Beyond Generic Alignment: Diagnostic Principle–Guided Instance-Concept Alignment for Few-shot Whole Slide Image Classification

Pengyu Guo
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
pengyu.guo@connect.hkust-gz.edu.cn

Liping Wang
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
lwang347@connect.hkust-gz.edu.cn

Jiachuan Wang✉
University of Tsukuba
Tsukuba, Japan
wangjc@slis.tsukuba.ac.jp

Zhao Chen
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
chenzhao@hkust-gz.edu.cn

Chen Cao
The Hong Kong University of Science and Technology
Hong Kong, China
caochen.hkust@gmail.com

Lei Chen
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science and Technology
Hong Kong, China
leichen@cse.ust.hk

Abstract

Few-shot classification of whole-slide images (WSIs) is fundamentally challenging due to extreme data scarcity. Recent approaches enhance supervision by injecting multi-granular textual concept priors. However, they typically rely on generic image–text alignment, treating all instance–concept pairs as equally plausible. Such alignment overlooks the diagnostic principles inherent in these textual priors, leading to spurious matches that misguide instance aggregation: (i) cross-magnification regional coherence: tissue morphology remains semantically consistent across magnification levels; and (ii) microenvironment consistency: leveraging the spatial continuity of lesion regions to filter misleading isolated artifacts. To address this gap, we introduce C^{2}UOT, a novel alignment strategy that reformulates instance–concept alignment as a principle-constrained unbalanced optimal transport (UOT) problem. C^{2}UOT integrates two biologically inspired inductive bias modules: (i) enforcing parent–child semantic consistency across magnifications via a guidance–penalty mechanism, and (ii) applying concept-family–aware spatial smoothing to suppress isolated artifacts while preserving lesion boundaries. Experiments across four datasets demonstrate that C^{2}UOT yields substantial gains by refining the generic alignment, achieving results that exceed SOTA baselines. Code is available at https://github.com/hkustgary/C2UOT.git.

CCS Concepts

• **Computing methodologies** → *Computer vision.*

Keywords

Whole Slide Images, Few-shot Learning

ACM Reference Format:
Pengyu Guo, Liping Wang, Jiachuan Wang, Zhao Chen, Chen Cao, and Lei Chen. 2026. Beyond Generic Alignment: Diagnostic Principle–Guided Instance-Concept Alignment for Few-shot Whole Slide Image Classification. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3767308.3836235

1 Introduction

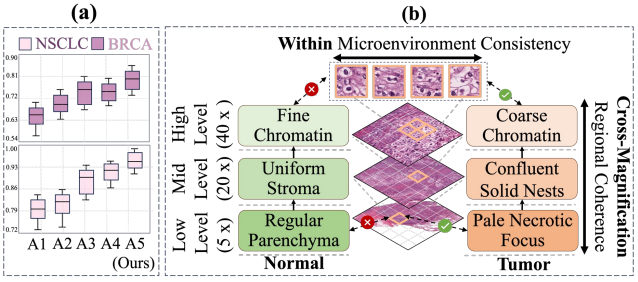


Figure 1: (a) Effect of Alignment Strategy. Generic principle-agnostic VLM-MIL alignment strategies (A1-A4) exhibit sub-optimal performance. The denotations are Cosine Similarity (A1), Optimal Transport (A2), Cross Attention (A3), Unbalanced OT (A4) and our C^{2}UOT (A5). (b) Spurious Alignments Detection via C^{2}UOT.

Digital whole-slide images (WSIs) play a crucial role in cancer diagnosis and treatment planning [6, 32]. Each WSI contains billion-pixels that comprehensively capture diagnostically relevant morphological patterns across multiple magnification scales. Existing studies [12, 17, 26, 27, 34] focus on automating slide-level classification of WSIs by leveraging multiple instance learning (MIL) [21],

This work is licensed under a Creative Commons Attribution 4.0 International License.

MM '26, Rio de Janeiro, Brazil

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3836235

which treats each slide as a bag of instances and aggregates instance features under weak supervision. As data are often scarce due to rare diseases [16], in computational pathology it is essential to formulate the task under more realistic *few-shot settings*.

To mitigate data scarcity in few-shot slide-level WSI classification, recent Vision-Language MIL (VLM-MIL) frameworks [9, 23, 27, 33] commonly leverage LLMs [1, 2, 31] to generate multi-granularity textual concepts as additional textual priors. These priors are then aligned with multi-scale image instances using pretrained Vision-Language models [10, 19] to enhance representation learning. While establishing this multi-scale to multi-granularity alignment is essential for guiding the aggregation of instances, current approaches typically adopt the generic instance–concept alignment of pretrained VLMs and fail to utilize the inherent diagnostic principles within the multi-granularity textual concepts themselves to constrain the alignment process. As a result, these diagnostic principle-agnostic alignments assign substantial mass to semantically incompatible and spurious matches, producing alignments that are locally plausible yet globally inconsistent with diagnostic logic. Our empirical analysis in Figure 1 (a) confirms this vulnerability: when applying various generic alignment strategies (A1-A4) within the ViLa-MIL framework, we observe suboptimal and unstable performance, underscoring the need for diagnostic principle guided alignment mechanisms.

To bridge this gap, we aim to develop a paradigm to align pathological image instances and textual concepts, thereby, utilizes the diagnostic principles inherent in multi-granularity textual concepts to constrain the cross-modal matching process.

Recognizing that pathology diagnostic principles are inherently grounded in spatial and hierarchical logic within textual concepts, our paradigm mathematically translates these biological priors into structural consistency constraints. Specifically, we enforce this consistency in the *image instance-textual concept* alignment along two complementary dimensions: **(i) *Cross-Magnification Regional Coherence***: Pathologists transition from coarse, low-magnification assessments (e.g., tissue architecture) to fine-grained morphology (e.g., nuclear atypia) at higher levels, iteratively checking that interpretations remain semantically coherent within the same regional location [22]. As shown in Figure 1 (b), if *pale necrotic focus* is observed at low magnification, the subsequent high-magnification analysis is semantically constrained to this specific region. Pathologists will then search for fine-grained pathology concepts consistent with this finding in the coarse high-magnification, such as *coarse chromatin* (Orange Box), while disregarding concepts from irrelevant regions (Green Box). **(ii) *Microenvironment Consistency:*** Lesion regions exhibit spatial continuity and similarity, meaning instances of the same subtype typically form cohesive clusters with crucial boundaries [14]. This allows pathologists to filter out misleading isolated patches (e.g., artifacts or inflammation), which often represent non-lesion entities. As shown in Figure 1 (b), neighboring high-mag instances (at the top of the figure) within the same tumor microenvironment should be semantically consistent.

However, operationalizing these principles within a computational framework presents distinct technical challenges:

(i) The primary challenge is mathematically translating the inherent diagnostic principles of multi-granularity textual concepts into a principle guided alignment objective, encouraging the cross-modal alignment to respect this semantic coherence across magnifications.
(ii) Second, while multi-granularity textual concepts provide strong microenvironment consistency to filter artifacts, enforcing this spatial consistency risks misidentifying *genuine but spatially isolated* pathological signals (e.g., micrometastases) as noise. The challenge lies in designing a selective constraint that suppresses actual artifacts while rigorously protecting these rare yet diagnostically paramount signals.

