REPRESENTATION COMPOSITION MATTERS: ANALYZING AND FUSING PATHOLOGY FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The emergence of pathology foundation models (FMs) provides diverse visual representations for computational pathology, yet when and how to effectively combine multiple FMs remains unclear. Existing fusion approaches typically aggregate multiple foundation models using fixed aggregation or learned fusion mechanisms, without explicitly considering how encoder utility depends on representation composition and downstream task requirements. To address this gap, we conduct a three-perspective analysis of multi-foundation-model fusion, examining representation geometry, feature-wise statistical association, and downstream functional contribution. Through this analysis, we find that representation similarity alone is insufficient to characterize functional redundancy across the evaluated foundation-model compositions. Moreover, no single fusion strategy among those evaluated consistently dominates across compositions, and fusion performance does not improve monotonically as the number of integrated encoders increases, indicating that fusion effectiveness varies with representation composition. Motivated by these findings, we propose CoR-Fuse, a lightweight, plug-and-play consensus-aware fusion module for adaptive multi-foundation-model integration. Experimental results on TCGA-UVM and TCGA-BLCA show that CoR-Fuse achieves competitive performance across the evaluated foundation-model compositions, while requiring fewer parameters and FLOPs than attention-based fusion strategies. Code is provided in the supplementary material.

1 INTRODUCTION

Pathology foundation models (FMs) provide diverse histological representations learned from large-scale WSI datasets (Lu et al., 2024; Xu et al., 2024; Nechaev et al., 2024; Saillard et al., 2024; Scalbert et al., 2026; Karasikov et al., 2025; Filiot et al., 2023; 2024; Chen et al., 2024; Vorontsov et al., 2024; Zimmermann et al., 2024). However, these models are trained under heterogeneous paradigms, yielding varied representations.

Nevertheless, no single pathology foundation model universally dominates all clinical diagnostic tasks (Boutaj et al., 2026; Vanderbilt et al., 2026). This limitation raises a critical challenge: how to effectively integrate heterogeneous foundation model representations. Although different foundation models provide diverse feature views of WSI data, simply increasing the number of integrated models or applying naive fusion strategies does not necessarily improve fusion performance.

Existing fusion approaches, including feature concatenation, attention-based methods (Vaswani et al., 2017), and mixture-of-experts (MoE) routing (Shazeer et al., 2017), have explored increasingly adaptive mechanisms for integrating representations. However, the relationship between foundation-model representation characteristics and downstream fusion behavior remains insufficiently understood.

In this work, we formulate multi-FM fusion as estimating sample-specific encoder contributions to the downstream task, rather than applying fixed weights. Each foundation model may contribute useful information, but its relative contribution can vary across models and samples. To better understand the underlying factors affecting multi-foundation-model fusion, we establish a systematic analysis framework that investigates foundation model relationships from three perspectives.

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.

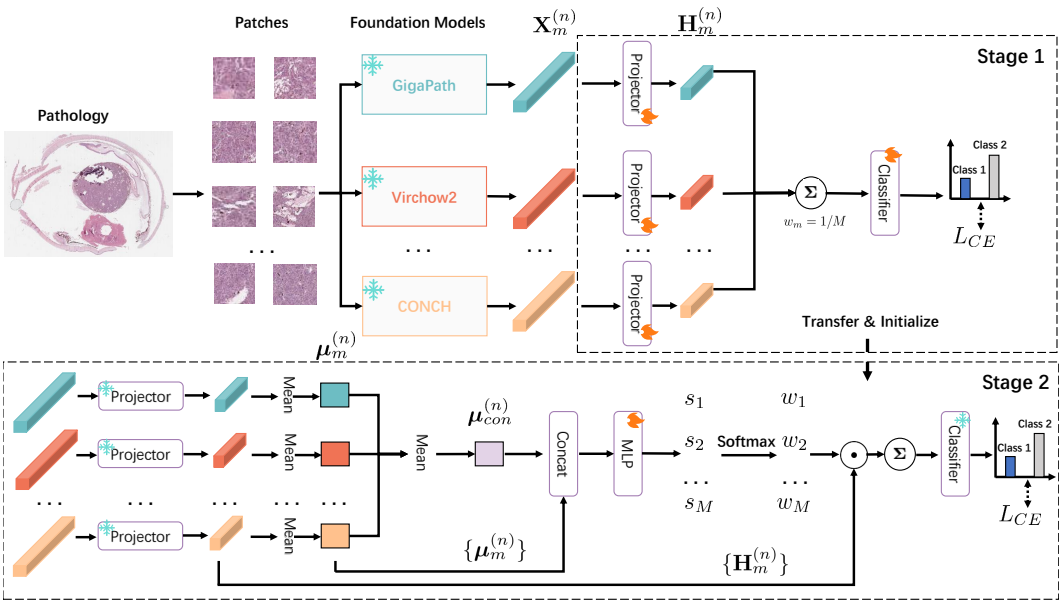

Figure 1: Overview of CoR-Fuse. Stage 1 initializes projection heads and the downstream classifier with mean fusion, while Stage 2 uses a consensus representation for adaptive routing. For slide n and foundation model m, patch representations $\mathbf{X}_m^{(n)}$ are projected to $\mathbf{H}_m^{(n)}$ and mean-pooled into $\mu_m^{(n)}$. CoR-Fuse produces one set of slide-level routing weights, shared across all patch embeddings to preserve the spatial bag structure, and is compatible with MIL, MLP, and GNN downstream architectures.

On TCGA-UVM, through systematic analyses based on this three-perspective framework, we derive several insightful observations: 1. Fusion effectiveness varies with foundation-model composition rather than the number of integrated models alone. 2. Representation similarity alone does not reliably indicate whether an encoder contributes redundant or complementary task-relevant information. 3. No single fusion strategy consistently dominates, the relative effectiveness of fusion strategies varies with the composition of the integrated representations. 4. Per-model marginal contribution is sample-dependent, with substantial variation across slides.

Driven by these key insights, we further develop CoR-Fuse (Consensus-aware Routing Fusion), a lightweight, plug-and-play routing module for multi-foundation-model fusion that combines encoder-specific representations with cross-encoder consensus to estimate task-adaptive encoder weights through a two-stage optimization strategy (Figure 1). CoR-Fuse constructs a fused representation compatible with downstream architectures, including multiple instance learning (MIL) (Ilse et al., 2018), graph neural networks (GNNs) (Scarselli et al., 2008) and MLP-based classifiers. We extensively evaluate CoR-Fuse across multiple TCGA cohorts (The Cancer Genome Atlas Research Network, 2013), investigating its robustness under different foundation-model compositions, fusion strategies, and computational settings.

Our main contributions are summarized as follows:

- We identify model-specific and sample-specific variation in encoder contributions in multi-foundation-model fusion.
- We develop a three-perspective analysis of representation geometry, feature association, and task-level contribution, suggesting that representation similarity alone does not fully characterize functional redundancy or fusion behavior.
- We show that CoR-Fuse provides competitive predictive performance while exhibiting lower empirical composition-level performance dispersion across the evaluated encoder compositions, suggesting more consistent behavior under changes in foundation-model composition.

2 Related Work

Recent studies have benchmarked pathology foundation models across downstream predictive performance, transferability, and clinical tasks (Lei et al., 2026a; Boutaj et al., 2026; Vanderbilt et al., 2026). With the growing number of pathology FMs, multi-FM fusion has emerged as a complementary direction, ranging from fixed aggregation such as concatenation and averaging to learnable gated and attention-based fusion. Other approaches explore adaptive ensemble paradigms, including logit-level fusion (Huang et al., 2026), group-based encoder selection (Quan et al., 2025), and ensemble learning for precision oncology (Luo et al., 2026). Recent methods further investigate unified or adaptive integration of multiple FMs, including Meta-Encoder, AdaFusion, Shazam, and FuseCPath (Gao et al., 2026; Xiao et al., 2025; Lei et al., 2026b; Yang et al., 2025). These methods primarily focus on learning fusion or selection mechanisms, but do not systematically analyze how representation composition affects individual FM task-level contributions.

Mixture-of-experts (MoE) approaches provide another paradigm for adaptive expert selection. Recent studies have explored multimodal MoE for immune biomarker prediction and structured routing based on region-graph optimal transport for WSI classification (Liu et al., 2026; Tian et al., 2026). These works highlight the potential of expert specialization and adaptive routing, but primarily study expert selection or structured routing rather than how the composition of foundation-model representations affects downstream fusion behavior.

Representation similarity has been widely studied for characterizing relationships among learned feature spaces, with centered kernel alignment (CKA) being a common measure (Kornblith et al., 2019). Recent work has also analyzed pathology FM representations, reporting slide-dependent representation structure and differences between vision-only and vision-language models (Mishra & Lotter, 2025). However, these studies mainly characterize individual representations rather than their task-conditioned interactions within multi-FM fusion. We therefore conduct a multi-perspective analysis of representation geometry, feature association, and task-level contribution, and examine how representation composition relates to fusion behavior.

