gain of 3.70 percentage points over KEEP. MQP reaches 88.82% macro-F1, improving upon MUSK by 3.65 percentage points and also exceeding KEEP alone (86.83%). All three fused sets improve AUC by 0.50-0.72 percentage points. Relative to $K = 16$, macro-F1 is lower by 0.79 percentage points for KQP and 1.10 for KPB, compared with 2.37 for MQP.

Table 9: Accuracy, macro F1, and AUC (%) at $K = 8$ shots on `LungHist700`, `HeidelbergSkin`, `Kather2016`, for the vision-language backbones and the fused sets whose members are all vision-language models. Mean ± standard deviation over 5 random seeds. **Bold** = best, underlined = second best in each column.

Method	LungHist700			HeidelbergSkin			Kather2016		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Vision-language foundation models (GCN adapter)									
KEEP	78.54 ±5.63	75.62 ±6.87	95.68 ±0.98	89.33 ±0.67	87.90 ±0.61	99.18 ±0.17	86.80 ±0.83	86.83 ±1.08	98.38 ±0.12
MUSK	70.34 ±3.39	67.60 ±3.87	92.74 ±0.35	83.19 ±3.02	79.86 ±3.38	98.60 ±0.32	85.16 ±1.80	85.17 ±2.03	98.36 ±0.44
QuiltNet	61.83 ±8.05	57.93 ±9.33	89.03 ±2.52	76.07 ±1.09	70.89 ±1.12	96.80 ±0.27	83.59 ±1.55	83.56 ±1.41	97.12 ±0.42
PLIP	62.54 ±4.49	60.28 ±4.97	90.18 ±1.22	76.68 ±1.47	71.52 ±1.32	97.53 ±0.20	82.75 ±1.62	82.66 ±1.74	97.68 ±0.54
BiomedCLIP	60.10 ±3.37	56.28 ±3.78	89.78 ±1.18	69.56 ±1.42	64.68 ±1.38	95.13 ±0.30	77.70 ±2.14	77.77 ±2.36	96.52 ±0.44
Fused sets (ours: OP alignment + combined graph + GCN)									
KEEP+QuiltNet+PLIP	**79.63** ±2.58	**77.40** ±2.95	96.58 ±0.16	91.45 ±1.32	90.74 ±1.12	99.67 ±0.08	**90.84** ±0.99	**90.87** ±1.11	**99.10** ±0.18
Δ	+1.09	+1.78	+0.90	+2.12	+2.84	+0.49	+4.04	+4.04	+0.72
KEEP+PLIP+BiomedCLIP	79.39 ±2.44	76.98 ±2.67	**96.61** ±0.14	**92.02** ±1.34	**91.48** ±1.23	**99.69** ±0.07	90.53 ±0.83	90.53 ±1.02	99.03 ±0.11
Δ	+0.85	+1.36	+0.93	+2.69	+3.58	+0.51	+3.73	+3.7	+0.65
MUSK+QuiltNet+PLIP	74.88 ±2.50	72.21 ±2.87	94.78 ±0.15	86.40 ±0.72	84.25 ±0.34	99.12 ±0.07	88.81 ±0.79	88.82 ±0.97	98.86 ±0.19
Δ	+4.54	+4.61	+2.04	+3.21	+4.39	+0.52	+3.65	+3.65	+0.5

LubLung. KQP achieves the highest mean accuracy (87.81%), macro-F1 (83.48%), and AUC (98.51%), improving upon KEEP by 5.16, 5.95, and 1.43 percentage points, respectively. KPB also improves all three metrics over KEEP, although its macro-F1 (82.36%) is 1.12 percentage points below KQP. MQP produces the largest gains over its own constituents: 5.46, 6.59, and 1.84 percentage points in accuracy, macro-F1, and AUC, respectively. Its final accuracy (83.60%) and macro-F1 (78.59%) exceed KEEP alone.

Table 10: Accuracy, macro F1, and AUC (%) at $K = 8$ shots on LubLung, BRACS, BACH, for the vision-language backbones and the fused sets whose members are all vision-language models. Mean ± standard deviation over 5 random seeds. **Bold** = best, underlined = second best in each column.