To address these challenges, we propose C^{2}UOT (**C**onsistency and **C**oherence aware **U**nbalanced **O**ptimal **T**ransport), a novel alignment paradigm that reformulates instance-concept alignment as a principle-constrained Unbalanced Optimal Transport (UOT) problem. C^{2}UOT treats image instances and textual concepts as discrete distributions, refining the generic VLM-based alignment through two biologically-inspired inductive bias modules: **(i) Cross-scale Coherence (CSC) module**. The CSC module explicitly preserves semantic consistency between instances across multiple magnification levels. This module penalizes assignments in the alignments that violate predefined cross-scale hierarchy through two steps: *guidance* - construct a soft target by leveraging coupled hierarchical parent–child dependencies and *penalty* - penalize hierarchical violations. **(ii) Microenvironment Semantic Consistency (MSC) module**. The MSC module imposes a concept family-aware spatial smoothing strategy to suppress isolated signals. This module discourages inconsistent semantic alignments among spatially adjacent instances while preserving *genuine but spatially isolated* pathological signals that are diagnostically crucial. Together, the CSC and MSC modules enhance instance–concept alignment under microenvironmental and cross-scale diagnostic constraints, providing a principle guidance signal for instance aggregation.

The contributions of this paper are summarized as follows:

- We introduce C^2UOT, a novel alignment paradigm that utilizes diagnostic principles inherent in multi-granularity textual concepts to explicitly constrain the cross-modal alignment process. By reformulating alignment as a principle-constrained Unbalanced Optimal Transport (UOT) problem, we provide a rigorous mathematical foundation to resolve the spurious matching bottleneck in few-shot WSI analysis.
- We operationalize complex diagnostic principles by integrating two novel inductive bias modules, CSC and MSC, into the UOT solver. These modules effectively integrate hierarchical cross-scale coherence and spatial microenvironment consistency into the UOT framework via a guidance-penalty mechanism and family-aware spatial smoothing.
- Extensive experiments on four benchmarks demonstrate that C^{2}UOT-MIL delivers superior performance, while additionally providing interpretable and topology-aware attention maps that remain coherent across scales.

2 Related Work

2.1 MIL in Pathology

Given the prohibitive cost of pixel-level annotations [4], WSI analysis widely adopts multiple instance learning (MIL), where each

slide is modeled as a bag of non-overlapping patches (instances) and only slide-level labels are available during training [20]. MIL methods [17, 20, 26, 34] typically comprise two stages: *instance encoding*, which extracts feature representations for each patch, and *instance aggregation*, which combines these representations into a slide-level embedding. Recent studies have begun exploring WSI-specific priors to enhance MIL. For instance, HiDisc [13] utilizes the inherent WSI hierarchy for representation learning, while PANTHER [28] employs Optimal Transport to aggregate morphological prototypes. However, these methods operate exclusively within the visual domain. In few-shot scenarios, relying solely on limited visual data provides insufficient supervision and suffers from severe performance degradation under extreme data scarcity.

2.2 VLM-MIL in Pathology

Recent efforts address the data scarcity in few-shot WSI classification by shifting towards Vision-Language MIL (VLM-MIL) frameworks [8, 9, 25, 27]. Building on vision–language models (VLMs) such as PLIP [10], QuiltNet [11], and CONCH [19], this line of work utilizes LLM-generated multi-granular textual concepts as explicit semantic priors to guide instance aggregation. For example, ViLa-MIL [27] employs a dual-scale prompt design to aggregate coarse- and fine-grained slide features with descriptive texts, and MSCPT [9] performs graph-based prompt tuning to capture inter-instance context prior to slide-level prediction. FOCUS [8] leverages semantic priors to filter out massive amounts of irrelevant information, enabling the model to focus on the key features.

Despite these advances, most methods still treat patch–concept alignments independently, without jointly modeling their diagnostic dependencies across scales. HiVE-MIL [33] and our method share similar motivation of overcoming this limitation, but operationalize it differently. HiVE-MIL substantially advances hierarchical vision–language modeling by embedding cross-scale dependencies into graph connectivity, with TGDF dynamically selecting patch–text edges through thresholding. In contrast, our CSC and MSC terms regularize a single global UOT coupling over all patches and all textual concepts. Unlike HiVE-MIL, this formulation induces global competition among all patch-concept pairs and enforces diagnostic consistency through soft penalties, preserving potentially informative matches that threshold-based filtering may discard.

3 Preliminaries and Background

This work tackles few-shot slide-level WSI classification under weak (slide-level) supervision. Given a gigapixel WSI X and its corresponding slide-level label $Y \in \{0, 1\}$, the most fundamental WSI classification objective is to train a parameterized model f_θ that minimizes the expected cross-entropy (CE) loss:

$$\min_\theta \mathbb{E}_{(X,Y)}\left[\mathrm{CE}(f_\theta(X), Y)\right] \tag{1}$$

Specifically, a gigapixel WSI X is typically digitized as a multi-resolution image pyramid. We denote the set of available magnification levels as $\mathcal{L} = \{l_1, l_2, \ldots, l_L\}$, where lower levels capture coarser macroscopic patterns and higher levels represent finer microscopic details. However, due to the prohibitive resolution of X, direct optimization is computationally intractable. Standard MIL paradigms [17, 18, 20, 26, 34] alleviate this by partitioning X into a bag of patch instances I, as defined below:

Definition 3.1. Instance Features. For each magnification level ℓ, the WSI is partitioned into non-overlapping instances. A pre-trained image encoder $E_I(\cdot)$ extracts instance embeddings $\mathcal{I}^{(\ell)} = \{E_I(x_{(i)}^{(\ell)}) : x_{(i)}^{(\ell)} \in X^{(\ell)}\} = \{\mathbf{z}_{(i)}^{(\ell)}\}_{i=1}^{N_\ell}$, where N_ℓ is the number of instances at level ℓ. The complete set of instances across all levels is denoted as $\mathcal{I} = \{\mathcal{I}^{(\ell)}\}_{\ell=1}^{L}$.

We mathematically formulate the spatial and hierarchical dependencies among these instances as an *Instance Graph*, denoted as $\mathcal{G}_i = (\mathcal{V}_I, \mathcal{E}_{sp}, \Gamma_{sc})$. Here, $\mathcal{V}_I$ is the set of all instance embeddings I. The within-level undirected edges $\mathcal{E}_{sp}$ connect spatially grid-adjacent patches, capturing the local tissue microenvironment. The cross-level directed edges Γ_{sc} capture pyramidal containment, linking a parent patch at a lower magnification to all spatially enclosed child patches at a higher magnification.

Consequently, compared with the slide-level loss in Eq. (1), the generic MIL objective is reformulated to optimize the aggregation of these instances via a function Φ:

$$\min_\theta \mathbb{E}_{(X,Y)}\left[\mathrm{CE}(f_\theta(\Phi(I)), Y)\right]. \tag{2}$$

However, in the few-shot regime, optimizing Eq. (2) is particularly challenging due to the limited training data, therefore additional supervision is necessary. To solve this problem, recent VLM-MIL approaches [27, 33] introduce multi-granular textual concepts U to enrich instance representations and guide the aggregation process.

Definition 3.2. Concept Features. Each magnification level ℓ is associated with a set of morphological concepts $C^{(\ell)}$. A text encoder $E_T(\cdot)$ maps each concept $c_{(j)}^{(\ell)}$ into a semantic embedding $\mathcal{J}^{(\ell)} = \{E_T(c_{(i)}^{(\ell)}) : c_{(i)}^{(\ell)} \in C^{(\ell)}\} = \{\mathbf{u}_{(i)}^{(\ell)}\}_{i=1}^{M_\ell}$, where M_ℓ is the number of concepts at level ℓ. The complete set of concepts is denoted as $U = \{\mathbf{u}_{(j)}\}_{j=1}^{M}$.

Similar to the visual domain, we structure these textual concepts into a *Concept Graph*, denoted as $\mathcal{G}_c = (\mathcal{V}_c, \mathcal{E}_c, \Gamma_c)$. Here, $\mathcal{V}_c$ contains all semantic embeddings. The cross-level directed edges Γ_c map established diagnostic dependencies, linking a coarse architectural parent concept to its specific fine-grained cytological descendants. The within-level undirected edges $\mathcal{E}_c$ connect sibling concepts that share the exact same parent node, defining a cohesive semantic *concept family*.