3 Diagnosing Foundation-Model Relationships for Fusion

To understand why different foundation-model compositions lead to different fusion performance, we analyze encoder relationships at three complementary levels: global representation geometry, feature-wise statistical association, and task-conditioned encoder attribution and pairwise interaction.

3.1 Representation Geometry Reveals Non-uniform Similarity Structures

We investigate representation geometry by computing coordinate-aligned linear CKA between encoder pairs (Figure 2). The analysis results reveal non-uniform similarity structures: CKA values range from 0.65 to 0.97, with encoders exhibiting varied representation profiles. Hierarchical clustering (Appendix Figure 3) further reveals non-uniform grouping structures across the 13 foundation models: 1. Hibou-L and Virchow2 form relatively distinct branches, exhibiting lower CKA similarity with other models. 2. CONCH, Phikon, UNI, GigaPath, and H-optimus share more similar feature spaces, with fine-grained sub-clusters among model variants.

While CKA measures global geometric similarity, it does not reveal feature-level statistical dependencies. We therefore compute cross-feature Spearman correlations between encoder pairs using coordinate-aligned patches. Let $\mathbf{X}_m^{(n)} \in \mathbb{R}^{P_n \times d_m}$ denote the embedding matrix produced by encoder m for slide n, where P_n is the number of spatially aligned patches and d_m is the embedding dimension of encoder m. For the a-th feature of encoder i and the b-th feature of encoder j, we define

$$\mathbf{x}_{m,a}^{(n)} = \mathbf{X}_m^{(n)}[:, a] \in \mathbb{R}^{P_n}. \tag{1}$$

The Spearman rank correlation across aligned patches is then computed as

$$\rho_{ijab}^{(n)} = \mathrm{corr}\left(\mathrm{rank}\left(\mathbf{x}_{i,a}^{(n)}\right), \mathrm{rank}\left(\mathbf{x}_{j,b}^{(n)}\right)\right). \tag{2}$$

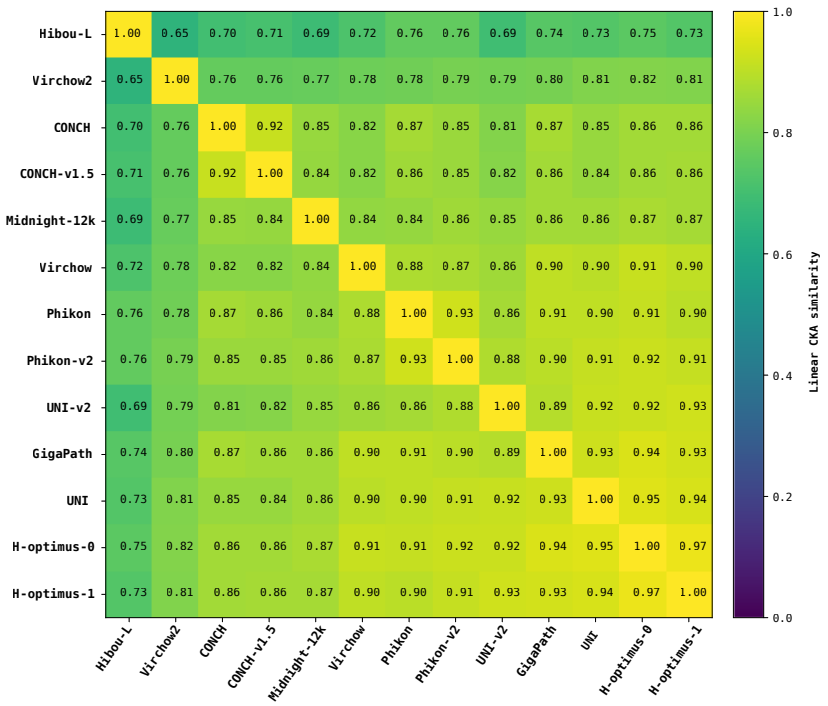

Figure 2: Representation similarity structures among pathology foundation models measured by Linear CKA.

Each encoder pair i, j yields $d_i d_j$ cross-feature correlations for slide n. We summarize the feature-level statistical association between the two encoders by averaging the absolute correlations over all feature pairs and slides:

$$R_{ij} = \frac{1}{N} \sum_{n=1}^{N} \left[\frac{1}{d_i d_j} \sum_{a=1}^{d_i} \sum_{b=1}^{d_j} \left| \rho_{ijab}^{(n)} \right| \right] \tag{3}$$

The absolute value measures the strength of monotonic association regardless of correlation direction. Finally, to quantify the relationship between CKA-based representation similarity and feature-wise statistical association, let $\overline{C}_{ij}$ denote the slide-averaged Linear CKA between encoders i and j, and let R_{ij} denote their mean absolute feature-wise Spearman association. We collect all unique encoder pairs from the upper-triangular part of the pairwise matrices ($i < j$) and compute:

$$\rho_{\text{CKA-FS}} = \text{Spearman}\left(\left\{ \overline{C}_{ij} \right\}_{i<j}, \left\{ R_{ij} \right\}_{i<j} \right) \tag{4}$$

This statistic measures the rank-order association between global representation similarity and feature-wise statistical association. We observe a weak negative rank association between global representation similarity and mean absolute cross-feature Spearman association. This association suggests that, for the evaluated encoder compositions, geometric representation similarity and cross-feature statistical association capture different aspects of encoder relationships.

3.2 Variation in Fusion Performance Across Representation Compositions

Although foundation models exhibit diverse representation structures, it remains unclear whether increasing the number of integrated encoders alone can reliably improve fusion performance. We observe non-monotonic and method-dependent trends: adding an encoder may improve performance under one representation composition while providing limited benefit or degradation under another.

Therefore, encoder quantity alone is insufficient to explain fusion behavior; effective fusion requires understanding representation composition. Detailed experimental results are discussed in Section 5.2.1.

3.3 Sample-specific Encoder Contributions Motivate Adaptive Routing

While representation-level and feature-level relationships characterize encoder similarities, they do not quantify whether an encoder improves downstream task performance when added to a multi-encoder fusion. Therefore, we perform task-driven analysis to quantify the variation in encoder-specific contributions across individual slides. We adopt a leave-one-encoder-out (LOO) attribution strategy, where the representation of the held-out encoder is replaced with an encoder-specific mean baseline. For a trained reference classifier using the complete encoder set $\mathcal{E}$, let $f_y(\mathcal{E}^{(n)})$ denote the target-class one-vs-rest logit margin for slide n. For encoder m, its LOO attribution is defined as:

$$C_m^{\mathrm{LOO}}(n) = f_y\big(\mathcal{E}^{(n)}\big) - f_y\big(\mathcal{E}^{(n)}_{\setminus m}\big), \tag{5}$$

where $\mathcal{E}^{(n)}_{\setminus m}$ denotes the encoder set in which the embedding of encoder m is replaced by an encoder-specific mean baseline, while the representations of all other encoders remain unchanged. The LOO analysis reveals that different foundation models exhibit distinct contribution patterns across samples. To further investigate whether encoder pairs exhibit non-additive interactions, we extend the analysis to pairwise task-level interaction. For a pair of encoders (i, j), we define the pairwise interaction attribution as:

$$I_{ij}(n) = f_y(\mathcal{E}^{(n)}) - f_y(\mathcal{E}^{(n)}_{\setminus i}) - f_y(\mathcal{E}^{(n)}_{\setminus j}) + f_y(\mathcal{E}^{(n)}_{\setminus \{i,j\}}) \tag{6}$$

This inclusion–exclusion score quantifies the non-additive task-specific interaction between two encoders, where positive and negative values indicate synergistic and interfering effects, respectively. The LOO attribution and pairwise interaction analyses provide offline diagnostic evidence motivating adaptive routing. The interaction analysis further suggests that representation similarity does not always align with task-level interaction (Appendix Figure 6). Therefore, representation similarity alone may not fully characterize encoder redundancy, motivating adaptive routing that accounts for encoder utility at the sample level.

4 CoR-Fuse: Consensus-aware Routing Fusion

4.1 Problem Formulation

For each slide n, we denote the patch-level representation extracted by foundation model m as:

$$\mathbf{X}_m^{(n)} = [\mathbf{x}_{m,1}^{(n)}, \ldots, \mathbf{x}_{m,P_n}^{(n)}]^\top \in \mathbb{R}^{P_n \times d_m}, \tag{7}$$

where P_n denotes the number of patches in slide n, and d_m is the embedding dimension of encoder m. All encoders operate on coordinate-aligned patches extracted from the same WSI, enabling patch-level correspondence across heterogeneous representations. Given a set of M foundation models $\mathcal{E} = \{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$, our goal is to learn a slide-level predictor:

$$p_\theta(y^{(n)} | \mathbf{X}_1^{(n)}, \ldots, \mathbf{X}_M^{(n)}) \tag{8}$$

The final prediction is obtained from the fused representation generated by the proposed fusion module.

4.2 Two-Stage Optimization for Adaptive Routing

To stabilize adaptive multi-encoder fusion, we adopt a two-stage optimization strategy consisting of mean-fusion warm start and consensus-guided adaptive routing. CoR-Fuse performs routing at the slide level to capture semantic-level encoder utility rather than local texture patterns. The detailed discussion of slide-level routing design choices is provided in Appendix J.