Method	LubLung			BRACS			BACH		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Vision-language foundation models (GCN adapter)									
KEEP	82.65 ±0.99	77.53 ±1.20	97.08 ±0.55	53.07 ±1.78	50.96 ±2.34	84.74 ±1.07	89.37 ±3.15	89.35 ±3.14	98.50 ±0.02
MUSK	78.14 ±2.29	72.00 ±2.19	95.24 ±0.72	46.71 ±2.13	44.99 ±0.49	81.68 ±0.43	81.67 ±4.11	81.47 ±4.08	94.63 ±1.62
QuiltNet	77.24 ±0.86	71.76 ±0.55	94.71 ±0.19	44.43 ±1.53	42.31 ±2.24	76.32 ±1.68	67.71 ±5.09	67.82 ±4.94	85.45 ±4.05
PLIP	76.07 ±1.33	70.84 ±1.19	94.89 ±0.24	42.24 ±2.47	40.61 ±1.85	76.51 ±2.94	66.67 ±9.59	66.91 ±9.57	85.50 ±4.00
BiomedCLIP	69.02 ±1.16	63.30 ±0.84	92.05 ±0.68	43.46 ±2.39	43.44 ±2.33	78.76 ±1.13	70.63 ±1.73	70.52 ±1.70	85.19 ±0.41
Fused sets (ours: OP alignment + combined graph + GCN)									
KEEP+QuiltNet+PLIP	**87.81** ±1.41	**83.48** ±0.83	**98.51** ±0.07	**55.83** ±2.48	**53.23** ±3.31	**86.71** ±0.34	89.58 ±2.20	89.53 ±2.29	98.77 ±0.33
Δ	+5.16	+5.95	+1.43	+2.76	+2.27	+1.97	+0.21	+0.18	+0.27
KEEP+PLIP+BiomedCLIP	86.66 ±1.10	82.36 ±1.30	98.36 ±0.22	54.74 ±2.00	52.30 ±2.19	86.35 ±0.36	**89.79** ±2.68	**89.78** ±2.77	**98.89** ±0.24
Δ	+4.01	+4.83	+1.28	+1.67	+1.34	+1.61	+0.42	+0.43	+0.39
MUSK+QuiltNet+PLIP	83.60 ±1.17	78.59 ±1.14	97.08 ±0.78	50.53 ±1.43	48.47 ±1.16	83.84 ±0.70	82.71 ±4.11	82.60 ±4.35	96.31 ±1.20
Δ	+5.46	+6.59	+1.84	+3.82	+3.48	+2.16	+1.04	+1.13	+1.68

BRACS. KQP achieves the highest mean accuracy (55.83%), macro-F1 (53.23%), and AUC (86.71%), with gains of 2.76, 2.27, and 1.97 percentage points over KEEP. KPB follows with 54.74% accuracy, 52.30% macro-F1, and 86.35% AUC. The ordering of these two configurations differs from $K = 16$, where KPB leads across all three metrics. MQP achieves the largest gains over its constituents (3.82, 3.48, and 2.16 percentage points), but its final scores remain below KEEP alone. In particular, its macro-F1 increases from MUSK's 44.99% to 48.47%, compared with 50.96% for KEEP.

BACH. KPB achieves the highest mean accuracy (89.79%), macro-F1 (89.78%), and AUC (98.89%), but its gains over KEEP are modest at 0.42, 0.43, and 0.39 percentage points, respectively. KQP similarly yields small gains of 0.21, 0.18, and 0.27 percentage points. The macro-F1 gains of these KEEP-based fusions are therefore substantially smaller than at $K = 16$ (2.86-3.06 percentage points). MQP produces the largest gains over its own constituents across all three metrics, including 1.13 percentage points in macro-F1 and 1.68 in AUC.

Tables 11 and 12 report accuracy, macro-F1, and AUC at $K = 4$ across the six datasets, complementing the main-text macro-F1 results in Table 3.

LungHist700. KQP achieves the highest mean accuracy (70.63%) and macro-F1 (66.76%), improving upon KEEP by 2.04 and 1.43 percentage points, respectively. KPB achieves the highest AUC (94.61%), with a gain of 1.43 percentage points over KEEP. MQP produces the largest gains over its constituents across all three metrics.