4 Method

As established in Section 3, the efficacy of VLM-MIL methods relies heavily on instance-concept alignment. However, current generic alignment strategies treat instance-concept pairs independently, leading to diagnostically implausible spurious alignments. To address this bottleneck, we formulate the instance-concept alignment as a principle-constrained Unbalanced Optimal Transport problem (Sec. 4.1). Specifically, we introduce C^{2}UOT, which mathematically encodes diagnostic principles into the alignment process via two inductive bias modules: the Cross-scale Coherence (CSC) module (Sec. 4.2) and the Microenvironment Semantic Consistency (MSC) module (Sec. 4.3). An overview of C^{2}UOT is illustrated in Figure 2.

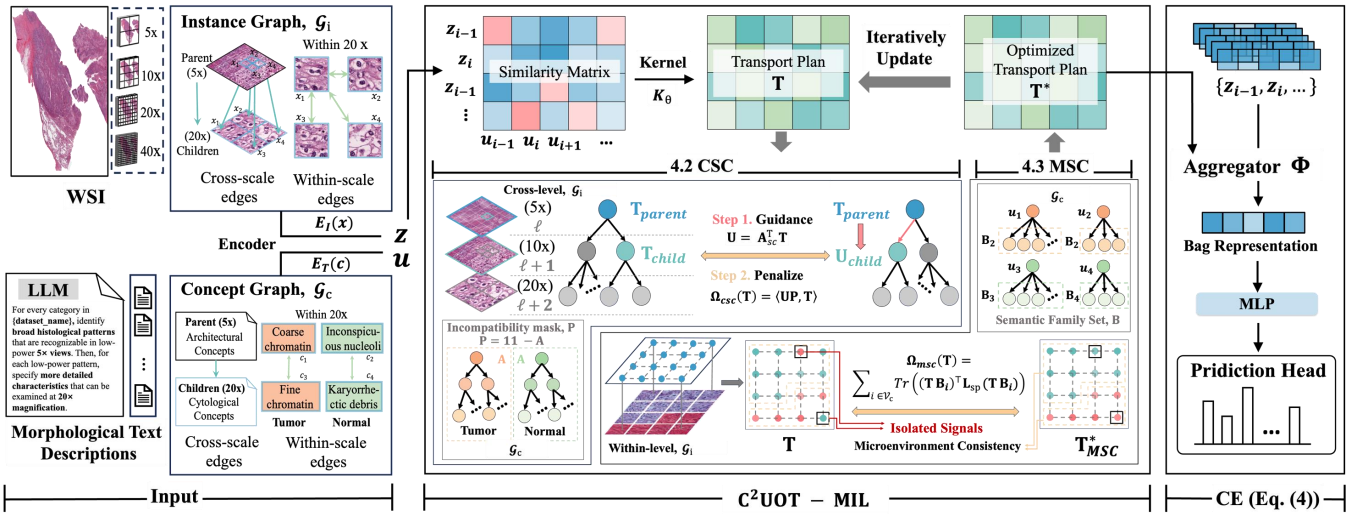


Figure 2: Framework overview of C²UOT-MIL

4.1 Problem Formulation

To incorporate additional supervision (Concept Graph $\mathcal{G}_c$), existing VLM-MIL methods [9, 23, 27, 33] guide the instance aggregation by computing a generic cross-modal alignment matrix $\hat{\mathbf{T}} \in \mathbb{R}_+^{N\times M}$. Each entry $\hat{\mathbf{T}}_{i,j}$ represents the pairwise similarity between an instance z_i and a textual concept u_j, derived independently in the shared VLM latent space. The objective is formulated as:

$$\min_{\theta} \mathbb{E}_{(X,Y)}[\mathrm{CE}(f_\theta(\Phi(I,\hat{\mathbf{T}})), Y)], \tag{3}$$

where Φ aggregates instance features into slide-level representations: $\mathbf{s}_{\text{slide}} = (\hat{\mathbf{T}}\mathbf{W}_\phi(\mathbf{1}^\top(\hat{\mathbf{T}}\mathbf{W}_\phi))^{-1})^\top\mathbf{z}$. $\mathbf{W}_\phi \in \mathbb{R}_+^{M\times 1}$ is a trainable concept-importance weight vector that projects the multi-granularity textual alignments into scalar instance-level attention scores, and $\mathbf{z} \in \mathbb{R}^{N\times d}$ is the stacked instance feature matrix.

The fundamental limitation of current approaches lies in the diagnostic principle-agnostic nature of alignment (mathematically, the construction of $\hat{\mathbf{T}}$). While current methods compute pairwise similarity independently, this alignment ignores the diagnostic structures $(\Gamma_c, \mathcal{E}_c)$ inherent in the multi-granular concept graph. Consequently, $\hat{\mathbf{T}}$ permits a high volume of spurious matches that are locally plausible in the VLM embedding space but globally violate established diagnostic logic.

To resolve this bottleneck, our objective is to reformulate the instance-concept alignment from a naive similarity computation (i.e., $\hat{\mathbf{T}}$) into a principle-constrained Unbalanced Optimal Transport (UOT) problem between the instance distribution and the concept distribution. We aim to replace the principle-agnostic $\hat{\mathbf{T}}$ with an optimal transport plan $\mathbf{T}^* \in \mathbb{R}_+^{N\times M}$. This transforms the overall WSI classification into a bi-level optimization framework:

$$\min_{\theta} \mathbb{E}_{(X,Y)}\left[\mathrm{CE}\left(f_\theta\left(\Phi\left(\mathcal{I}, \mathbf{T}^*\right)\right), Y\right)\right], \tag{4}$$

$$\mathbf{T}^* = \arg\min\left[\langle \mathbf{T}, \mathbf{D}\rangle + \tau\mathcal{D}_{KL}(\mathbf{T}) + \lambda\Omega(\mathbf{T}, \mathcal{G}_i, \mathcal{G}_c)\right], \tag{5}$$

where $\mathbf{D} = -log K_\Theta$ denotes the pairwise cost matrix between visual instances and textual concepts, computed using the kernel $K_\Theta(r,q) = \exp(-\frac{1}{\sigma^2}\left\|W_z\,\mathbf{z_r} - W_u\,\mathbf{u}_q\right\|_2^2)$. W_z and W_u are projection heads of instances and concepts, respectively. σ is a learnable bandwidth that controls the sharpness of the cost matrix. Larger values of $\mathbf{D}_{(r,q)}$ impose stronger penalties on matching instance $\mathbf{z}_r$ with concept $\mathbf{u}_q$. $\mathcal{D}_{KL}(\mathbf{T}) = \mathrm{KL}(\mathbf{T}\mathbf{1}_M||\mathbf{a}) + \mathrm{KL}(\mathbf{T}^\top\mathbf{1}_N||\mathbf{b})$ serves as the marginal relaxation term, where $\mathbf{a}$ and $\mathbf{b}$ represent the marginal distributions of the visual instances and textual concepts, respectively. Unlike standard OT, which strictly enforces mass conservation and forces all visual instances to align with all concepts, this relaxation term parameterized by τ allows the UOT solver to naturally discard pathologically irrelevant instances or unmatchable concepts.

The inner optimization in Eq. (5) seeks an optimal transport plan $\mathbf{T}^*$ that minimizes the cross-modal transport cost of the original transport plan $\mathbf{T}$ while being constrained by inductive bias terms $\Omega(\cdot)$, which ensure that $\mathbf{T}^*$ adheres to diagnostic principles encoded in the concept graph.