4.2.1 STAGE 1: MEAN-FUSION INITIALIZATION

Since pathology foundation models produce representations with different embedding dimensions, we first project all embeddings into a shared latent space:

$$\mathbf{H}_m^{(n)} = P_m(\mathbf{X}_m^{(n)}) \in \mathbb{R}^{P_n \times 512}, \tag{9}$$

where $P_m(\cdot)$ denotes an encoder-specific projection head that projects the features into a shared 512-dimensional latent space. We first train a reference fusion model using uniform mean fusion. The fused patch representations are subsequently aggregated by an MIL classifier to produce slide-level predictions. The reference model is trained by minimizing the cross-entropy loss:

$$\mathcal{L}_T = \mathrm{CE}(p_\theta(\mathbf{X}_1^{(n)}, \ldots, \mathbf{X}_M^{(n)}), y^{(n)}) \tag{10}$$

where $y^{(n)}$ denotes the ground-truth label of slide n, and $p_\theta(\cdot)$ denotes the predicted class probabilities produced by the reference model. We use a strictly nested evaluation protocol to prevent information leakage from the outer test fold. Within each outer training split, the checkpoint is selected based on inner-validation AUC and is subsequently frozen. At the end of the warm-start stage, the projection heads and MIL classifier are transferred to initialize the adaptive fusion model.

4.2.2 STAGE 2: CONSENSUS-GUIDED ADAPTIVE ROUTING

Given the initialized shared representation space, we introduce a lightweight encoder-level router to estimate slide-specific fusion weights. Importantly, this pooling operation is only used for routing and does not replace the patch-level representations used by the downstream MIL classifier. For each encoder m, we aggregate its projected patch embeddings into a slide-level descriptor through mean pooling:

$$\mu_m^{(n)} = \frac{1}{P_n} \sum_{p=1}^{P_n} \mathbf{H}_{m,p}^{(n)} \in \mathbb{R}^{512}. \tag{11}$$

We then construct a consensus representation as:

$$\mu_{con}^{(n)} = \frac{1}{M} \sum_{m=1}^{M} \mu_m^{(n)} \in \mathbb{R}^{512}. \tag{12}$$

The consensus representation summarizes the average semantic content across all encoders for that slide. These consensus-conditioned encoder descriptors are subsequently used by the adaptive router to estimate slide-dependent routing scores. Each encoder descriptor is concatenated with the consensus representation and fed into a two-layer MLP router:

$$s_m^{(n)} = \mathrm{MLP}([\mu_m^{(n)}, \mu_{con}^{(n)}]) + b_m, \tag{13}$$

where b_m is a learnable encoder-specific bias capturing the global preference of each foundation model. The routing scores $s_m^{(n)}$ are converted into normalized weights $w_m^{(n)}$ through softmax, and the final fused patch representation is obtained as:

$$\mathbf{Z}_p^{(n)} = \sum_{m=1}^{M} w_m^{(n)} \mathbf{H}_{m,p}^{(n)} \tag{14}$$

Although routing decisions are made at the slide level, fusion is performed on patch embeddings to preserve the spatial correspondence required by downstream MIL aggregation. The router and fusion model are optimized jointly using only the downstream classification objective. Unlike distillation-based routing approaches, our method does not require additional teacher-student matching objectives or attribution supervision.

4.3 FUSION PREDICTION

The routing weights are applied to the patch-level representations from all foundation models to construct the fused patch bag. The fused representations are subsequently processed by an ABMIL classifier (Ilse et al., 2018), which performs patch-level attention aggregation and generates slide-level predictions. The entire CoR-Fuse module is optimized with the same downstream classification objective.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. To validate the observations mentioned above and evaluate multi-foundation-model fusion across different fusion strategies, we conduct experiments on two TCGA datasets, including TCGA-UVM (Jager et al., 2018) and TCGA-BLCA (The Cancer Genome Atlas Research Network, 2014). To examine whether above insights hold across diverse foundation-model compositions, we use TCGA-UVM as a diagnostic study setting and extract representations from 13 pathology foundation models. This broad model coverage enables us to systematically evaluate how representation relationships and encoder composition relate to downstream fusion behavior. Following previous molecular characterization studies, we formulate the task as D3/M3 subtype binary classification (Robertson et al., 2017), with 38 D3 cases and 42 M3 cases. We additionally evaluate CoR-Fuse on TCGA-BLCA, a larger dataset with 457 slides. The task is formulated as binary prediction of TP53 mutation status from WSI representations, distinguishing TP53-mutant and TP53-wildtype cases (The Cancer Genome Atlas Research Network, 2014), with 238 mutant and 219 wildtype cases.

Implementation Details. For both TCGA-UVM and TCGA-BLCA datasets, we perform patient-level stratified five-fold cross-validation and report the average performance across five folds. All foundation-model embeddings are extracted using the same preprocessing pipeline. Performance is evaluated using the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPRC). Computational efficiency is assessed in terms of trainable parameters and floating-point operations (FLOPs). Detailed training configurations and hyperparameter settings are provided in Appendix E.

5.2 EXPERIMENTAL RESULTS

5.2.1 MULTI-FOUNDATION-MODEL FUSION UNDER DIFFERENT REPRESENTATION COMPOSITIONS

Individual Foundation Model Performance. We first evaluate all 13 FMs on TCGA-UVM (Appendix Table 5). The individual performance varies substantially across foundation models, suggesting that individual FMs provide varying levels of task relevance and motivating representation-aware fusion.

Plug-and-Play Compatibility. To evaluate the plug-and-play capability of CoR-Fuse, we examine its compatibility with both different foundation-model compositions and downstream prediction heads. On the downstream side, we combine CoR-Fuse with four downstream heads, including ABMIL, TransMIL (Shao et al., 2021), MLP, and GNN-based classifiers. The results in Appendix Table 14 show that CoR-Fuse can be effectively combined with different architectures. On the encoder side, CoR-Fuse can adapt to different numbers and combinations of foundation models without requiring a fixed encoder set.

Comparison with Existing Fusion Strategies. Based on the above analysis, we hypothesize that fusion effectiveness depends on the composition of integrated representations rather than the number of encoders alone. To examine this hypothesis, we evaluate representative fusion strategies across different foundation-model compositions. The compared representative fusion paradigms (detailed in Appendix G), include: **Fixed aggregation:** Mean fusion, Concatenation; **Global learned weighting:** Global weighting fusion; **Slide-specific weighting:** Gated fusion; **Interaction-based fusion:** Cross-attention, Self-attention.

We further compare CoR-Fuse with recently proposed multi-foundation-model fusion methods such as Meta-Encoder (Gao et al., 2026) and FuseCPath (Yang et al., 2025). Results in Appendix Table 7 show that CoR-Fuse provides competitive performance against both generic fusion strategies and task-oriented fusion frameworks. We additionally compare with ELF (Luo et al., 2026) using its prescribed fixed encoder composition, {H-optimus-0, Virchow2, UNI, GigaPath, CONCH-v1.5}. Under this setting, CoR-Fuse achieves the highest AUC, while remaining competitive in AUPRC and F1 (Appendix Table 8).

We fix the number of fused encoders to five and evaluate three representative encoder compositions spanning high, intermediate, and heterogeneous representation similarity, compositions were

selected according to pre-defined CKA ranges before evaluating fusion performance (Appendix Table 6). We also report the mean and best individual-FM AUC within each composition to characterize the predictive strength of the selected encoders. For the high-similarity encoder composition {H-optimus-0, H-optimus-1, UNI, GigaPath, Phikon-v2}, attention-based fusion yields higher performance than simpler aggregation strategies (Appendix Table 10). For the heterogeneous composition {Hibou-L, Virchow2, CONCH-v1.5, UNI-v2, H-optimus-0}, simpler fusion strategies outperform attention-based methods (Appendix Table 9), indicating that complex interaction modeling is not universally beneficial. Finally, we evaluate an intermediate-similarity composition {H-optimus-1, H-optimus-0, Virchow2, UNI-v2, CONCH-v1.5} (Appendix Table 11). As shown in Table 1, across all three representative compositions, CoR-Fuse achieves competitive performance with low additional computational overhead at the fusion stage.

Table 1: Summary of fusion performance across three representative five-encoder compositions on TCGA-UVM. Values are reported as mean ± standard deviation across the three compositions.