HeidelbergSkin. KPB achieves the highest mean accuracy (90.00%), macro-F1 (88.38%), and AUC (99.65%), with gains of 2.50, 2.12, and 0.61 percentage points over KEEP. KQP follows closely, reaching 88.30% macro-F1 and 99.61% AUC. MQP achieves larger macro-F1 and AUC gains over its constituents (3.38 and 0.99 percentage points), but its final scores remain below KEEP alone. The two KEEP-based fusions therefore retain the strongest absolute performance despite the larger relative improvements obtained by MQP.

Kather2016. MQP achieves the highest mean accuracy (82.94%), macro-F1 (82.88%), and AUC (97.85%), improving upon MUSK by 3.51, 3.44, and 0.77 percentage points, respectively. The three fused configurations are closely grouped, with differences of at most 0.41 percentage points in accuracy and macro-F1 and 0.22 in AUC. KQP and KPB achieve larger macro-F1 gains of 6.07 and 6.29 percentage points. These gains are measured against KEEP's lower baseline macro-F1 (76.40%), whereas MQP is compared with MUSK (79.44%). Thus, the larger gains of the KEEP-based fusions coexist with slightly lower final scores.

Table 11: Accuracy, macro F1 and AUC (%) at $K = 4$ shots on `LungHist700`, `HeidelbergSkin`, `Kather2016`, for the vision-language backbones and the fused sets whose members are all vision-language models. Mean ± standard deviation over 5 random seeds. **Bold** = best, underlined = second best in each column.

Method	LungHist700			HeidelbergSkin			Kather2016		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Vision-language foundation models (GCN adapter)									
KEEP	68.59 ±3.54	65.33 ±3.02	93.18 ±0.92	87.50 ±0.85	86.26 ±0.83	99.04 ±0.26	76.70 ±1.18	76.40 ±1.66	96.63 ±0.78
MUSK	64.10 ±3.92	60.52 ±4.39	89.91 ±1.61	78.97 ±1.90	74.49 ±2.75	97.62 ±0.67	79.43 ±3.27	79.44 ±3.30	97.08 ±1.22
QuiltNet	57.93 ±1.02	53.15 ±1.89	84.89 ±3.78	69.30 ±1.40	63.89 ±1.27	94.92 ±0.95	75.64 ±3.65	75.42 ±3.93	94.88 ±1.24
PLIP	55.37 ±3.97	50.94 ±4.08	86.97 ±1.68	69.67 ±0.96	64.76 ±0.57	96.28 ±0.44	76.53 ±2.68	76.34 ±2.79	96.13 ±0.82
BiomedCLIP	57.32 ±6.56	52.62 ±7.02	88.92 ±2.45	62.74 ±2.37	58.00 ±3.83	93.81 ±0.64	71.29 ±4.31	70.72 ±5.14	94.83 ±1.38
Fused sets (ours: OP alignment + combined graph + GCN)									
KEEP+QuiltNet+PLIP	**70.63** ±6.52	**66.76** ±6.63	94.17 ±0.98	89.58 ±1.16	88.30 ±1.71	99.61 ±0.08	82.53 ±3.65	82.47 ±3.62	97.63 ±0.89
Δ	+2.04	+1.43	+0.99	+2.08	+2.04	+0.57	+5.83	+6.07	+1.00
KEEP+PLIP+BiomedCLIP	70.24 ±6.34	66.19 ±6.87	**94.61** ±0.79	**90.00** ±1.47	**88.38** ±2.23	**99.65** ±0.05	82.79 ±3.51	82.69 ±3.48	97.82 ±0.77
Δ	+1.65	+0.86	+1.43	+2.50	+2.12	+0.61	+6.09	+6.29	+1.19
MUSK+QuiltNet+PLIP	66.34 ±3.75	62.08 ±4.37	92.50 ±1.14	81.40 ±1.21	77.87 ±1.24	98.61 ±0.11	**82.94** ±4.29	**82.88** ±4.36	**97.85** ±0.93
Δ	+2.24	+1.56	+2.59	+2.43	+3.38	+0.99	+3.51	+3.44	+0.77

`LubLung`. KQP achieves the highest mean accuracy (82.40%), macro-F1 (76.08%), and AUC (97.45%), with gains of 3.05, 3.98, and 1.82 percentage points over KEEP. KPB performs similarly, differing by at most 0.26 percentage points across the three metrics. MQP yields the largest gains over its constituents: 6.49, 6.54, and 2.39 percentage points in accuracy, macro-F1, and AUC, respectively. For MQP, the strongest constituent is QUILTNET for accuracy and macro-F1, but MUSK for AUC. Despite these larger gains, MQP's final macro-F1 (68.98%) remains below KEEP alone (72.10%).