Specifically, to operationalize these principles, we decouple $\Omega(\cdot)$ into two complementary modules:

$$\Omega(\mathbf{T}, \mathcal{G}_i, \mathcal{G}_c) = \Omega_{csc}(\mathbf{T}, \mathcal{G}_i, \mathcal{G}_c) + \Omega_{msc}(\mathbf{T}, \mathcal{G}_i, \mathcal{G}_c). \tag{6}$$

The first inductive bias term Ω_{csc} enforces cross-scale coherence, and the second term Ω_{msc} preserves microenvironment semantic consistency. Details of these two modules are described in the following sections.

4.2 Cross-scale Coherence (CSC) Module

Recall that when pathologists examine a WSI, they inherently follow a top-down diagnostic principle: microscopic details must align with the macroscopic context. For example, if a broad region is identified as a *pale necrotic focus* at a low magnification (e.g., 5x), the high-magnification patches (e.g., 20x) within this specific region should not be confidently matched to *healthy stroma* concepts. The CSC module is designed to encode this diagnostic principle inherent in multi-granularity textual concepts into the alignment process. We treat the alignment results from the low-magnification parent patches as a *soft target* and propagate them down to their child patches. Any transport assignments that attempt to violate this established diagnostic principle are penalized.

Specifically, CSC enforces the diagnostic principle guided alignment through a two-step penalty mechanism on the transport plan: (1) *Guidance.* First, it propagates the parent's (low magnification) diagnostic information to create a *soft target* $\mathbf{U}$. (2) *Penalty.* Then, this *soft target* $\mathbf{U}$ is used in conjunction with a graph-based *incompatibility mask* to compute the final penalty Ω_{csc}, which penalizes any assignments in $\mathbf{T}$ that violate this cross-scale hierarchy.

Step 1: Construct the Soft Target (Guidance). The transport plan $\mathbf{T}_{(p)}$ of a parent instance is propagated to its children $\mathbf{z}_{(s)}$, to serve as a diagnostic prior. Building on a common assumption in hierarchical Bayesian [15] modeling that parent nodes provide priors for their descendants, the *soft target* for child instances is computed as:

$$\mathbf{U} \;=\mathbf{A}_{\text{sc}}^{\top}\mathbf{T} \;\in\; \mathbb{R}_{+}^{N\times M}, \tag{7}$$

where $\mathbf{A}_{\text{sc}} \in \{0,1\}^{N\times N}$ is an adjacency matrix induced by Γ_{sc} in $\mathcal{G}_i$. Each row of $\mathbf{U}$ is the result of this propagation: $\mathbf{U}_{(s,:)} = \sum_p (\mathbf{A}_{\text{sc}})_{(p,s)} \cdot \mathbf{T}_{(p,:)}$. This operation encodes the expected concept distribution (the soft target) for a child instance $\mathbf{z}_{(s)}$ by summing the transport plan of all its parent instances $\mathbf{z}_{(p)}$, where $(\mathbf{A}_{\text{sc}})_{(p,s)} = 1$.

Step 2: Penalize Hierarchical Violations (Penalty). This step penalizes semantic inconsistencies between parent and child instances by integrating the soft target $\mathbf{U}$ with the predefined concept hierarchy structure $\mathcal{G}_c$. To achieve this, we introduce a boolean incompatibility mask $\mathbf{P}$, which mathematically represents concept pairs that are not in a hierarchical parent–child relationship.

Intuitively, since $\mathbf{U}$ encodes the expected concept distribution propagated from a parent to its children, the matrix product $\mathbf{UP}$ highlights concept alignments that are incompatible in terms of the hierarchical structure Γ_c. The **CSC** module penalizes cases where the transport plan $\mathbf{T}$ aligns with these incompatible alignments:

$$\Omega_{\text{csc}}(\mathbf{T}, \mathcal{G}_i, \mathcal{G}_c) \;=\; \langle \mathbf{UP},\; \mathbf{T}\rangle_F \;=\; \sum_{i=1}^{N}\sum_{i'=1}^{M} (\mathbf{UP})_{(i,i')}\, \mathbf{T}_{(i,i')}. \tag{8}$$

Here, $\mathbf{P} \;=\; \mathbf{1}\mathbf{1}^{\top} - \mathbf{R}$, where $\mathbf{R} \in \{0,1\}^{M\times M}$ is the reachability matrix derived from the edge set Γ_c. A penalty is incurred when a child instance $\mathbf{z}_{(i)}$ is assigned mass to a hierarchically incompatible concept $\mathbf{u}_{(i')}$ (i.e., $\mathbf{T}_{(i,i')} > 0$).

Formally, this mechanism can be formalized by the following proposition, which proves that Ω_{csc} suppresses alignments that violate the cross-scale coherence.

Proposition 4.1. *The **CSC** module Ω_{csc} is a non-negative penalty term that is minimized if and only if for every child instance $\mathbf{z}_s$ and its parent $\mathbf{z}_p$, the transport assignment $\mathbf{T}_{(s,q')}$ is non-zero only for concepts $\mathbf{u}_{q'}$ that are descendants of the concepts $\mathbf{u}_q$ to which the parent has non-zero assignment $\mathbf{T}_{(r,q)}$.*

4.3 Microenvironment Semantic Consistency (MSC) Module

Tumor microenvironments are usually spatially continuous, so an isolated tumor patch within healthy tissue is more likely to be an artifact than a true lesion. The MSC module exploits this continuity to filter out such misleading isolated signals. However, applying a naive global spatial smoothing would be catastrophic, as it would blur sharp tissue boundaries and erase *genuine but spatially isolated* pathological signals, such as crucial micrometastases. To solve this problem, MSC introduces a concept-family-aware smoothing strategy. We strictly confine the spatial smoothing process to occur only within the same semantic concept family, thereby eliminating random high-frequency noise while preserving the sharp boundaries between distinct tissue types.

Mathematically, we operationalize this family-wise smoothness on the transport plan $\mathbf{T}$ by computing the Dirichlet energy for each concept family independently. For each parent concept $\mathbf{u}_{(i)}$, we define a diagonal mask $\mathbf{B}_{(i)} \in \{0,1\}^{M\times M}$, which identifies the subset of concepts in $\mathbf{T}$ corresponding to a single semantic concept family, comprising all child concepts that share $\mathbf{u}_{(i)}$ as their parent node. This inductive bias term is then defined as the sum of these independent family energies:

$$\Omega_{\text{msc}}(\mathbf{T}) \;=\; \sum_{i\in\mathcal{V}_c} \text{Tr}\Big((\mathbf{T}\mathbf{B}_{(i)})^{\top}\mathbf{L}_{\text{sp}}(\mathbf{T}\mathbf{B}_{(i)})\Big) \tag{9}$$

where $\mathbf{L}_{\text{sp}} := \text{diag}(\mathbf{A}_{\text{sp}}\mathbf{1}) - \mathbf{A}_{\text{sp}}$ is the combinatorial Laplacian [5] induced by $\mathcal{E}_{sp}$. The effectiveness of Ω_{msc} is guaranteed by the following proposition:

Proposition 4.2. *Minimizing Ω_{msc} is mathematically equivalent to applying a low-pass denoising filter to each concept family. The filter $H_\eta = (I + \eta L_{sp})^{-1}$, with spectral response $g(\lambda_i) = (1+\eta\lambda_i)^{-1}$, is able to suppress spatially isolated high-frequency noise. Therefore, minimizing Ω_{msc} effectively suppresses spurious alignments.*

4.4 Solver and Complexity Analysis

The formulation in Eq. (5) can be efficiently solved using the scaling-iteration Sinkhorn method [3]. The complexity of solving Eq. (5) consists of kernel construction and Sinkhorn iterations. Kernel construction computes the cost matrix $\mathbf{D}$, which costs $O(NMd)$. Each Sinkhorn iteration incorporates two inductive bias modules yields a per-iteration cost of $O(NM^2)$ for L_{sink} iterations, leading to a total complexity of $O(NMd + L_{\text{sink}}NM^2)$. As d and L_{sink} are constants, the effective complexity scales linearly to $O(NM^2)$. In practice, M is usually small ($< 10^2$), so the computation remains manageable in realistic applications.