Method	AUC	AUPRC	F1	Params(M)	FLOPs(G)
Mean Fusion	0.82 ± 0.03	0.86 ± 0.03	0.73 ± 0.02	4.01	72.34
Concatenation	0.81 ± 0.04	0.86 ± 0.03	0.73 ± 0.04	5.32	96.03
Global Weight	0.82 ± 0.03	0.86 ± 0.04	0.74 ± 0.03	4.01	72.34
Gated Fusion	0.81 ± 0.02	0.85 ± 0.01	0.73 ± 0.02	4.01	72.34
Cross Attention	0.83 ± 0.03	0.86 ± 0.02	0.74 ± 0.06	6.11	148.55
Self Attention	0.81 ± 0.04	0.85 ± 0.03	0.73 ± 0.03	24.19	2235.86
CoR-Fuse	$\mathbf{0.83 \pm 0.01}$	$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.75 \pm 0.02}$	4.07	72.34

Fusion Performance with Different Foundation Model Set Sizes. We systematically evaluate fusion strategies under different foundation-model set sizes by increasing the number of integrated encoders from two to ten, following the ranking of individual encoder performance on the validation set. As shown in Appendix Figure 4, increasing the number of foundation models does not necessarily improve fusion performance, and the benefit of additional encoders depends on the encoder composition and fusion strategy. Across encoder-set sizes, CoR-Fuse exhibits the smallest empirical AUC dispersion among the evaluated methods, suggesting more consistent performance across different FM compositions (Appendix Table 4).

Table 2: Fusion performance under a representative five-encoder composition on TCGA-BLCA.

Method	AUC	AUPRC	F1	Params(M)	FLOPs(G)
Mean Fusion	0.74 ± 0.06	0.74 ± 0.10	0.60 ± 0.10	4.01	74.81
Concatenation	0.74 ± 0.06	0.75 ± 0.08	0.57 ± 0.23	5.32	99.31
Global Weight	0.74 ± 0.05	0.75 ± 0.09	0.65 ± 0.07	4.01	74.81
Gated Fusion	0.73 ± 0.06	0.74 ± 0.10	$\mathbf{0.68 \pm 0.06}$	4.01	74.81
Cross Attention	0.74 ± 0.03	0.73 ± 0.06	0.57 ± 0.17	6.11	153.62
Meta-Encoder	0.64 ± 0.06	0.63 ± 0.16	0.63 ± 0.11	24.19	2235.86
FuseCPath	0.67 ± 0.06	0.67 ± 0.16	0.59 ± 0.15	7.16	3.21
CoR-Fuse	$\mathbf{0.75 \pm 0.07}$	$\mathbf{0.76 \pm 0.12}$	0.68 ± 0.08	4.04	74.81

5.2.2 Evaluation on TCGA-BLCA Cohort

To evaluate fusion under a practically motivated, limited yet diverse encoder composition, we select five representative pathology foundation models spanning different architectures, pretraining strategies, and model scales: {UNI, Virchow2, Midnight-12k, GigaPath, Phikon}. More detailed information is provided in Appendix D. As shown in Table 2, CoR-Fuse achieves competitive performance and lower computational cost than Cross Attention and Meta-Encoder.

Table 3: Ablation study of CoR-Fuse components.

Method	TCGA-UVM		TCGA-BLCA	
	AUC	AUPRC	AUC	AUPRC
w/o Consensus	0.79 ± 0.22	0.83 ± 0.20	0.72 ± 0.03	0.72 ± 0.10
w/o Warm Start	0.73 ± 0.23	0.77 ± 0.22	0.72 ± 0.02	0.71 ± 0.11
CoR-Fuse	$\mathbf{0.81 \pm 0.09}$	$\mathbf{0.84 \pm 0.10}$	$\mathbf{0.73 \pm 0.03}$	$\mathbf{0.72 \pm 0.09}$

5.3 Ablation Study

To evaluate the contribution of each design component in CoR-Fuse, we conduct ablation studies on TCGA-UVM and TCGA-BLCA using the encoder composition: {UNI, Virchow2, Midnight-12k, GigaPath, Phikon} (Table 3).

Effect of Consensus Representation. Removing the consensus component leaves the overall concentration of slide-level routing weights largely unchanged, with the mean Gini coefficient changing slightly. (Appendix H). Nevertheless, it induces a modest redistribution of routing mass across foundation models: removing consensus decreases the mean weights assigned to Phikon and UNI, while increasing those assigned to Virchow2 and Midnight-12k. Thus, consensus conditioning primarily affects encoder-level weight allocation, while having limited influence on the overall concentration of the routing distribution. Thus, the consensus component primarily affects encoder-level weight allocation rather than the overall concentration of the routing distribution.

Effect of Task-Driven Warm Start. Although final performance was comparable, warm-start initialization substantially improved training reliability. It reduced validation AUC variance, yielding more stable and predictable optimization trajectories. More detailed discussion is provided in Appendix I.

Overall, the ablation results suggest that consensus and warm-start initialization may play different roles: consensus appears to influence routing behavior, whereas warm start can facilitate optimization during training. Their effects on final predictive performance vary across the evaluated compositions and datasets.

6 Conclusion

In this work, we demonstrate that multi-foundation-model fusion performance is not simply improved by increasing the number of integrated models. Fusion effectiveness varies with representation composition, while representation similarity alone is insufficient to characterize functional redundancy or task-level utility. These findings motivate the design of CoR-Fuse: consensus conditioning provides cross-encoder context for handling heterogeneous representation compositions, while sample-specific routing accounts for variation in encoder contributions across slides. Our study has several limitations. First, experiments are conducted on relatively small TCGA cohorts, particularly TCGA-UVM, where performance differences should be interpreted in light of cross-validation variance. Second, although we evaluate CoR-Fuse on two TCGA cohorts, these datasets do not constitute fully independent external validation across institutions or acquisition settings. Future studies will assess the method on larger, multi-institutional cohorts and under domain-shift scenarios. Third, our evaluation focuses on diagnostic and molecular classification tasks, and future work will extend CoR-Fuse to prognosis-oriented endpoints such as survival prediction.

References

Sofiène Boutaj, Leo Fillioux, Maria Vakalopoulou, Stergios Christodoulidis, and Pierre Marza. From patches to patients: A study of the tile-to-slide performance transferability in digital pathology, 2026.

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, 3 2024. doi: 10.1038/s41591-024-02857-3. URL https://www.nature.com/articles/s41591-024-02857-3.

Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. medRxiv, 2023. URL https://www.medrxiv.org/content/early/2023/07/26/2023.07.21.23292757v2.

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction, 2024. URL https://arxiv.org/abs/2409.09173.

Ruitian Gao, Zhaochang Yang, Xin Yuan, Yifan Wang, Yujia Xia, Yufei Zhang, Bingyan Zheng, Yuqiao Gong, Yiping Yue, Yansong Tu, and Zhangsheng Yu. Meta-Encoder: a unified integration framework for multiple pathological foundation models in cancer detection. *Nature Communications*, 17:5738, 2026. doi: 10.1038/s41467-026-71558-x.

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, 2026.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 2018. doi: 10.26302/tmlr.v80i1.2127.

Martine J. Jager, Niels J. Brouwer, and Bita Esmaeli. The Cancer Genome Atlas Project: An Integrated Molecular View of Uveal Melanoma. *Ophthalmology*, 125(8):1139–1142, 2018. doi: 10.1016/j.ophtha.2018.03.011.

Mikhail Karasikov, Joost van Doorn, Nicolas Känzig, Melis Erdal Cesur, Hugo Mark Horlings, Robert Berke, Fei Tang, and Sebastian Otálora. Training state-of-the-art pathology foundation models with orders of magnitude less data. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15967 of *Lecture Notes in Computer Science*, pp. 573–583. Springer Nature Switzerland, 10 2025. doi: 10.1007/978-3-032-04984-1_55.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Tianhao Lei, Parsa Esmaeilkhani, Saghir Alfasly, Wataru Uegami, Judy C. Boughey, Matthew P. Goetz, Krishna Kalari, and H. R. Tizhoosh. Validation of whole-slide foundation models for TCGA data, 2026a.

Wenhui Lei, Yusheng Tan, Anqi Li, Hanyu Chen, Hengrui Tian, Ruiying Li, Zhengqun Jiang, Fang Yan, Xiaofan Zhang, and Shaoting Zhang. Shazam: Unifying multiple foundation models for advanced computational pathology, 2026b.

Tianyu Liu, Ziqing Wang, Zhaokang Liang, Tong Ding, Peter Humphrey, Lorraine Colón-Cartagena, Emily Ling-Lin Pai, Kenneth Tou En Chang, Mohamed Kahila, Jonathan Chong Kai Liew, Tinglin Huang, Rex Ying, Kaize Ding, Faisal Mahmood, and Wengong Jin. Predicting immune biomarkers with multimodal mixture-of-expert pathology foundation models empowers precision oncology, 2026.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 3 2024. doi: 10.1038/s41591-024-02856-4. URL https://www.nature.com/articles/s41591-024-02856-4.

Xiangde Luo, Xiyue Wang, Feyisope Eweje, Xiaoming Zhang, Juan Luis Gomez Marti, Sarah Cascarino, Sen Yang, Yuchen Li, Ryan Quinton, Jinxi Xiang, Yuanfeng Ji, Zhe Li, Yijiang Chen, Colin Bergstrom, Ted Kim, Francesca Maria Olguin, Kelley Yuan, Matthew Abikenari, Andrew

Heider, Sierra Willens, Sanjeeth Rajaram, Robert West, Joel Neal, Adam Schoenfeld, Maximilian Diehn, Chad Vanderbilt, and Ruijiang Li. Ensemble learning of pathology foundation models for precision oncology. *Cancer Cell*, September 2026. ISSN 15356108. doi: 10.1016/j.ccell.2026.08.008. URL https://linkinghub.elsevier.com/retrieve/pii/S1535610826003855.