`BRACS`. KPB achieves the highest mean accuracy (50.95%), macro-F1 (49.53%), and AUC (85.64%), improving upon KEEP by 1.48, 0.89, and 0.72 percentage points, respectively. KQP improves accuracy and AUC by 0.53 percentage points each, but its macro-F1 decreases from 48.64% to 48.33%. MQP improves all three metrics over MUSK, including a macro-F1 increase of 3.55 percentage points. Its final macro-F1 (42.22%), however, remains below KEEP and both KEEP-based fusions.

Table 12: Accuracy, macro F1, and AUC (%) at $K = 4$ shots on `LubLung`, `BRACS`, `BACH`, for the vision-language backbones and the fused sets whose members are all vision-language models. Mean ± standard deviation over 5 random seeds. **Bold** = best, underlined = second best in each column.

Method	LubLung			BRACS			BACH		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Vision-language foundation models (GCN adapter)									
KEEP	79.35 ±3.12	72.10 ±4.99	95.63 ±1.79	49.47 ±5.62	48.64 ±4.91	84.92 ±0.57	**86.46** ±9.73	**86.26** ±9.94	97.59 ±1.54
MUSK	68.80 ±2.79	61.51 ±3.18	92.73 ±1.20	40.44 ±3.58	38.67 ±2.09	79.54 ±1.10	75.42 ±6.99	75.66 ±6.89	92.22 ±2.61
QuiltNet	69.15 ±1.62	62.44 ±1.68	91.68 ±0.27	36.49 ±2.17	34.83 ±2.90	73.45 ±3.21	52.92 ±5.67	52.17 ±6.43	78.74 ±2.26
PLIP	66.75 ±3.59	61.50 ±1.50	92.04 ±0.70	33.64 ±3.01	32.35 ±2.63	73.34 ±1.22	56.25 ±6.61	55.72 ±7.00	78.51 ±2.36
BiomedCLIP	63.17 ±2.03	55.96 ±0.75	90.48 ±1.21	38.99 ±3.12	37.65 ±2.81	77.50 ±1.08	62.71 ±7.92	62.04 ±8.48	83.05 ±3.09
Fused sets (ours: OP alignment + combined graph + GCN)									
KEEP+QuiltNet+PLIP	**82.40** ±3.21	**76.08** ±4.62	**97.45** ±0.43	50.00 ±3.53	48.33 ±2.32	85.45 ±0.85	84.83 ±3.43	84.61 ±3.57	97.33 ±1.30
Δ	+3.05	+3.98	+1.82	+0.53	−0.31	+0.53	−1.63	−1.65	−0.26
KEEP+PLIP+BiomedCLIP	82.14 ±2.02	76.02 ±3.67	97.40 ±0.72	**50.95** ±4.00	**49.53** ±2.69	**85.64** ±0.97	86.17 ±2.64	85.94 ±2.78	**97.61** ±1.24
Δ	2.79	+3.92	+1.77	+1.48	+0.89	+0.72	−0.29	−0.32	+0.02
MUSK+QuiltNet+PLIP	75.64 ±1.22	68.98 ±2.28	95.12 ±0.83	44.28 ±4.20	42.22 ±3.48	81.54 ±0.69	74.67 ±5.29	74.23 ±5.80	92.80 ±3.45
Δ	+6.49	+6.54	+2.39	+3.84	+3.55	+2.00	−0.75	−1.43	+0.58

`BACH`. KEEP retains the highest mean accuracy (86.46%) and macro-F1 (86.26%). All three fused sets yield lower accuracy and macro-F1 than their strongest constituent, with macro-F1 decreases of 1.65, 0.32, and 1.43 percentage points for KQP, KPB, and MQP, respectively. AUC follows a different pattern: KPB achieves the highest value (97.61%), although its gain over KEEP is

only 0.02 percentage points. MQP also improves AUC by 0.58 percentage points despite its lower accuracy and macro-F1. For the KEEP-based fusions, macro-F1 standard deviations decrease from 9.94 for KEEP to 3.57 for KQP and 2.78 for KPB.