5 Experiment

We conduct experiments to investigate the proposed framework and answer the following research questions.

- **RQ1:** How does the overall performance of our framework compare against SOTA baselines?
- **RQ2:** How effectively do the proposed modules mitigate spurious alignments and improve the final performance?
- **RQ3:** How robust is the proposed framework and what is its practical computational cost?

5.1 Experimental Setup

Datasets. We evaluate the proposed framework on four publicly available datasets [7, 30]: TCGA-NSCLC, TCGA-BRCA, TCGA-RCC and CAMELYON, covering slide-level subtyping and survival risk classification tasks in histology. Following the experimental protocol of current works [8, 9, 27], we evaluate our method under K-shot settings, where $K \in \{4, 8, 16\}$. For each K-shot configuration, we sample exactly K WSIs per class. Results are reported as the mean ± standard deviation of ACC, F1, and AUC over five-fold cross-validation. For fair comparison, the exact same few-shot splits are utilized across all methods.

Table 1: Few-shot WSI classification under 16 -shot setting across 3 datasets.

Mod.	Method	TCGA-NSCLC AUC	TCGA-NSCLC ACC	TCGA-NSCLC F1	TCGA-BRCA AUC	TCGA-BRCA ACC	TCGA-BRCA F1	TCGA-RCC AUC	TCGA-RCC ACC	TCGA-RCC F1
V	Mean Pooling	$0.8648_{\pm 0.037}$	$0.7877_{\pm 0.039}$	$0.7868_{\pm 0.043}$	$0.7556_{\pm 0.045}$	$0.6414_{\pm 0.043}$	$0.6165_{\pm 0.043}$	$0.9617_{\pm 0.026}$	$0.8678_{\pm 0.059}$	$0.8515_{\pm 0.033}$
	ABMIL [12]	$0.9094_{\pm 0.017}$	$0.8601_{\pm 0.021}$	$0.8579_{\pm 0.014}$	$0.7525_{\pm 0.012}$	$0.6917_{\pm 0.031}$	$0.6681_{\pm 0.021}$	$0.9671_{\pm 0.019}$	$0.8819_{\pm 0.021}$	$0.8621_{\pm 0.024}$
	DSMIL [17]	$0.9219_{\pm 0.027}$	$0.8269_{\pm 0.034}$	$0.8266_{\pm 0.031}$	$0.7325_{\pm 0.035}$	$0.6952_{\pm 0.035}$	$0.6339_{\pm 0.031}$	$0.9745_{\pm 0.024}$	$0.9091_{\pm 0.025}$	$0.8957_{\pm 0.022}$
	CLAM [20]	$0.9428_{\pm 0.023}$	$0.8742_{\pm 0.025}$	$\underline{0.8817}_{\pm 0.022}$	$0.7651_{\pm 0.015}$	$0.6963_{\pm 0.012}$	$0.6752_{\pm 0.016}$	$0.9754_{\pm 0.027}$	$0.9127_{\pm 0.047}$	$\underline{0.9029}_{\pm 0.028}$
	Trans-MIL [26]	$0.9318_{\pm 0.019}$	$0.8635_{\pm 0.012}$	$0.8671_{\pm 0.013}$	$0.7561_{\pm 0.021}$	$0.6869_{\pm 0.025}$	$0.6621_{\pm 0.022}$	$0.9775_{\pm 0.023}$	$\underline{0.9169}_{\pm 0.029}$	$0.8956_{\pm 0.017}$
	DTFD-MIL [34]	$0.9524_{\pm 0.029}$	$\underline{0.8743}_{\pm 0.026}$	$0.8703_{\pm 0.022}$	$0.7727_{\pm 0.032}$	$0.6986_{\pm 0.023}$	$0.6709_{\pm 0.024}$	$0.9779_{\pm 0.021}$	$0.9024_{\pm 0.014}$	$0.8876_{\pm 0.011}$
	WiKG [18]	$0.9165_{\pm 0.041}$	$0.8394_{\pm 0.038}$	$0.8375_{\pm 0.038}$	$0.7491_{\pm 0.039}$	$0.6607_{\pm 0.027}$	$0.6291_{\pm 0.024}$	$\underline{0.9815}_{\pm 0.025}$	$0.8991_{\pm 0.025}$	$0.8716_{\pm 0.031}$
	RRT-MIL [29]	$0.9379_{\pm 0.013}$	$0.8697_{\pm 0.016}$	$0.8719_{\pm 0.018}$	$0.7672_{\pm 0.029}$	$0.7041_{\pm 0.031}$	$0.6782_{\pm 0.032}$	$0.9793_{\pm 0.035}$	$0.8426_{\pm 0.039}$	$0.8261_{\pm 0.041}$
VL-S	FOCUS [8]	$\underline{0.9526}_{\pm 0.012}$	$0.8734_{\pm 0.019}$	$0.8638_{\pm 0.018}$	$\underline{0.7828}_{\pm 0.031}$	$\underline{0.7342}_{\pm 0.033}$	$\underline{0.6903}_{\pm 0.032}$	$0.9776_{\pm 0.028}$	$0.9097_{\pm 0.022}$	$0.8815_{\pm 0.025}$
	TOP [25]	$0.9067_{\pm 0.011}$	$0.8436_{\pm 0.009}$	$0.8417_{\pm 0.012}$	$0.7562_{\pm 0.017}$	$0.6918_{\pm 0.017}$	$0.6629_{\pm 0.015}$	$0.9786_{\pm 0.026}$	$0.9128_{\pm 0.024}$	$0.8962_{\pm 0.021}$
VL-M	ViLa-MIL [27]	$0.9174_{\pm 0.026}$	$0.8487_{\pm 0.021}$	$0.8422_{\pm 0.020}$	$0.7567_{\pm 0.014}$	$0.6824_{\pm 0.016}$	$0.6661_{\pm 0.021}$	$0.9756_{\pm 0.021}$	$0.8987_{\pm 0.029}$	$0.8758_{\pm 0.021}$
	MSCPT [9]	$0.9316_{\pm 0.032}$	$0.8672_{\pm 0.031}$	$0.8615_{\pm 0.031}$	$0.7819_{\pm 0.029}$	$0.7068_{\pm 0.023}$	$0.6681_{\pm 0.022}$	$0.9754_{\pm 0.021}$	$0.9111_{\pm 0.022}$	$0.8992_{\pm 0.022}$
	HiVE-MIL [33]	$0.9443_{\pm 0.023}$	$0.8604_{\pm 0.038}$	$0.8602_{\pm 0.040}$	$0.7428_{\pm 0.066}$	$0.7255_{\pm 0.016}$	$0.6701_{\pm 0.054}$	$0.9713_{\pm 0.011}$	$0.9018_{\pm 0.021}$	$0.8974_{\pm 0.025}$
	C^{2}UOT-MIL	$\mathbf{0.9671}_{\pm 0.013}$	$\mathbf{0.8896}_{\pm 0.018}$	$\mathbf{0.8894}_{\pm 0.017}$	$\mathbf{0.8038}_{\pm 0.014}$	$\mathbf{0.7422}_{\pm 0.018}$	$\mathbf{0.7033}_{\pm 0.017}$	$\mathbf{0.9826}_{\pm 0.018}$	$\mathbf{0.9205}_{\pm 0.024}$	$\mathbf{0.9124}_{\pm 0.041}$

V: vision-only MIL; **VL-S**: vision–language single-scale; **VL-M**: vision–language multi-scale. We report AUC/ACC/F1, mean±std(5-fold). **best** result in bold, second-best result with an underline.