Vaibhav Mishra and William Lotter. Comparing computational pathology foundation models using representational similarity analysis, 2025.

Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: A family of foundational vision transformers for pathology, 2024. URL https://arxiv.org/abs/2406.05074.

Peiran Quan, Zifan Gu, Zhuo Zhao, Qin Zhou, Donghan M. Yang, Ruichen Rong, Yang Xie, and Guanghua Xiao. GAS-MIL: Group-aggregative selection multi-instance learning for ensemble of foundation models in digital pathology image analysis, 2025.

Alan G. Robertson, Jonathan Shih, Christina Yau, Ewan A. Gibb, Jun Oba, Karen L. Mungall, and et al. Integrative analysis identifies four molecular and clinical subsets in uveal melanoma. *Cancer Cell*, 32(2):204–220.e15, 2017. doi: 10.1016/j.ccell.2017.07.003.

Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0. GitHub repository, 2024. URL https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.

Marin Scalbert, Charlie Saillard, Thomas Peeters, Liam Gonzalez, Dasha Valter, Felipe Llinares-López, Zelda E. Mariet, and Rodolphe Jenatton. H-optimus-1: A foundation model for computational histopathology. AACR Annual Meeting 2026 Late-Breaking Abstract, 2026.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2136–2147. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/10c272d06794d3e5785d5e7c5356e9ff-Paper.pdf.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017.

The Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013. doi: 10.1038/ng.2764.

The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507:315–322, 2014. doi: 10.1038/nature12965.

Xin Tian, Jiuliu Lu, Ephraim Tsalik, Bart Wanders, Colleen Knoth, and Julian Knight. Region-graph optimal transport routing for mixture-of-experts whole-slide image classification, 2026.

Chad Vanderbilt, Gabriele Campanella, Siddharth Singi, Swaraj Nanda, Jie-Fu Chen, Ali Kamali, Amir Momeni Boroujeni, David Kim, Mohamed Yakoub, Jamal Benhamida, Meera Hameed, Neeraj Kumar, and Gregory Goldgof. GOLDMARK: Governed outcome-linked diagnostic model assessment reference kit, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, 2024. doi: 10.1038/s41591-024-03141-0.

Yuxiang Xiao, Yang Hu, Bin Li, Tianyang Zhang, Zexi Li, Huazhu Fu, Jens Rittscher, and Kaixiang Yang. AdaFusion: Prompt-guided inference with adaptive fusion of pathology foundation models, 2025.

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*, 630(8015):181–188, 6 2024. doi: 10.1038/s41586-024-07441-w. URL https://www.nature.com/articles/s41586-024-07441-w.

Zhidong Yang, Xiuhui Shi, Wei Ba, Zhigang Song, Haijing Luan, Taiyuan Hu, Senlin Lin, Jiguang Wang, Shaohua Kevin Zhou, and Rui Yan. Fusecpath: Fusion of multi-scale heterogeneous pathology foundation models for whole slide image analysis, 2025.

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, 2024. URL https://arxiv.org/abs/2408.00738.

A AI USE STATEMENT

During the development of this work, we utilized LLMs for grammar correction and occasional text rewriting to improve clarity. LLMs were not used to generate any scientific content, experimental design, data analysis, or intellectual contributions, nor did their assistance influence any of the reported results.

B ANALYSIS OF FOUNDATION MODELS IN TCGA-UVM

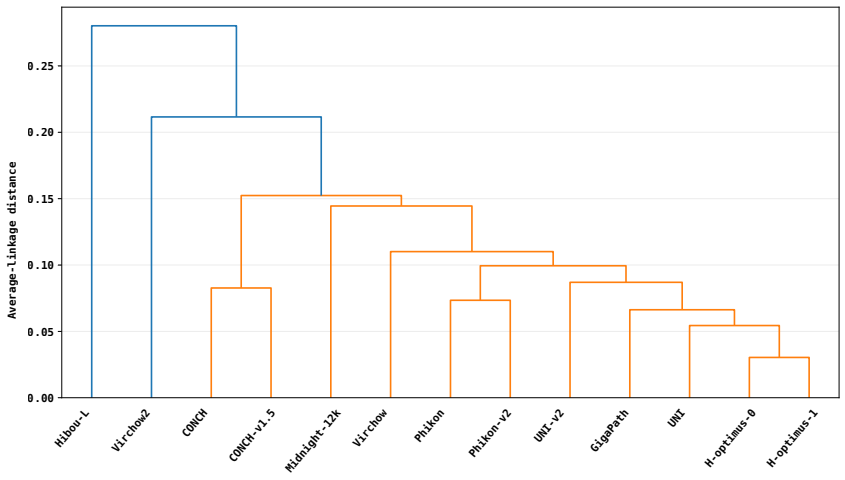

Figure 3: Hierarchical clustering based on CKA similarity across 13 foundation models. The vertical axis represents the average-linkage distance; lower merge distances indicate greater similarity between learned representation spaces. Blue and orange denote two major model clusters.

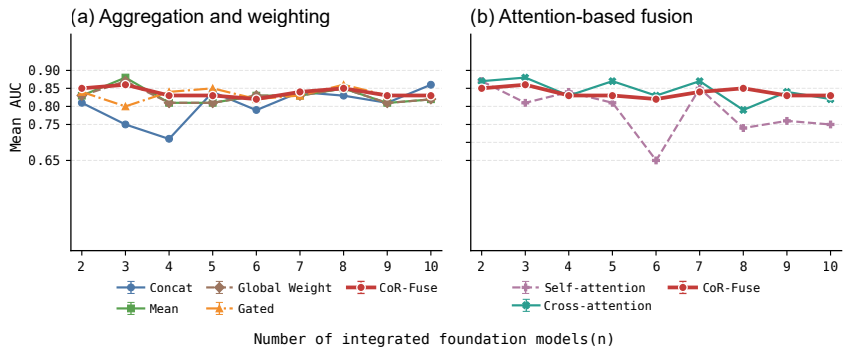

Figure 4: Fusion performance on TCGA-UVM as encoders are added from 2 to 10, ordered by individual validation AUC. Concatenation achieves its highest AUC with 10 integrated foundation models, providing further evidence that fusion performance depends not only on the number of encoders but also on the composition of the representations.

Table 4: Comparison of fusion methods on TCGA-UVM. Results are aggregated across encoder-set sizes $n = 2, \ldots, 10$, with each composition evaluated using 5-fold cross-validation.

Method	AUC	AUPRC	F1	Params(M)	FLOPs(G)
Mean Fusion	0.83 ± 0.02	0.87 ± 0.02	0.74 ± 0.01	4.71	52.35
Concatenation	0.80 ± 0.05	0.85 ± 0.04	0.73 ± 0.03	6.28	69.85
Global Weight	0.83 ± 0.02	0.86 ± 0.02	0.74 ± 0.01	4.71	52.35
Gated Fusion	0.83 ± 0.02	0.87 ± 0.02	0.75 ± 0.02	4.71	52.35
Cross Attention	0.84 ± 0.03	0.89 ± 0.02	0.74 ± 0.03	6.81	105.16
Self Attention	0.79 ± 0.07	0.83 ± 0.05	0.72 ± 0.05	8.38	209.96
CoR-Fuse	0.84 ± 0.01	0.87 ± 0.01	0.75 ± 0.02	4.77	52.35

Table 5: Single foundation model performance on TCGA-UVM D3/M3 classification. The results, ordered by AUC and the results are reported as mean $\pm$ standard deviation over five folds.

Model	AUC	AUPRC	F1
H-optimus-1	$\mathbf{0.84 \pm 0.15}$	0.86 ± 0.14	0.75 ± 0.18
Virchow2	0.83 ± 0.12	0.85 ± 0.14	$\mathbf{0.78 \pm 0.11}$
H-optimus-0	0.83 ± 0.14	$\mathbf{0.86 \pm 0.13}$	0.76 ± 0.14
GigaPath	0.79 ± 0.16	0.83 ± 0.14	0.70 ± 0.17
CONCH	0.78 ± 0.14	0.82 ± 0.15	0.73 ± 0.17
CONCH-v1.5	0.76 ± 0.16	0.78 ± 0.17	0.72 ± 0.19
Phikon	0.76 ± 0.21	0.81 ± 0.18	0.70 ± 0.16
UNI-v2	0.76 ± 0.24	0.81 ± 0.19	0.69 ± 0.18
Virchow	0.74 ± 0.17	0.75 ± 0.18	0.70 ± 0.14
Phikon-v2	0.74 ± 0.19	0.78 ± 0.20	0.62 ± 0.20
Hibou-L	0.73 ± 0.17	0.82 ± 0.13	0.67 ± 0.12
UNI	0.72 ± 0.19	0.75 ± 0.21	0.67 ± 0.27
Midnight-12k	0.59 ± 0.25	0.67 ± 0.20	0.62 ± 0.15

C EXPERIMENTS ON TCGA-UVM

Table 6: Representation similarity and individual-FM predictive performance for three representative five-encoder compositions on TCGA-UVM. Notably, representation similarity does not directly track individual predictive strength. For example, the high-similarity and heterogeneous compositions exhibit substantially different mean pairwise CKA values (0.930 vs 0.777), yet their best individual-FM AUCs remain similar (0.840 vs 0.830).

Encoder composition	Mean pairwise CKA	Mean individual AUC	Best individual AUC
High-similarity	0.930	0.784	0.840
Intermediate-similarity	0.854	0.804	0.840
Heterogeneous	0.777	0.782	0.830

Table 7: Fusion performance of existing methods under {H-optimus-1, H-optimus-0, Virchow2, UNI-v2, CONCH-v1.5} composition on TCGA-UVM.

Method	AUC	AUPRC	F1	Params(M)	FLOPs(G)
Mean Fusion	0.83 ± 0.04	0.87 ± 0.04	0.73 ± 0.06	4.27	47.47
Cross Attention	0.79 ± 0.04	0.84 ± 0.03	0.70 ± 0.05	6.37	94.40
Meta-Encoder	0.60 ± 0.17	0.66 ± 0.10	0.60 ± 0.19	25.76	533.97
FuseCPath	0.70 ± 0.03	0.76 ± 0.02	0.63 ± 0.11	7.42	3.34
CoR-Fuse	$\mathbf{0.84 \pm 0.01}$	$\mathbf{0.87 \pm 0.03}$	$\mathbf{0.77 \pm 0.03}$	4.34	47.47

Table 8: Fusion performance on TCGA-UVM under the fixed encoder composition of {H-optimus-0, Virchow2, UNI, GigaPath, CONCH-v1.5}.

Method	AUC	AUPRC	F1
Mean Fusion	0.81 ± 0.01	$\mathbf{0.84 \pm 0.02}$	0.72 ± 0.02
Concatenation	0.74 ± 0.02	0.77 ± 0.03	0.63 ± 0.11
Global Weight	0.81 ± 0.00	$\mathbf{0.84 \pm 0.02}$	0.72 ± 0.02
Gated Fusion	0.81 ± 0.03	0.82 ± 0.05	$\mathbf{0.74 \pm 0.04}$
Cross Attention	0.79 ± 0.02	0.83 ± 0.02	0.68 ± 0.05
ELF	0.80 ± 0.02	0.83 ± 0.03	$\mathbf{0.74 \pm 0.02}$
CoR-Fuse	$\mathbf{0.82 \pm 0.04}$	0.83 ± 0.02	0.73 ± 0.03

C.1 Experimental Results of Five Heterogeneous Encoders

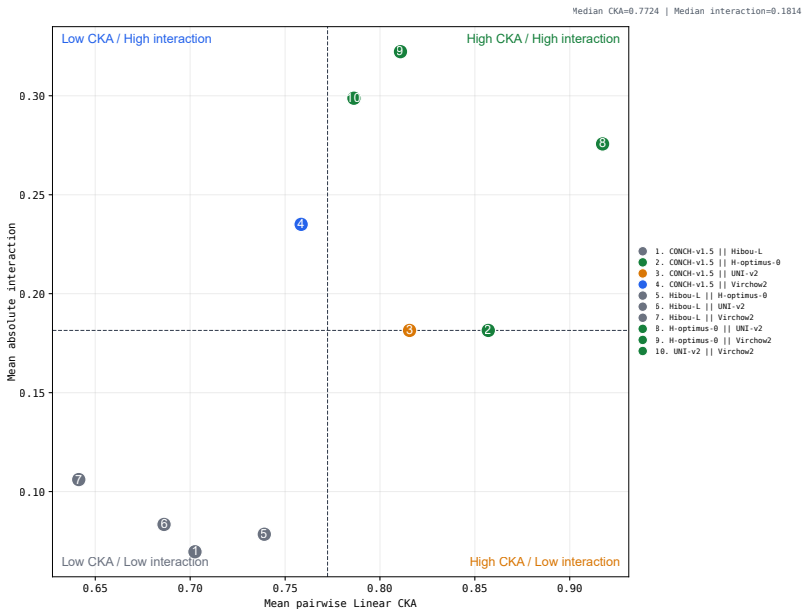

Figure 5: Pairwise CKA versus mean absolute task-level interaction $|I_{ij}|$ for the heterogeneous five-encoder composition on TCGA-UVM. The scatter plot includes encoder pairs with relatively high CKA similarity (> 0.77) and varying interaction magnitudes (> 0.18), suggesting that geometric similarity alone may not fully characterize task-level relationships between encoder representations.

Table 9: Fusion performance under the heterogeneous five-encoder composition {Hibou-L, Virchow2, CONCH-v1.5, UNI-v2, H-optimus-0} on TCGA-UVM. In contrast to the high-similarity composition (Table 10), where attention-based methods perform better, the observed differences suggest that fusion effectiveness may vary across representation compositions.

Method	AUC	AUPRC	F1
Mean Fusion	0.84 ± 0.11	0.88 ± 0.09	0.75 ± 0.15
Concatenation	0.84 ± 0.09	0.87 ± 0.09	0.73 ± 0.09
Global Weight	$\mathbf{0.85 \pm 0.11}$	$\mathbf{0.89 \pm 0.09}$	$\mathbf{0.78 \pm 0.11}$
Cross Attention	0.81 ± 0.13	0.84 ± 0.12	0.70 ± 0.24
Self Attention	0.78 ± 0.13	0.83 ± 0.10	0.72 ± 0.13
Gated Fusion	0.81 ± 0.13	0.85 ± 0.11	0.74 ± 0.16
CoR-Fuse	0.83 ± 0.11	0.88 ± 0.10	0.73 ± 0.15

C.2 Experimental Results of Five High-Similarity Encoders

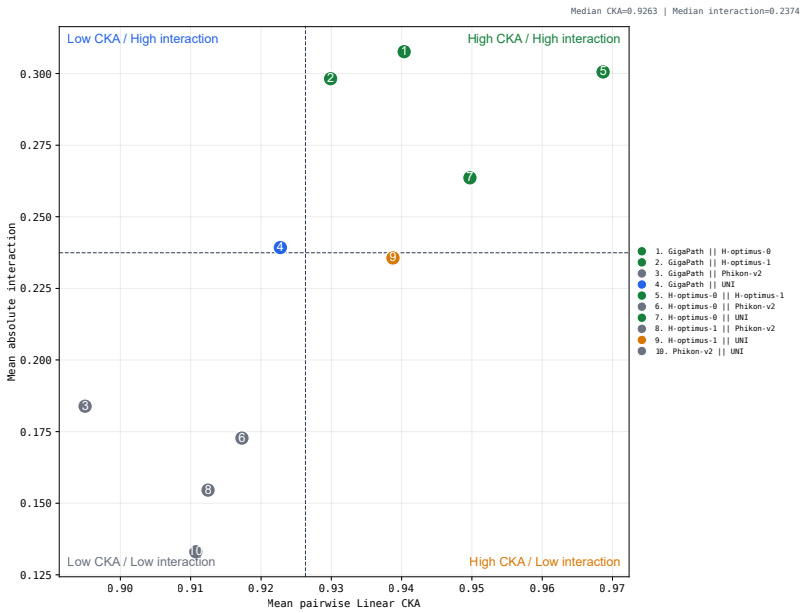

Figure 6: Pairwise CKA versus mean absolute task-level interaction $|I_{ij}|$ for the high-similarity five-encoder composition. The highly similar H-optimus-0 and H-optimus-1 pair exhibits a large interaction magnitude, illustrating that high geometric similarity can coexist with strong task-level interaction. The direction of this interaction is examined separately using signed interaction statistics.

Table 10: Fusion performance under the high-similarity five-encoder composition {H-optimus-0, H-optimus-1, UNI, GigaPath, Phikon-v2}. The ranking of fusion strategies reverses relative to Table 9: attention-based methods perform best in this setting, whereas simpler aggregation methods lag behind. This pattern supports our central observation that no single fusion strategy consistently dominates across encoder compositions.