Baselines. We compare our framework with SOTA methods across various MIL approaches, including (1) **Vision-only (V)** methods: standard MIL methods that do not incorporate any language priors: Mean-Pooling, ABMIL [12], DSMIL [17], CLAM [20], TransMIL [26], DTFD-MIL [34], WiKG [18] and RRT-MIL [29] (2) **Vision–Language Single-Scale (VL-S)** methods: evaluates the contribution of language priors under a single (20x) magnification setting: FOCUS [8] and TOP [25]. (3) **Vision–Language Multi-Scale (VL-M)** methods incorporate multi-granular textual concepts in conjunction with multi-level image instances: ViLa-MIL [27], MSCPT [9] and HiVE-MIL [33].

Implementation Details. Our C^{2}UOT-MIL is implemented by using C^{2}UOT as a principled enhancement atop the ViLa-MIL [27]. Specifically, C^{2}UOT replaces the image-text alignment component while retaining the visual prototype-attention branch of ViLa-MIL. We train both the baselines and C^{2}UOT-MIL using the Adam optimizer with $lr = 1 \times 10^{-4}$ and weight decay 1×10^{-5}. For a fair comparison, all vision-language baselines employ the same CONCH model [19] as the feature extractor for both image and text embeddings. We follow the most commonly used setting in **VL-M** research [9, 27, 33] to obtain the instance graph, which sets $L = 2$ and operates on both the 5× and 20× magnification levels of the WSI. At each level, CLAM [20] is used for WSI pre-processing, cropping patches to 512×512 pixels. Following standard practice in current VLM-MIL works [9, 23, 27, 33], we employ an LLM with the prompt design adopted from HiVE-MIL to obtain textual concepts. Specifically, all vision-language baselines use GPT-5 [24] to generate the concept graph.

5.2 RQ1: Overall Performance Comparison

We first analyze the performance under the 16-shot setting. As summarized in Table 1, C^{2}UOT-MIL achieves new state-of-the-art results on all datasets, surpassing the second-best VL baseline by 1.45% on TCGA-NSCLC (average AUC 0.9671) and 2.1% on TCGA-BRCA, which shows the efficiency of the proposed method. Under smaller few-shot settings, as shown in Figure 3, the advantage of C^{2}UOT-MIL becomes more pronounced. For example, on TCGA-NSCLC with 4-shot supervision, C^{2}UOT-MIL attains an AUC of 0.9492, whereas ViLa-MIL reaches only 0.8115. This 13.77% gap far exceeds the 4.97% gap in the 16-shot setting. This widening margin suggests that with very limited supervision, spurious alignments increasingly undermine general VL models. In contrast, C^{2}UOT-MIL introduces diagnostic principles to prevent pathologically implausible alignments that harm the model performance.

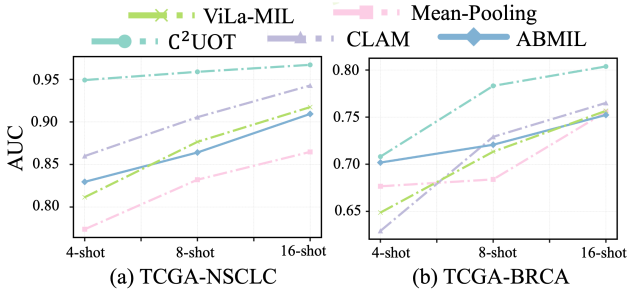


Figure 3: Other Few-shot Settings.

5.3 RQ2: Effectiveness Analysis

This section validates our core premise that generic VLM-MIL models suffer from generic instance-concept alignments which misguide slide-level classification. First, we provide empirical evidence demonstrating that alignment quality is the fundamental bottleneck in few-shot WSI classification (Sec 5.3.1 and Sec 5.3.2). Second, we validate how explicitly enforcing diagnostic principles via our proposed modules actively refine spurious matches and improve

Table 2: Instance-level tumor discrimination and bag-level classification on CAMELYON.

Methods	CAMELYON$_{Inst}$		CAMELYON$_{Bag}$	
	AUC	F1	AUC	F1
CLAM	$0.6729_{\pm 0.078}$	$0.5211_{\pm 0.192}$	$0.8324_{\pm 0.163}$	$0.7756_{\pm 0.201}$
ViLa-MIL	$\underline{0.7244}_{\pm 0.049}$	$\underline{0.6137}_{\pm 0.121}$	$\underline{0.9339}_{\pm 0.024}$	$\underline{0.8917}_{\pm 0.067}$
C^{2}UOT-MIL	$\mathbf{0.7687}_{\pm 0.046}$	$\mathbf{0.6542}_{\pm 0.043}$	$\mathbf{0.9536}_{\pm 0.062}$	$\mathbf{0.9125}_{\pm 0.061}$

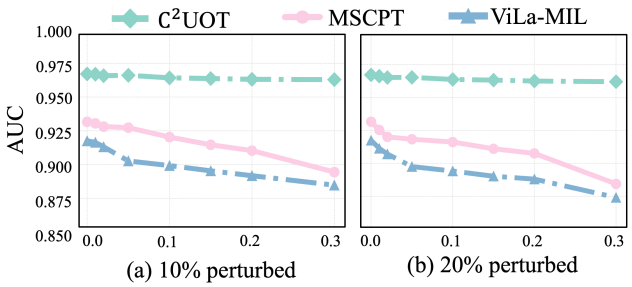


Figure 4: Experiments of Alignments Perturbation.

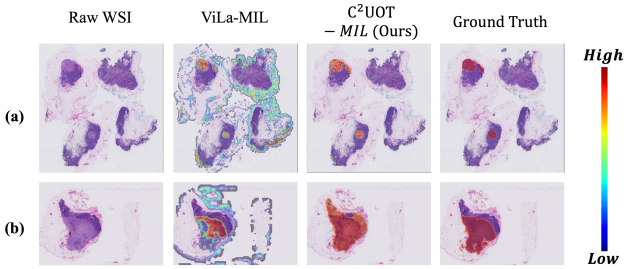


Figure 5: Qualitative visualization of tumor localization.

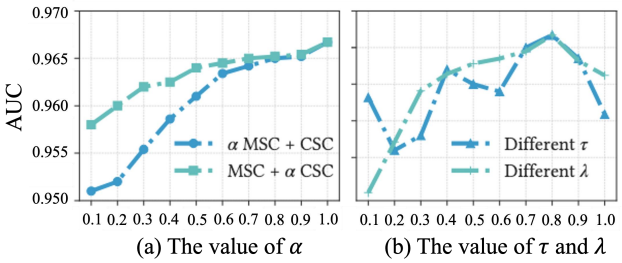


Figure 6: Effects of diagnostic principles & hyperparameters.

classification performance (Sec 5.3.3 and Sec 5.3.4). Finally, we provide qualitative visual proof of the refined alignments (Sec 5.3.5).

5.3.1 ***Impact of Spurious Alignments on WSI Prediction.*** We first investigate whether instance-concept alignment quality directly affects slide-level classification performance. To this end, we evaluate on the CAMELYON dataset under 16-shot, which uniquely provides binary labels for each individual patch (tumor vs. normal). This enables us to examine how alignment quality at the instance level (CAMELYON$_{Inst}$) translates into slide-level prediction quality (CAMELYON$_{Bag}$). In our framework, an instance is considered positive if the total transport mass assigned to the positive concept set (i.e., all textual concepts describing the tumor) in the optimal transport plan T exceeds that assigned to the negative concept set. Patch labels are only used during testing.