Method	AUC	AUPRC	F1
Mean Fusion	0.79 ± 0.18	0.82 ± 0.17	0.72 ± 0.18
Concatenation	0.83 ± 0.09	0.88 ± 0.10	0.76 ± 0.13
Global Weight	0.79 ± 0.18	0.82 ± 0.17	0.72 ± 0.18
Cross Attention	$\mathbf{0.86 \pm 0.12}$	0.88 ± 0.11	$\mathbf{0.81 \pm 0.09}$
Self Attention	0.85 ± 0.11	$\mathbf{0.88 \pm 0.08}$	0.76 ± 0.10
Gated Fusion	0.83 ± 0.15	0.86 ± 0.12	0.74 ± 0.16
CoR-Fuse	0.83 ± 0.11	0.87 ± 0.11	0.74 ± 0.17

C.3 Analysis Results of Five Intermediate-Similarity Encoders

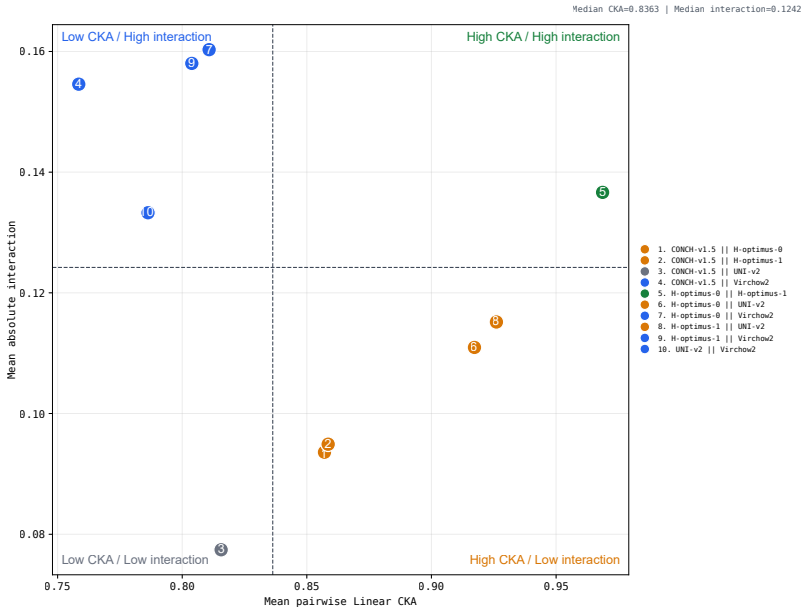

Figure 7: Pairwise CKA versus mean absolute task-level interaction $|I_{ij}|$ for the intermediate-similarity five-encoder composition. The distribution across different CKA-interaction regimes suggests that geometric similarity alone may not fully characterize task-level relationships between pathology foundation encoders.

Table 11: Fusion performance under an intermediate-similarity five-encoder composition {H-optimus-1, H-optimus-0, Virchow2, UNI-v2, CONCH-v1.5} on TCGA-UVM.

Method	AUC	AUPRC	F1
Mean Fusion	0.83 ± 0.04	0.87 ± 0.04	0.73 ± 0.06
Concatenation	0.77 ± 0.05	0.83 ± 0.03	0.69 ± 0.04
Global Weight	0.83 ± 0.04	0.87 ± 0.05	0.73 ± 0.06
Gated Fusion	0.83 ± 0.03	0.86 ± 0.03	0.72 ± 0.05
Cross Attention	0.79 ± 0.04	0.84 ± 0.03	0.70 ± 0.05
Self Attention	0.80 ± 0.08	0.85 ± 0.09	0.70 ± 0.04
CoR-Fuse	$\mathbf{0.84 \pm 0.01}$	$\mathbf{0.87 \pm 0.03}$	$\mathbf{0.77 \pm 0.03}$

D EXPERIMENTS ON TCGA-BLCA

Table 12: Single foundation model performance on TCGA-BLCA TP53 mutation status classification. Results averaged over 5-fold cross-validation.

Model	AUC	AUPRC	F1
Phikon	$\mathbf{0.73 \pm 0.01}$	$\mathbf{0.72 \pm 0.09}$	$\mathbf{0.66 \pm 0.08}$
Virchow2	0.71 ± 0.06	0.70 ± 0.07	0.60 ± 0.09
GigaPath	0.68 ± 0.04	0.65 ± 0.12	0.64 ± 0.11
UNI	0.67 ± 0.04	0.68 ± 0.14	0.61 ± 0.11
Midnight-12k	0.62 ± 0.06	0.63 ± 0.13	0.61 ± 0.13

Table 13: Representation similarity and individual-model performance for {UNI, Virchow2, Midnight-12k, GigaPath, Phikon} composition on TCGA-BLCA.

Mean pairwise CKA	Mean individual AUC	Best individual AUC
0.750	0.682	0.730

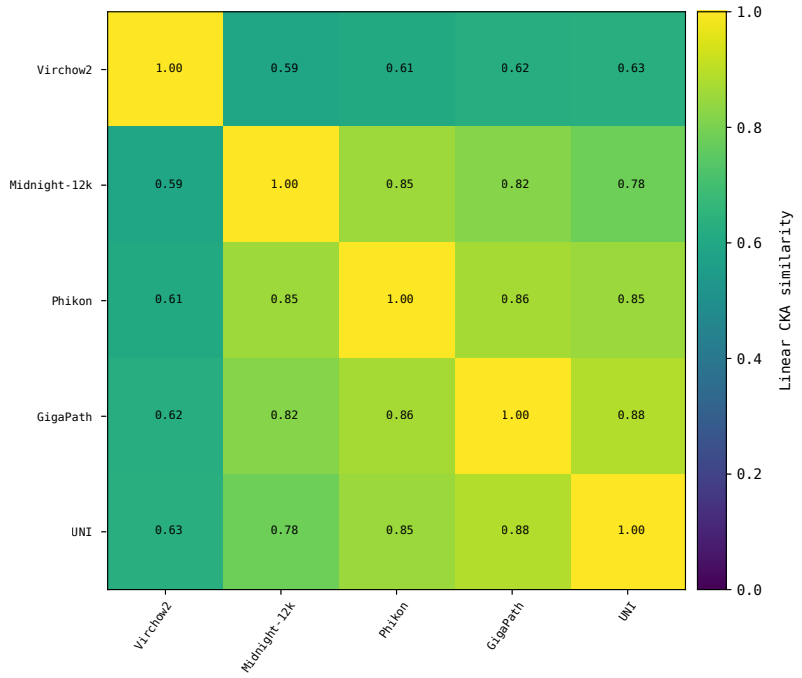

Figure 8: Pairwise CKA similarity matrix for five-encoder composition on TCGA-BLCA. Each entry reports the CKA similarity between a pair of foundation-model representations.

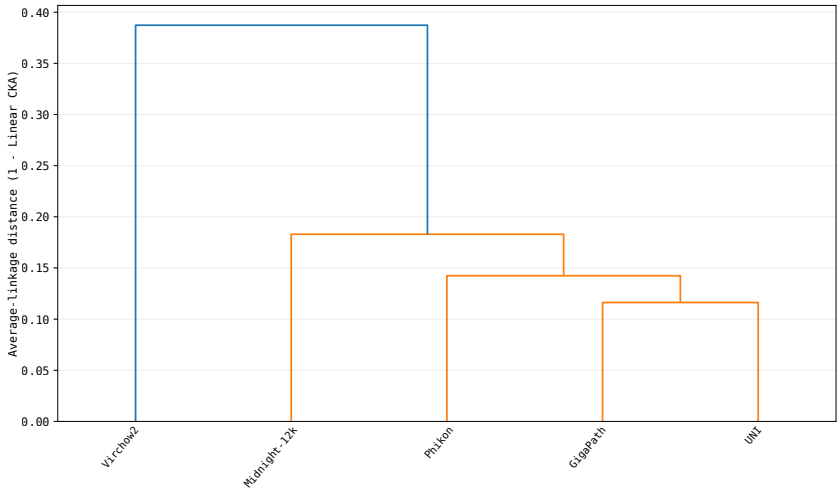

Figure 9: Hierarchical clustering of the five foundation-model representations on TCGA-BLCA based on pairwise CKA similarity. Virchow2 forms a relatively distinct branch, indicating lower representational similarity with the other evaluated encoders.

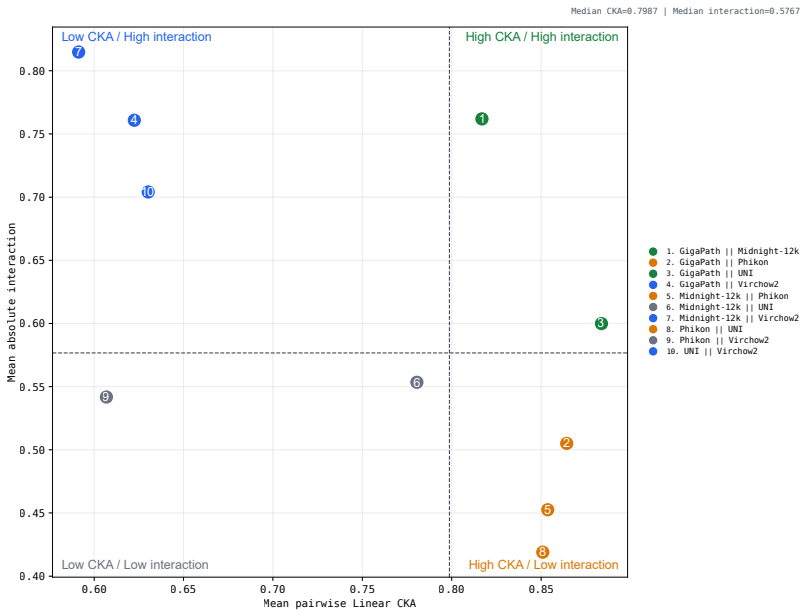

Figure 10: Pairwise CKA versus mean absolute task-level interaction $|I_{ij}|$ for the five-encoder composition on TCGA-BLCA. Each point represents a pair of foundation models, with CKA measuring representational similarity and $|I_{ij}|$ measuring the magnitude of their task-level interaction.

E Implementation Details

All experiments were implemented in PyTorch and run on one NVIDIA GeForce RTX 4090 GPU. We conducted experiments using three random seeds (42, 35, and 2026) with 5-fold cross-validation to assess reproducibility. The warm start stage was optimized for 20 epochs with a learning rate of 1×10^{-4}. The subsequent fusion stage was trained for 80 epochs with the same learning rate. For scoring, we used a lightweight two-layer MLP ($1024 \to 64 \to 1$).

To align heterogeneous foundation-model embeddings, we employ an encoder-specific lightweight projection head for each encoder. Each head consists of a linear projection to a shared 512-dimensional space, followed by LayerNorm and dropout ($p = 0.1$); no additional nonlinear activation is used. The projection heads are independently parameterized across encoders. Linear layers are Xavier-normal initialized with zero biases, while LayerNorm scales and biases are initialized to one and zero, respectively. The reported FLOPs account only for downstream projection, routing, and aggregation operations applied to pre-extracted embeddings, excluding the forward passes of the frozen foundation-model backbones.

For a fair comparison across ablation variants, we keep the total optimization budget consistent across all variants. The full CoR-Fuse model is trained for 20 epochs in Stage 1 and 80 epochs in Stage 2, for a total of 100 epochs. For the w/o Warm Start variant, the model is trained for the same total of 100 epochs from random initialization, with the corresponding routing-stage optimization performed throughout training. Thus, differences between the warm-start and w/o Warm Start variants are attributable to the initialization strategy rather than a difference in the total number of training epochs.

F Experimental Results of Multiple Downstream Heads

Table 14: Performance of different fusion strategies and downstream heads on TCGA-UVM using the representative three-encoder composition {H-optimus-0, Virchow2, CONCH}. CoR-Fuse achieves the highest AUC and AUPRC with ABMIL among the evaluated CoR-Fuse variants, while simple fusion strategies remain competitive with this three-encoder composition.

Method	AUC	AUPRC	F1
Mean Fusion	0.83 ± 0.11	0.86 ± 0.09	0.68 ± 0.16
Concatenation	0.81 ± 0.10	0.87 ± 0.05	$\mathbf{0.75 \pm 0.08}$
Global Weight	0.83 ± 0.11	0.86 ± 0.10	0.69 ± 0.18
Cross Attention	0.78 ± 0.09	0.83 ± 0.09	0.72 ± 0.11
Gated Fusion	0.70 ± 0.27	0.74 ± 0.24	0.73 ± 0.12
CoR-Fuse(ABMIL)	$\mathbf{0.85 \pm 0.08}$	$\mathbf{0.89 \pm 0.09}$	0.74 ± 0.19
CoR-Fuse(TransMIL)	0.80 ± 0.16	0.84 ± 0.13	0.72 ± 0.17
CoR-Fuse(MLP)	0.71 ± 0.20	0.76 ± 0.17	0.61 ± 0.23
CoR-Fuse(GNN)	0.75 ± 0.18	0.76 ± 0.18	0.70 ± 0.16

G Detailed Information of Compared Fusion Methods

G.1 Mean Fusion

Uniform mean fusion assigns an equal weight of $1/M$ to every encoder:

$$\mathbf{Z}_p = \frac{1}{M} \sum_{m=1}^{M} \mathbf{H}_{m,p}. \tag{15}$$

This parameter-free baseline assumes that all encoders contribute equally.

G.2 CONCATENATION

Feature concatenation is performed on the projected latent representations $\mathbf{H}_{M,p}$ into a joint 512 $M-$ dimensional vector prior to downstream bag aggregation:

$$\mathbf{Z}_p = [\mathbf{H}_{1,p}; \mathbf{H}_{2,p}; \ldots; \mathbf{H}_{M,p}] \in \mathbb{R}^{512M}. \tag{16}$$

G.3 GLOBAL WEIGHT FUSION

This method learns a single set of encoder weights shared across all slides and patches:

$$\mathbf{Z}_p = \sum_{m=1}^{M} w_m \mathbf{H}_{m,p}, \qquad \mathbf{w} = \mathrm{softmax}(\mathbf{b}), \quad \mathbf{b} \in \mathbb{R}^M. \tag{17}$$

The encoder-weight logits are optimized jointly with the downstream network. This baseline tests whether a fixed global preference over encoders can improve upon uniform averaging.

G.4 GATED FUSION

Gated fusion produces slide-specific encoder weights from slide-level encoder descriptors. For encoder m, its slide-level descriptor is obtained by mean pooling its patch embeddings:

$$\mu_m^{(n)} = \frac{1}{P_n} \sum_{p=1}^{P_n} \mathbf{H}_{m,p}^{(n)}. \tag{18}$$

A shared linear gate produces an encoder score:

$$g_m = \sigma\left(\mathbf{a}^\top \mu_m + b\right), \tag{19}$$

where $\mathbf{a} \in \mathbb{R}^{512}$ and $b \in \mathbb{R}$ are shared across encoders. The scores are normalized across encoders:

$$w_m = \frac{g_m}{\sum_{j=1}^{M} g_j}. \tag{20}$$

The resulting weights are applied to all patches of the corresponding slide. Unlike global weighting, this formulation allows encoder weights to vary across slides while remaining shared across patches within each slide. The gate computes each encoder score from its own slide-level descriptor without explicitly modeling pairwise relationships among encoders.

H ABLATION STUDY

Table 15: Comparison of routing weights with and without the consensus component. The Gini coefficient measures the concentration of routing weights across encoders, while the per-encoder weights characterize encoder utilization.

Overall Routing Concentration		
Setting	Mean Gini	Normalized Gini
With Consensus	0.2269 ± 0.0752	0.2836
Without Consensus	0.2291 ± 0.1095	0.2863
Per-Encoder Routing Weight		
Foundation Model	With Consensus	Without Consensus
Phikon	0.3194 ± 0.0973	0.3048 ± 0.1303
GigaPath	0.2048 ± 0.0521	0.2101 ± 0.0723
UNI	0.1929 ± 0.0484	0.1768 ± 0.0630
Virchow2	0.1459 ± 0.0783	0.1588 ± 0.0757
Midnight-12k	0.1371 ± 0.0441	0.1494 ± 0.0630

I Understanding the Effect of Mean-Fusion Warm-Start

We further investigate why initializing CoR-Fuse from a mean-fusion model facilitates optimization. During Stage 1, encoder-specific projection heads and the downstream prediction head are optimized using a uniform aggregation strategy.

Let $\theta^{(1)}$ and $\phi^{(1)}$ denote the encoder-specific projection parameters and downstream prediction-head parameters, respectively, after Stage 1. Let $\theta^{(2)}$ and $\phi^{(2)}$ denote the corresponding parameters at the beginning of Stage 2. We initialize Stage 2 by transferring the parameters learned in Stage 1:

$$\theta^{(2)} \leftarrow \theta^{(1)}, \qquad \phi^{(2)} \leftarrow \phi^{(1)}. \tag{21}$$

The uniform aggregation is then replaced by slide-specific routing.

This initialization provides two optimization advantages. First, the projection heads already map heterogeneous encoder features into a task-aligned latent space. Second, the downstream prediction head starts from parameters adapted to a predictive fused representation. Therefore, the routing stage mainly focuses on refining encoder contributions rather than jointly learning representation projection, prediction, and routing from random initialization.

Moreover, mean-fusion initialization provides a permutation-symmetric starting point, where all encoders receive equal initial importance before task-specific evidence is introduced. This reduces the risk of unstable early routing behavior caused by randomly initialized projections and routing scores, which is particularly relevant in the low-sample regime.

J Rationale for Slide-Level Routing

CoR-Fuse performs routing at the slide level rather than the patch level because the objective of the router is to estimate the relative contribution of different foundation models for a given diagnostic task, rather than to identify discriminative tissue regions. First, the available supervision is provided only at the slide level, where each WSI is associated with a single diagnostic label. Patch-level routing would require learning $P \times M$ encoder preferences for individual patches without direct patch-level supervision, potentially introducing additional uncertainty.

Second, our analysis in Section 3 focuses on slide-level encoder relationships and demonstrates that foundation-model contributions vary across slides rather than explicitly across tissue regions. Extending routing to the patch level would therefore require additional evidence for region-dependent encoder specialization.

Third, slide-level routing naturally complements MIL-based downstream architectures. Specifically, CoR-Fuse estimates which foundation models are informative for a slide, while the MIL attention module identifies which patches are relevant for prediction. Patch-level routing would couple encoder weighting with instance selection, potentially introducing unnecessary interaction between the two optimization processes.

Finally, slide-level routing substantially reduces the number of routing decisions. For a slide with P_n patches and M encoders, slide-level routing produces M encoder weights that are shared across patches, whereas patch-level routing requires $P_n \times M$ routing scores. This design therefore provides a simpler and computationally more efficient routing mechanism while remaining compatible with the coordinate-aligned patch representations used for fusion. Patch-level adaptive routing, which could capture region-dependent encoder specialization, remains an interesting direction for future investigation.