As shown in Table 2 and Figure 5, the generic alignment used in ViLa-MIL yields suboptimal accuracy, suggesting prevalent spurious alignments. C^{2}UOT-MIL rectifies this, boosting CAMELYON$_{Inst}$ AUC by 4.43%. Consequently, CAMELYON$_{Bag}$ improves from 0.9339 to 0.9536, confirming that correcting instance-level alignments directly enhances global WSI classification.

5.3.2 ***Robustness to Alignment Perturbation.*** To further validate whether alignment quality is a critical bottleneck of classification performance, we simulate aggravated spurious alignments by injecting additive noise $\epsilon \in [0.0, 0.3]$ into a subset (10%, Figure 4 (a) and 20%, Figure 4 (b)) of the initial instance-concept cost matrix elements. As shown, the degradation of baseline models confirms their fatal vulnerability to low quality alignments. In contrast, C^{2}UOT-MIL maintains a flat performance curve. Even under high perturbation, our framework effectively filters the injected noise by constraining the transport plan to comply with the immutable structural constraints inherent in the concept graph. This proves that our framework possesses an inherent corrective capability that refines spurious matches into diagnostically plausible ones.

5.3.3 ***Effects of key components in C^{2}UOT-MIL.*** We conduct an ablation study on two TCGA datasets under the 16-shot setting, as shown in Table 3. It can be observed that UOT without MSC and CSC obtains the worst performance across all datasets. Adding the MSC module yields a further +1.06% to +1.7% improvement, confirming the effectiveness of enforcing microenvironment consistency. Finally, enabling CSC and MSC together achieves the best overall results. The complementary effects of diagnostic principle-guided inductive bias modules suppress spurious alignments during instance aggregation, confirming the effectiveness of each component and their synergy within C^{2}UOT-MIL.

Table 3: Key components ablation on NSCLC and BRCA for slide-level classification.

Components			TCGA-NSCLC		TCGA-BRCA	
UOT	MSC	CSC	AUC	F1	AUC	F1
✓			$0.9428_{\pm 0.069}$	$0.8697_{\pm 0.122}$	$0.7622_{\pm 0.054}$	$0.6850_{\pm 0.027}$
✓	✓		$\underline{0.9591}_{\pm 0.009}$	$\underline{0.8826}_{\pm 0.018}$	$\underline{0.7728}_{\pm 0.046}$	$\underline{0.6918}_{\pm 0.032}$
✓	✓	✓	$\mathbf{0.9671}_{\pm 0.013}$	$\mathbf{0.8894}_{\pm 0.017}$	$\mathbf{0.8038}_{\pm 0.014}$	$\mathbf{0.7033}_{\pm 0.017}$

5.3.4 ***Effects of Diagnostic Principle Constraints.*** To rigorously verify that the performance gains of C^{2}UOT-MIL stem specifically from enforcing diagnostic principles, we introduce a *Constraint Coverage Ratio* $\alpha \in [0.1, 1]$. We rank instance-concept pairs by their *Violation Scores*, which are defined as hierarchical contradictions for CSC and spatial isolation for MSC. We then apply penalties only to the top α fraction. As shown in Figure 6(a), increasing α from 10% (less penalty) to 100% (full penalty) yields a clear, monotonic rise in AUC, with the peak attained at 100%. This shows that C^{2}UOT-MIL detects and suppresses spurious alignments via explicit constraints thus improving performance.

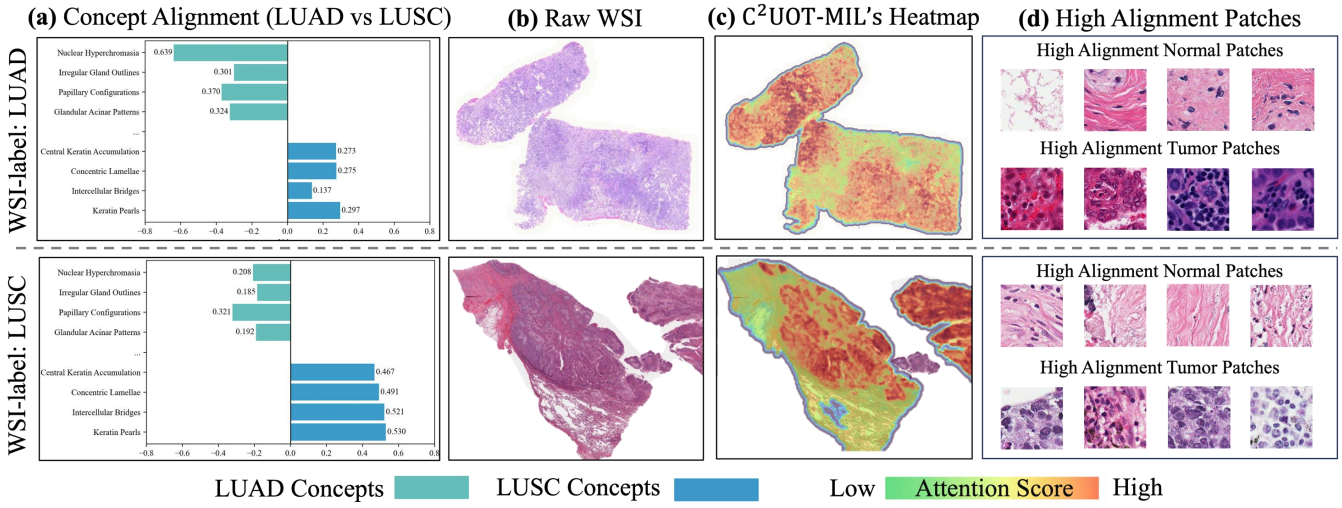


Figure 7: Interpretability analysis on the TCGA-NSCLC dataset. Including: (a) Concept contributions of WSIs, (b) Visualization of the original WSI and (c) Attention heatmaps of C²UOT-MIL. (d) High alignment patches focused by C²UOT-MIL

5.3.5 ***Explainability Analysis.*** To qualitatively demonstrate the efficacy of C²UOT in refining spurious alignments, we visualize two representative WSIs from distinct classes in TCGA-NSCLC. As shown in Figure 7 (a), the LUAD slide exhibits stronger alignments to LUAD-related concepts and vice versa. Figure 7 (b) and (c) show the raw WSI and the attention heatmap derived from C²UOT-MIL, respectively. The model assigns high attention to true tumor regions while effectively suppressing normal tissue. Finally, we sample representative instances with high and low attention values, respectively. As shown in Figure 7 (d), tumor instances display significant nuclear atypia, whereas the normal instances exhibit features of benign tissue. This experiment indicates C²UOT-MIL produces pathologically meaningful alignments, thereby delivering strong interpretability.

5.4 RQ3: Robustness and Algorithmic Efficiency

A practical clinical framework must be resilient to system-level variations. Here, we evaluate the robustness of C²UOT-MIL on the TCGA-NSCLC dataset under the 16-shot setting with respect to hyperparameter variations.

5.4.1 ***Robustness to Hyperparameters.*** Here, we analyze C²UOT-MIL's sensitivity to its two hyperparameters in Figure 6 (b). The coefficient λ controls the weight of inductive bias terms in Eq. (5). As λ increases from 0.1 to 0.3, model performance increases significantly, indicating MSC and CSC are essential. Beyond $\lambda > 0.3$, performance remains flat over a wide range (0.3 to 1.0), demonstrating C²UOT-MIL's robustness. We also assess the sensitivity to τ in Eq. (5) and observe that performance is also stable across a broad span of values.

5.4.2 ***Algorithmic Efficiency.*** We compare the computational cost of different VL-M methods in Table 4 on the TCGA-NSCLC dataset. We report the FLOPs of a single forward pass and the mean and peak GPU memory allocated during the corresponding inference process. While ViLa-MIL has the lowest overhead, its principle-agnostic alignment severely caps its performance. Among other models, C²UOT-MIL demonstrates remarkable efficiency, requiring only 712.25G FLOPs and 712.28 MB of peak GPU memory, which is nearly a 4× reduction in GPU memory usage compared to MSCPT [9].

Table 4: Comparison of computational cost.

Method	FLOPs(G)	Average GPU Memory (MB)	Peak GPU Memory (MB)
ViLa-MIL [27]	119.72	518.84	769.15
MSCPT [9]	983.28	1988.23	2736.62
C²UOT-MIL (Ours)	712.25	554.02	712.28

6 Conclusion

In this work, we introduce C²UOT, a diagnostic principle-guided alignment paradigm that overcomes the limitations of principle-agnostic generic alignment in VLM-based methods. By casting instance–concept matching as a principle-constrained UOT problem, C²UOT explicitly embeds pathologists' diagnostic principles into the alignment process. It further refines generic VLM-based alignment with two biologically inspired inductive bias modules: Cross-Scale Coherence (CSC), which enforces parent–child semantic consistency across magnification levels, and Microenvironment Semantic Consistency (MSC), which performs concept-family-aware spatial smoothing to suppress isolated artifacts. Experiments on four benchmarks demonstrate that instance-concept alignment quality is a critical bottleneck for slide-level prediction. By effectively refining generic alignments, C²UOT-MIL consistently outperforms SOTA baselines across all shot settings, delivering profound gains in the most data-scarce regimes.

Acknowledgments

Lei Chen's work is partially supported by National Key Research and Development Program of China Grant No. 2023YFF0725100, National Science Foundation of China (NSFC) under Grant No. U22B2060, Guangdong-Hong Kong Technology Innovation Joint Funding Scheme Project No. 2024A0505040012, the Hong Kong RGC GRF Project 16213620, RIF Project R6020-19, AOE Project AoE/E-603/18, Theme-based project TRS T41-603/20R, CRF Project C2004-21G, Key Areas Special Project of Guangdong Provincial Universities 2024ZDZX1006, Guangdong Province Science and Technology Plan Project 2023A0505030011, Guangzhou municipality big data intelligence key lab, 2023A03J0012, Hong Kong ITC ITF grants MHX/078/21 and PRP/004/22FX, Hong Kong ITC TC-SKLCRCC26EG01, Zhujiang scholar program 2021JC02X170, Microsoft Research Asia Collaborative Research Grant, HKUST-Webank joint research lab, 2025 HKUST Shenzhen-Hong Kong Collaborative Innovation Institute Green Sustainability Special Fund from Shui On Xintiandi and the InnoSpace GBA, and HKUST(GZ) - CMCC(Guangzhou Branch) Metaverse Joint Innovation Lab under Grant No. P00659. Jiachuan WANG's work is supported in part by JST CREST (JPMJCR22M2).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Wanxing Chang, Ye Shi, and Jingya Wang. 2023. Csot: Curriculum and structure-aware optimal transport for learning with noisy labels. *Advances in Neural Information Processing Systems* 36 (2023), 8528–8541.

[4] Chi-Long Chen, Chi-Chung Chen, Wei-Hsiang Yu, Szu-Hua Chen, Yu-Chan Chang, Tai-I Hsu, Michael Hsiao, Chao-Yuan Yeh, and Cheng-Yu Chen. 2021. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nature communications* 12, 1 (2021), 1193.

[5] Fan RK Chung and Robert P Langlands. 1996. A combinatorial Laplacian with vertex weights. *journal of combinatorial theory, Series A* 75, 2 (1996), 316–327.

[6] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine* 24, 10 (2018), 1559–1567.

[7] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, CAME-LYON16 consortium, Meyke Hermsen, Quirine F Manson, et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* 318, 22 (2017), 2199–2210.

[8] Zhengrui Guo, Conghao Xiong, Jiabo Ma, Qichen Sun, Lishuang Feng, Jinzhuo Wang, and Hao Chen. 2025. Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 15590–15600.

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

[10] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. 2023. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine* 29, 9 (2023), 2307–2316.

[11] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2023. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems* 36 (2023), 37995–38017.

[12] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. 2018. Attention-based Deep Multiple Instance Learning. In *ICML (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 2132–2141.

[13] Cheng Jiang, Xinhai Hou, Akhil Kondepudi, Asadur Chowdury, Christian W Freudiger, Daniel A Orringer, Honglak Lee, and Todd C Hollon. 2023. Hierarchical discriminative learning improves visual representations of biomedical microscopy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19798–19808.

[14] Si-yu Jing, He-qi Wang, Ping Lin, Jiao Yuan, Zhi-xuan Tang, and Hong Li. 2025. Quantifying and interpreting biologically meaningful spatial signatures within tumor microenvironments. *NPJ Precision Oncology* 9, 1 (2025), 1–14.

[15] John K Kruschke. 2010. Bayesian data analysis. *Wiley interdisciplinary reviews: cognitive science* 1, 5 (2010), 658–676.

[16] Junghwan Lee, Cong Liu, Junyoung Kim, Zhehuan Chen, Yingcheng Sun, James R Rogers, Wendy K Chung, and Chunhua Weng. 2022. Deep learning for rare disease: A scoping review. *Journal of biomedical informatics* 135 (2022), 104227.

[17] Bin Li, Yin Li, and Kevin W. Eliceiri. 2021. Dual-Stream Multiple Instance Learning Network for Whole Slide Image Classification With Self-Supervised Contrastive Learning. In *CVPR*. Computer Vision Foundation / IEEE, 14318–14328.

[18] Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. 2024. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11323–11332.

[19] 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. 2024. A visual-language foundation model for computational pathology. *Nature medicine* 30, 3 (2024), 863–874.

[20] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* 5, 6 (2021), 555–570.

[21] Oded Maron and Tomás Lozano-Pérez. 1997. A framework for multiple-instance learning. *Advances in neural information processing systems* 10 (1997).

[22] Tianhang Nan, Song Zheng, Siyuan Qiao, Hao Quan, Xin Gao, Jun Niu, Bin Zheng, Chunfang Guo, Yue Zhang, Xiaoqin Wang, et al. 2025. Deep learning quantifies pathologists' visual patterns for whole slide image diagnosis. *Nature Communications* 16, 1 (2025), 5493.

[23] 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. 2025. MGPATH: vision-language model with multi-granular prompt learning for few-shot WSI classification. *arXiv preprint arXiv:2502.07409* (2025).

[24] OpenAI. 2025. Introducing GPT-5. *OpenAI Blog* (2025).

[25] Linhao Qu, Kexue Fu, Manning Wang, Zhijian Song, et al. 2023. The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. *Advances in Neural Information Processing Systems* 36 (2023), 67551–67564.

[26] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. 2021. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. In *NeurIPS*. 2136–2147.

[27] Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. 2024. Vilamil: Dual-scale vision-language multiple instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11248–11258.

[28] Andrew H Song, Richard J Chen, Tong Ding, Drew FK Williamson, Guillaume Jaume, and Faisal Mahmood. 2024. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11566–11578.

[29] Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. 2024. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11343–11352.

[30] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. 2015. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia* 2015, 1 (2015), 68–77.

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[32] Jason W Wei, Laura J Tafe, Yevgeniy A Linnik, Louis J Vaickus, Naofumi Tomita, and Saeed Hassanpour. 2019. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific reports* 9, 1 (2019), 3358.

[33] Bryan Wong, Jongwoo Kim, Huazhu Fu, and Mun Yi. 2026. Few-shot learning from gigapixel images via hierarchical vision-language alignment and modeling. *Advances in Neural Information Processing Systems* 38 (2026), 71313–71354.

[34] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. 2022. DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification. In *CVPR*. IEEE, 18780–18790.